Principles of Forecasting: A Handbook for Researchers and Practitioners,
J. Scott Armstrong (ed.): Norwell, MA: Kluwer Academic Publishers, 2001.
Reprinted permission of Kluwer/Springer

Combining Forecasts

J. Scott Armstrong
The Wharton School, University of Pennsylvania

ABSTRACT

To improve forecasting accuracy, combine forecasts derived from methods that differ substantially and draw from different sources of information. When feasible, use five or more methods. Use formal procedures to combine forecasts: An equal-weights rule offers a reasonable starting point, and a trimmed mean is desirable if you combine forecasts resulting from five or more methods. Use different weights if you have good domain knowledge or information on which method should be most accurate. Combining forecasts is especially useful when you are uncertain about the situation, uncertain about which method is most accurate, and when you want to avoid large errors. Compared with errors of the typical individual forecast, combining reduces errors. In 30 empirical comparisons, the reduction in ex ante errors for equally weighted combined forecasts averaged about 12.5% and ranged from 3 to 24 percent. Under ideal conditions, combined forecasts were sometimes more accurate than their most accurate components.

KEYWORDS: Consensus, domain knowledge, earnings forecasts, equal weights, group discussion, rule-based forecasting, uncertainty.

Assume that you want to determine whether Mr. Smith murdered Mr. Jones, but you have a limited budget. Would it be better to devote the complete budget to doing one task well, for example, doing a thorough DNA test? Or should you spread the money over many small tasks such as finding the murder weapon, doing ballistic tests, checking alibis, looking for witnesses, and examining potential motives? The standard practice in matters of life and death is to combine evidence from various approaches. Although it is not a matter of life and death, combining plays a vital role in forecasting.

Combining has a long history that predates its use in forecasting. In 1818, Laplace claimed ?In combining the results of these two methods, one can obtain a result whose probability law of error will be more rapidly decreasing@ (as quoted in Clemen 1989). The value of combining was also appreciated by Galton (1878). Using photographic equipment to combine many portraits, he concluded (p. 135) that ?All of the composites are better looking than their components because the averaged portrait of many persons is free from the irregularities that variously blemish the look of each of them.@ Langlois and Roggman (1990), using computer composites of portrait photographs, added support for Galton; raters found composite faces to be more attractive as more faces were added. The more average, the better looking. Levins (1966), a biologist, suggested that rather than building one master model, it is often better to build several simple models that, among them, use all the information available and then average them. It has become respectable in the social sciences to combine results from different approaches. Important papers on this approach date back at least as far as the mid-1950s (e.g., Cronbach and Meehl 1955).

Combining forecasts, sometimes referred to as composite forecasts, refers to the averaging of independent forecasts. These forecasts can be based on different data or different methods or both. The averaging is done using a rule that can be replicated, such as to take a simple average of the forecasts.

Some researchers object to the use of combining. Statisticians object because combining plays havoc with traditional statistical procedures, such as calculations of statistical significance. Others object because they believe

there is one right way to forecast. Another argument against combining is that developing a comprehensive model that incorporates all of the relevant information might be more effective.

Despite the objections, combining forecasts is an appealing approach. Instead of trying to choose the single best method, one frames the problem by asking which methods would help to improve accuracy, assuming that each has something to contribute. Many things affect the forecasts and these might be captured by using alternative approaches. Combining can reduce errors arising from faulty assumptions, bias, or mistakes in data.

Over the past half-century, practicing forecasters have advised firms to use combining. For example, the National Industrial Conference Board (1963) and Wolfe (1966) recommended combined forecasts. PoKempner and Bailey (1970) concluded that combining was a common practice among business forecasters. Dalrymple=s (1987) survey on sales forecasting revealed that, of the 134 U.S. companies responding, 20% ?usually combined,@ 19% ?frequently combined,@ 29% ?sometimes combined,@ and 32% ?never combined.@ I suspect however, that they are referring to an informal averaging, which does not conform with the definition in this paper. In recent years, combining has also been adopted by weather forecasters, who call it *ensemble forecasting*. They improve accuracy by combining forecasts made at different lead times. For example, they combine Wednesday=s forecast for the coming weekend with the forecasts made on Monday and Tuesday (*Science*, Dec. 23, 1994, p. 1940).

Below, I summarize results from empirical studies of combining. I was aided by prior reviews, such as Clemen=s (1989) annotated bibliography, which included 209 papers. I did not find computer searches to be very useful. A search of the *Social Science Citation Index*, using “combining and forecasts,” produced 115 papers from 1988 to 2000. Only nine of these were among the 57 empirical studies I had located elsewhere. None of the remaining 106 proved to be relevant. To see why computer searches are unrewarding for this topic, examine the references at the end of this paper. Most titles provide no indication that the paper deals with combining forecasts.

To help ensure that my list of studies was complete, I put it on the Principles of Forecasting website (hops.wharton.upenn.edu/forecast) in January 1999, and appealed to researchers through e-mail lists to notify me about omissions. Many researchers responded with studies, some of which were relevant. Below, I summarize the relevant empirical studies that I could find on combining. I omitted a few studies that were hard to obtain or difficult to understand. However, their results were consistent with those included.

PROCEDURES FOR COMBINING FORECASTS

Combining forecasts improves accuracy to the extent that the component forecasts contain useful and independent information. Ideally, forecast errors would be negatively related so that they might cancel each other, but this is rare in practice. Lacking a negative correlation, one would hope to combine forecasts whose errors will be uncorrelated with one another. However, forecasts are almost always positively correlated and often highly so.

There are two ways to generate independent forecasts. One is to analyze different data, and the other is to use different forecasting methods. The more that data and methods differ, the greater the expected improvement in accuracy over the average of the individual forecasts.

$ Use different data or different methods.

Using several sources of data can add useful information and may also adjust for biases. For example, to forecast the number of cars to be imported by the U.S. from Japan, you could extrapolate the number that the U.S. counts as imports or from the number Japan reports having exported to the U.S. These numbers are likely to differ for such reasons as a desire to avoid import duties, an attempt to benefit from export subsidies, or differences in the definitions of the country of origin. For an example, using the same data and different procedures, consider the problem of whether it is better to forecast a quantity directly, say toothpaste, or whether to forecast each type of toothpaste by flavor and package size then, add them up. If good arguments can be made for either approach, you can use each and average their forecasts.

The use of different methods and different data go hand in hand. Baker et al. (1980) illustrated this in forecasts of the impact that offshore nuclear power plants would have on visits to adjacent beaches. One source of

forecasts was expert surveys, a second source was analogies (visits to beaches near land-based nuclear plants), and a third was surveys of nearby residents about their intentions to visit these beaches.

Batchelor and Dua (1995) examined forecasts of four variables (real GNP, inflation, corporate profits, and unemployment) for forecast horizons of 6, 12, and 18 months ahead. They had been prepared by 22 economists whose forecasts were summarized in the Blue Chip Economic Indicators. They reported reducing the Mean Square Error (MSE) by an average of 9.2 % by combining any two economists= forecasts, and by an average of 16.4% by combining ten of them. They also classified the forecasters based on the assumptions they reported using (43% Keynesian, 20% monetarism, and 12% supply side) and the methods (48% judgment, 28% econometric modeling, and 24% time-series analysis). When Batchelor and Dua combined forecasts based on diverse assumptions, they reduced the error more than when they combined forecasts based on similar assumptions. Similarly, the error reductions were larger when combining was based on different methods. The gains from dissimilar methods were not so pronounced as those from different assumptions, perhaps because many forecasters used more than one method in making their forecasts.

Lobo and Nair (1990) studied quarterly earnings forecasts for 96 firms from 1976 to 1983. They used two judgmental methods and two extrapolations to generate the forecasts. When they combined two judgmental methods (based on professional analysts= forecasts), the Mean Absolute Percentage Error (MAPE) decreased by 0.6% (compared with the average component error). When they combined two different extrapolation methods, the MAPE fell by 2.1%. When they combined a judgment method with an extrapolation method (I averaged the results of all four combinations), the MAPE fell by 5.2%.

While it is possible for a single forecaster to use different methods, objectivity is enhanced if forecasts are made by independent forecasters. One possible approach here is to find forecasts that have been published by others. Such a procedure was used by the Blue Chip Economic Indicators in collecting and summarizing macroeconomic forecasts. One might also use this procedure to obtain forecasts on key sectors such as automobiles, restaurants, or air travel. Some sources of published forecasts can be found on the Forecasting Principles website. On the negative side, published forecasts often fail to provide a description of how they were obtained.

$ Use at least five forecasts when possible.

When inexpensive, it is sensible to combine forecasts from at least five methods. As might be expected, adding more methods leads to diminishing rates of improvement. Makridakis and Winkler (1983) plotted the reduction in errors as more extrapolations were combined. They used extrapolations that had been prepared for the 1,001 series of the M-competition. The gains dropped exponentially. When they combined five methods, they had achieved most of the possible error reduction, but they obtained further small gains as they combined more than five forecasts.

The results for expert forecasting are similar to those for extrapolation. Using theoretical arguments, Hogarth (1978) advised using at least six experts but no more than 20. Libby and Blashfield (1978) conducted three empirical studies that showed substantial improvements in accuracy when going from one to three judges, and they concluded that the optimum would be between five and nine.

Ashton and Ashton (1985) studied judgmental forecasts of the number of advertising pages in *Time* magazine. Combining the forecasts of four experts reduced error by about 3.5%. While gains in accuracy continued as they included forecasts from up to 13 experts, they were small after the fifth expert.

A study of forecasts of four macroeconomic variables (Batchelor and Dua 1995) achieved nearly all the gains in accuracy by combining forecasts from 10 of 22 economists; however, they continued to make small gains as they added those of the remaining 12. Krishnamurti et al. (1999), in a study of short-term weather forecasts, concluded that forecasts are needed from six or seven models.

Lobo and Nair (1990) and Lobo (1991) studied quarterly earnings forecasts for 96 firms. They obtained average MAPEs by combining two methods (based on judgmental forecasts from two sources) and two extrapolation forecasts. Combinations of two methods, of which there were six, had an average MAPE of 57.4%. Combinations of three methods, of which there were four, had an average MAPE of 56.4%. Lastly, the combination of all four methods had a MAPE of 56.0%.

Combining also helps when the components themselves are combined forecasts. Winkler and Poses (1993) examined physicians= predictions of survival for 231 patients who were admitted to an intensive care unit. Here, physicians sometimes received unambiguous and timely feedback, so those with more experience were more accurate. They grouped the physicians into four classes based on their experience: 23 interns, four fellows, four attending physicians, and four primary care physicians. The group averages were then averaged. Accuracy improved substantially as they included two, three, and then all four groups. The error measure (the Brier score) dropped by 12% when they averaged all four groups across the 231 patients (compared to that of just one group).

$ Use formal procedures to combine forecasts.

Combining should be done mechanically and the procedure should be fully described. Equal weighting is appealing because it is simple and easy to describe. If judgment is used, it should be used in a structured way and details of the procedure should be recorded. Subjective weightings allow people to impose their biases. For example, assume that you were trying to forecast the effects of a state's use of capital punishment. People have strong biases about this issue and these would affect their forecasts. Biases are also common for forecasting in organizations. In Sanders and Manrodt's (1994) survey of sales forecasting practices, 70.4 % of the respondents said that they preferred to underforecast sales, while only 14.6% said they preferred to overforecast (the rest expressed no preference).

Avoid judgmental weights in cases where those doing the weighting lack information about the relative accuracy of alternative sources of forecasts. Fischer and Harvey (1999), in a laboratory study, asked subjects to combine forecasts from four forecasters. When the subjects had poor feedback about the accuracy of the components, judgmental weighting was less accurate than equal weights. However, judgmental weighting was more accurate when the subjects had good feedback about the accuracy of the sources.

Mechanical weighting schemes can help to protect against biases. Rowse, Gustafson and Ludke (1974) asked 96 experienced firemen, in groups of four, to estimate the likelihood of various situations, asking such questions as, "Is a fire more likely to be in a private dwelling or public building?" They compared a group consensus (the group discussed the question and reached agreement) against five mechanical procedures: equal weights, peer weights, self weights, group weights, and average weights (based on self weights and group weights). The mechanical schemes were all more accurate than the group consensus.

Lawrence, Edmundson and O=Connor (1986) compared judgmental combining with equal-weights combining. To do this, they first asked 35 subjects to make extrapolations for an 18-month horizon for 68 monthly time series. The subjects were provided with data tables and graphs. The researchers then asked each subject to combine the forecasts based on tables with those based on graphs. The judgmental combinations took longer and were less accurate than a mechanical combination.

Weinberg (1986) examined forecasts for attendance at performing-arts events. He used an econometric model to obtain ex ante forecasts for 15 events during 1977 and 1978. The events' managers used their judgment *and* the econometric model forecasts. On average, the econometric model alone was more accurate than the managers' forecasts (31.2% error versus 34.8%). The MAPE for a simple averaging of the manager's forecast and model, however, was more accurate at 28.9%.

$ Use equal weights unless you have strong evidence to support unequal weighting of forecasts.

Clemen (1989) conducted a comprehensive review of the evidence and found equal weighting to be accurate for many types of forecasting. However, the studies that he examined did not use domain knowledge. In the typical study, there was little reason to prefer any of the methods a priori. My conclusion from his review is that when you are uncertain about which method is best, you should weight forecasts equally.

Much of the evidence to support the use of equal weights has come from judgmental forecasting. For example, in examining forecasts of the outcomes of football games, Winkler (1967) found that weighting all judges equally was as accurate as weighting their forecasts according to their previous accuracy or according to their self-rated expertise. But in this case, the judges were students, not experts, and they did not receive good feedback.

Evidence from economics also supports the use of equal weights. MacLaughlin (1973) examined the accuracy of 12 econometric services in the U.S. The rankings of the most accurate methods for 1971 were *negatively* related to their rankings for 1972. In an analysis of five econometric models' forecasts for the UK from 1962 to 1967, Pencavel (1971) found no tendency for models that produced the most accurate forecasts in one year to do so in the next. Similarly, Batchelor and Dua (1990) concluded that "all forecasters are equal" in economics.

Weighting forecasts equally is useful when asking experts to forecast *change.* Armstrong (1985, pp. 91-96) reviewed this literature. However, expertise is useful for assessing the *current status* (level), so different weights might be appropriate if estimates of the level are an important source of error.

$ Use trimmed means.

Individual forecasts may have large errors because of miscalculations, errors in data, or misunderstandings. For this reason, it may be useful to throw out the high and low forecasts. I recommend the use of trimmed means when you have at least five forecasts. Forecasters have not studied the effects of trimming, so this principle is speculative. However, some researchers have compared means with medians (the ultimate trimmed mean), obtaining evidence that favors the use of medians.

McNees (1992), in his study of 22 experts who forecasted seven U.S. macroeconomic forecasts from the Blue Chip Economic Indicators, found little difference in accuracy between the mean and the median. The mean seemed superior when accuracy was measured by RMSE, while the median seemed to be a better measure when accuracy was measured by the Mean Absolute Error.

Agnew (1985) examined combined annual forecasts from the Blue Chip Economic Indicators for six variables: nominal GNP, real GNP, inflation, housing starts, corporate profits, and unemployment. Sixteen economists made one-year-ahead forecasts for the six years from 1977 through 1982. In total, the 16 economists each made 36 forecasts. The Mean Absolute Error for the median of the forecasts of the 16 experts was about 4.8% less than that based on the group mean.

Larreche and Moinpour (1983) asked business school students to make one-month-ahead market-share forecasts for eight different marketing plans in a simulation. In preparation, they gave the students data for 48 previous months. Twelve groups of five subjects made predictions. Larreche and Moinpour compared the 96 forecasts with the "true" values for this simulation. The group's median was more accurate than the group's mean for 56% of the comparisons.

X Use the track record to vary the weights if evidence is strong.

If you have good evidence that a particular method has been more accurate than others when the methods have been tested in a situation, you should give it a heavier weight. For example, in forecasting annual earnings for firms, ample evidence exists that judgmental forecasts are superior to extrapolations. This conclusion was based on a meta-analysis of the results from 14 studies that contained 17 comparisons (Armstrong 1983). If you were to combine forecasts of firms' annual earnings, then, you should give extrapolation forecasts less weight than judgmental ones.

Makridakis (1990), using 111 time series from the M-competition, found that individual methods that did better in ex ante forecast tests were more accurate in subsequent ex ante forecast tests. He compared four of the leading methods from the M-competition. Rather than use ex ante accuracy to weight the methods, he used it to select a single method. Lacking evidence to the contrary, I speculate that the optimum would lie somewhere between the equal-weights method and relying completely on the method that was most accurate in the validation tests.

Lobo (1991), extending the study of Lobo and Nair (1990), examined differential weighting for analysts= forecasts of company earnings. In a study of quarterly earnings forecasts for 1976 through 1983, he regressed actual values against component forecasts. Thus, he weighted the forecasts by their previous accuracy. He examined the accuracy of four different weighting schemes and of equal weights for a holdout period. The forecasts of all the weighted combinations were more accurate than the forecasts from equal weights. On average, equal-weights combining was off by 56% (using MAPE), whereas the average of the weighted combinations was off by 52.8%.

In a study of rainfall-runoff predictions in 11 regions, Shamseldin, O=Connor and Liang's (1997) equally weighted combined forecast reduced the MAPE by 9.4%. In contrast, the two procedures that weighted the forecasts according to the previous accuracy of the methods were more successful as they reduced the MAPE by 14.6% on average.

Krishnamurti et al. (1999) found that weather forecasts based on a combined forecast using weights based on regression were more accurate than combined forecasts with equal weights. These findings were based on short-term (one-to- three days-ahead) forecasts of wind and precipitation.

Given the track record of different forecasting methods, what do we do? The statistician's answer would be to weight by the inverse of the mean square error. I suspect that, given the instability of the MSE for forecasting, such a rule would not work well. The RMSE would seem preferable to the MSE and perhaps the Relative Absolute Error (RAE) would do even better. Little evidence exists on this issue. Whatever measure is used, I would shrink it toward equal weights, perhaps by using the average of the two weights.

$ Use domain knowledge to vary the weights on methods.

Those who are familiar with the situation may be able to make useful judgments about which methods are most appropriate. This would seem especially relevant when asked to judge from among a set of methods with which they have some experience. In particular, they should be able to identify methods that are *unlikely* to work well in the situation. To ensure that these weightings are reliable, you should obtain independent weightings from a group of experts. Five experts should be sufficient. If the weights show low inter-rater reliability, they should not be used to make differential weights. This advice is speculative, as I am not aware that the issue has been studied directly.

An alternative procedure is to structure the experts' domain knowledge. For example, they could be asked for their expectations about trends or whether they expect discontinuities. This approach was studied in Armstrong, Adya and Collopy (2001). Although the level of domain knowledge was low and only two experts were involved in specifying the knowledge, differential weights improved the accuracy of combined forecasts.

ASSESSING UNCERTAINTY

I have discussed the use of combining to forecast expected values. One can also combine alternative estimates of prediction intervals. For example, prediction intervals for a sales forecast can be obtained by asking more than one expert to provide 95% prediction intervals, and these intervals could be averaged. They could also be estimated using a holdout sample of ex ante forecast errors for an extrapolation method and combining the resulting prediction intervals with those from the judgmentally estimated intervals.

Uncertainty can also be assessed by examining the agreement among the components of a combined forecast. Close agreement among forecasts from *dissimilar methods* indicates construct validity, which should increase one=s confidence. Conversely, large differences should reduce confidence. However, it is difficult to convert these differences into prediction intervals. Furthermore, methods may produce similar results even when they are inaccurate. This occurs, for example, when experts forecast the probability of success of job applicants. Nevertheless, comparisons among forecasts are of some value, as the three studies below show.

The first study concerns estimation, not forecasting. Walker (1970) asked subjects to estimate the length of a line; the length, width, and height of a room; the weight of a book; the weight of a rock; the area of an irregularly shaped piece of paper; and the volume of a wastepaper bin. Four or more groups made estimates for each of the eight items. The groups consisted of an average of 16 subjects, each subject working independently. I reanalyzed the results to compare the agreement within each group and the group's accuracy. When the group=s judges were in agreement (coefficient of variation less than 10%), the MAPE was 7%. When they disagreed (coefficient of variation greater than 10%), the MAPE was 19%. So agreement did relate to accuracy.

Lobo (1992), in his study of professional analysts' forecasts of company earnings, put the forecasts into three equal-size groups based on their level of dispersion. The average MAPE for the high-agreement group was

10.8%, while for the moderate-agreement group it was 25.3%, and for the low-agreement group it was 55.6%. Here again, higher agreement was related to more accurate forecasts.

Plous (1995) conducted a series of experiments in which he asked subjects to specify the 90% confidence intervals for twenty almanac questions. Subjects working alone were poorly calibrated, as one would expect from prior research. Subjects working in groups were better calibrated, but were still much too overconfident. This occurred even if they were told that groups were overconfident, or if they used devil's advocate, or if they were asked to argue against their forecasts. The procedure that led to the best calibration was to collect confidence intervals from nominal groups (where three to four people worked alone); the highest and lowest estimates of confidence intervals among those in the group were then used to provide a confidence interval. This combining procedure led to estimates that were only slightly overconfident.

CONDITIONS FAVORING COMBINED FORECASTS

Combining is possible only if there is more than one sensible source of forecasts. Fortunately, it is often possible to use more than one forecasting method. It is assumed that each method has some validity, yet none of the methods provides perfect forecasts. If the best method is known a priori, it should receive more weight, perhaps all the weight. But generally, alternative methods are likely to add some value to the forecast.

High uncertainty calls for combining forecasts. For example, policy makers in some states in the U. S. are interested in predicting the effects of instituting nondiscretionary handgun laws. (These state that, once a person meets certain well-specified criteria for carrying a concealed handgun, he or she must be issued a permit upon request.) In this example, there is uncertainty about the types of data to use, what methods to employ, and who should do the analyses. Also, few people have experience with handguns and they typically have not examined data on the topic. Instead, they rely upon mass media reports and thus are subject to biases in these reports, in their selective attention to media, and in their interpretation. Assume that all states in the U.S. adopted nondiscretionary handgun laws. What changes in the annual number of murder cases in the U. S. would you predict? What would you predict about the change in deaths from mass shootings and in deaths from accidental shootings? For these issues, emotions are likely to affect the weights people assign to alternative forecasts. People are likely to put all of the weight on the forecast that supports their beliefs. They may also search for forecasts that support their views.

When emotions are involved, it is especially important to decide upon the weightings prior to examining the forecasts. Continuing with the gun control example, how would you weight forecasts from judgment and from econometric methods? In such a case, econometric methods seem less subject to biases and thus deserving more emphasis. But consider this. To forecast the impact of extending the nondiscretionary concealed-handgun law to all states, Lott (1998) used alternative econometric models. He drew upon economic theory in developing models and then derived parameter estimates based on county levels, time trends by state, and levels by state. He supplemented this by examining prior research. The component forecasts showed a consistent pattern and he combined them. Lott forecasted that murders would be reduced by at least 1,400 per year. He predicted an increase in the number of accidental deaths of nine (p. 112) and that mass shooting deaths would drop substantially (pp. 100-102). Lott's results upset many people and led to attacks on his integrity.

- **Combine forecasts from several methods when you are uncertain which forecasting method is most accurate.**

Combining is expected to be useful when you are uncertain as to which method is best. This may be because you encounter a new situation, have a heterogeneous set of time series, or expect the future to be especially turbulent.

Meade and Islam (1998) examined extrapolation models proposed for technological forecasting. Despite a large literature, it was not clear a priori, which method would be most accurate. Meade and Islam compared a selection rule (picking the best-fitting model) against a combined forecast. Using seven forecasting methods on 47 data sets, they found that the combined forecast was more accurate than the best fitting model for 77% of the ex ante forecasts.

Even if one can identify the best method, combining may still be useful if other methods contribute some information. If so, differential weights may improve accuracy.

- **Combine forecasts from several methods when you are uncertain about the forecasting situation.**

In situations of uncertainty, combining can reduce error. For example, Klugman (1945) found that combining judgments led to greater improvements for estimates of heterogeneous items (irregularly-shaped lima beans in a jar) than of homogeneous items (identically-sized marbles in a jar). In addition to ease of estimation, situations can be judged uncertain based on unexplained large variations in the past or on expected volatile changes in the future.

Because uncertainty increases with the forecast horizon, combining should be especially useful for long-range forecasts. Makridakis and Winkler (1983, Table 3) examined forecasts for periods 9 through 18 for 617 monthly series. The gains from combining increased as the forecast horizon increased. For example, combining two forecasts in horizon nine produced a 4.1% error reduction (MAPE reduced from 19.7% to 18.9%), whereas it produced a 10.1% reduction in horizon 18 (MAPE reduced from 45.5% to 40.9%).

Lobo (1992) analyzed quarterly earnings forecasts for 205 firms over the eight years from 1978 through 1985. Forecasts were made for four forecast horizons, producing a total of 6,560 forecasts. For one-quarter ahead forecasts, combining led to a corresponding reduction from 14.8% to 12.6%, a decrease of 2.2%. For four quarters ahead, the average MAPE for the components was 36.8% while it was 32.3% for the combined forecasts – a decrease of 4.5%.

Lobo (1992) also found combining to be more useful when analysts' forecasts differed more. Where they differed most (the top third of his 6,560 forecasts), the combined forecast MAPE averaged 57.6% versus 66.0% for the average component, a difference of 8.4%. For the low-dispersion group, the MAPE for the combined forecast was 12.6% versus 14.7% for the individual components, a difference of only 2.1%.

Schnaars (1986) combined seven extrapolations for 103 consumer products. For one-year-ahead forecasts, the MAPE for the combined forecast was less (11.5% vs. 9.7%). For the five-year-ahead forecasts, the combined MAPE was 38.3%, while the average forecast had a MAPE of 45.8%.

Sanders and Ritzman (1989) obtained contrary evidence. They used five extrapolation methods to make one-period-ahead forecasts for 22 time series from a public warehouse. They then split the series into two groups. For the 11 series having the most variation, combining reduced the MAPE by 10.4% (compared with the average of the individual techniques). Unexpectedly, the error reduction was greater (20.6%) for the low variability group.

New product forecasting involves much uncertainty, so combining should be useful there. Gartner and Thomas (1993) conducted a mail survey of new product forecasts of U.S. software firms and got responses from 103 of them. They divided them into two groups: 46 of them with fairly accurate forecasts and 57 with large errors. Firms in the more accurate group used more forecasting methods than those in the less accurate group. Assuming that they combined these forecasts in some way, the results are consistent with the hypothesis that combined forecasts improve accuracy.

- **Use combined forecasts when it is important to avoid large errors.**

Because the MAPE for a combined, equally weighted forecast is never greater than the typical forecast error, it will never be less accurate than the worst component. Thus, combining is useful when large errors might have especially serious consequences, such as when actions might lead to bankruptcy, death, or war.

On the other hand, if there is a premium on making the best forecast, as when bidding on contracts, it may be wise to avoid the crowd and be willing to tolerate large errors. For example, an economic forecaster who wants to be noticed might make an extreme forecast to gain recognition for foreseeing an unusual event whereas, if the forecast is wrong, it is likely to be ignored. Batchelor and Dua (1992) provide evidence that economic forecasters behave this way. In such a case, combining should not be used.

EVIDENCE ON THE VALUE OF COMBINING

Combined forecasts are more accurate than the typical component forecast in almost all situations studied to date. Sometimes the combined forecast will surpass the best method. This could be seen from a simple example involving offsetting errors. Assume that one forecast is 40 and another is 60, while the actual value turns out to be 50. Each of the components is off by 10, while the combined forecast has no error.

Evidence on the value of combining comes from studies on intentions, expert forecasts, extrapolation, and econometric forecasts. In addition, some evidence comes from studies of combining across different types of methods. I have focused on evidence from tests of ex ante forecasting that include comparisons to alternative procedures and, in particular, to combining equally weighted forecasts.

The gains in accuracy from equal-weights combining are influenced by many factors, including the number of forecasts combined, differences among the methods and data, accuracy of the component methods, amount of uncertainty about the methods selected, amount of uncertainty in the situation, choice of error measure, and length of the forecast horizon. Given the many sources of variation, it is difficult to estimate the reduction in error that combining forecasts can yield in a given situation.

To estimate the typical gain, I included all studies that provided ex ante forecasts and that reported on comparisons of combined forecasts and the average accuracy of their components. I did not include studies with less than five forecasts. When a study provided a number of comparisons, I used the one that exemplified the best practice. For example, if a researcher compared two forecasts, three forecasts, and four forecasts, I would use only the comparison of the four. When forecasts were made for different forecast horizons, I took an average across the horizons. In making comparisons across methods, I expressed errors in percentages. When possible, I reported on the proportion by which the error was reduced (e.g., the percentage reduction in the MAPE). To ensure that my summary is correct, I was able to contact authors of 22 of the studies, and received replies from 16 of them. This feedback led to corrections in summaries of two studies.

Exhibit 1 summarizes the findings. There were 30 comparisons and, on average, combining reduced forecast errors by 12.5%. Although the gains varied due to many factors, there were always gains.

Exhibit 1
Error Reductions from Combining Ex Ante Forecasts

Study	Methods	Components	Criterion	Data	Situation	Validation Forecasts	Forecast Horizon	Percent error reduction
Levine (1960)	intentions	2	MAPE	annual	capital expenditures	6	1	18.0
Okun (1960)	"	2	"	"	housing starts	6	1	7.0
Landefeld & Seskin (1986)	"	2	MAE	"	plant & equipment	11	1	20.0
Armstrong et al. (2000)	"	4	RAE	"	consumer products	65	varied	5.5
Winkler & Poses (1993)	expert	4	Brier	cross-section	survival of patients	231	varied	12.2
Thorndike (1938)	"	4 to 6	% wrong	"	knowledge questions	30	varied	6.6
Makridakis et al. (1993)	"	5	MAPE	monthly	economic time series	322	1 thru 14	19.0
Richards & Fraser (1977)	"	5	"	annual	company earnings	213	1	8.1
Batchelor & Dua (1995)	"	10	MSE	"	macroeconomic	40	1	16.4
Kaplan et al. (1950)	"	26	% wrong	cross-section	technology events	16	varied	13.0
Zarnowitz (1984)	"	79	RMSE	quarterly	macroeconomic	288	1	10.0
Sanders & Ritzman (1989)	extrapolation	3	MAPE	daily	public warehouse	260	1	15.1
Makridakis & Winkler (1983)	"	5	"	monthly	economic time series	617	18	24.2
Makridakis et al. (1993)	"	5	"	"	"	322	1 thru 14	4.3
Lobo (1992)	"	5	"	quarterly	company earnings	6,560	1 thru 4	13.6
Schnaars (1986)	"	7	"	annual	consumer products	1,412	1 thru 5	20.0
Landefeld & Seskin (1986)	econometric	2	MAE	annual	plant & equipment	7	1	21.0
Clemen & Winkler (1986)	"	4	MAD	quarterly	GNP (real & nominal)	45	1 thru 4	3.4
Shamseldin et al. (1997)	"	5	MAPE	annual	rainfall runoff	22	1	9.4
Lobo (1992)	expert/extrap	2	MAPE		company earnings	6,560	1 thru 4	11.0
Lawrence et al. (1986)	"	3	"	annual monthly	economic time series	1,224	1 thru 18	10.7
Sanders & Ritzman (1989)	"	3	"	daily	public warehouse	260	1	15.5
Lobo & Nair (1990)	"	4	"	annual	company earnings	768	1	6.4
Landefeld & Seskin (1986)	intentions/econ	2	MAE	annual	plant & equipment	11	1	11.5
Vandome (1963)	extrap/econ	2	MAPE	quarterly	macroeconomic	20	1	10.1
Armstrong (1985)	"	2	"	annual	photo sales by country	17	6	4.2
Weinberg (1986)	expert/econ	2	"	cross-section	performing arts	15	varied	12.5
Bessler & Brandt (1981)	exprt/extrap/econ	3	"	quarterly	cattle & chicken prices	48	1	13.6
Fildes (1991)	"	3	MAE	annual	construction	72	1 & 2	8.0
Brandt & Bessler (1983)	"	6	MAPE	quarterly	hog prices	24	1	23.5
							Unweighted average	**12.5**

Some of these studies were described earlier. The remaining studies are described here. The descriptions could help to assess the benefits to be expected from combining forecasts in a particular situation. Researchers might be interested in assessing the limitations of the evidence and determining how to improve the estimates. If you prefer to skip the details, you could go to the next major section, "Implications for Practitioners."

Intentions Studies

Levine (1960) presented forecasts of annual investment in U.S. plant and equipment from 1949 to 1954. The forecasts came from two intentions studies, one by the Securities Exchange Commission (SEC) and one by McGraw-Hill. Each survey asked company executives about their planned capital expenditures in the coming year. The SEC had a MAPE of 4.4% and McGraw-Hill=s was 4.0% – an average of 4.2%. Using their data, I calculated a combined forecast. The MAPE for the combined forecast was 3.5%, a reduction in error of about 18%.

Okun (1960) examined two intentions studies, *Fortune=s* survey of homebuilders and the Survey Research Centers= (SRC) survey of buyers= intentions, both used to forecast annual U.S. housing starts from 1951 through 1956. The SRC forecasts had a MAPE of 8.5% and *Fortune=s* was 7.5%. Using data in Okun's Table 2, I calculated a combined forecast using equal weights and the MAPE was 6.5%.

Landefeld and Seskin (1986) examined data from intentions surveys for next year=s plant and equipment expenditures conducted by the U.S. Dept. of Commerce=s Bureau of Economic Analysis (BEA), McGraw-Hill, and Merrill Lynch Economics. The BEA survey was the largest with 13,000 firms, while the other two each had less than 1,000 firms. The surveys were conducted in November and December prior to the years being forecast, and they covered the 11 years from 1970 through 1980. The BEA survey was the most accurate with a mean absolute error (MAE) of 1.9 percent, versus 2.7 for McGraw-Hill and 3.05 for Merrill Lynch. I calculated combined forecasts from their results. Combinations any two of the three forecasts reduced the MAE by 11.8% in comparison to the individual forecasts. When all three were combined, the error dropped by 20%, again showing the benefit of combining more than two forecasts.

Armstrong, Morwitz & Kumar (2000) combined forecasts for automobiles and wireless telephone service from four different intentions methods. The data, from the U.S. and France, covered forecasts with horizons ranging from 2 to 14 months. A total of 65 forecasts were made for various years from 1961 to 1996. Overall, the combined forecasts reduced the Relative Absolute Error (RAE) by 5.5%.

Expert Forecasts

Evidence on the value of combining experts= judgments goes back to Gordon (1924). She asked people to estimate weights as they lifted them. When she correlated the rankings with the true order, the average for 200 judges was .41. By averaging the rankings of any five judges chosen at random, she improved the average correlation to .68, and for 50 judges it was .94.

Stroop (1932) extended Gordon=s study by having a single individual make many estimates. Fifty estimates by the same individual led to more accurate rankings. Biased estimates were unlikely in this situation, so the gains were probably due to improved reliability.

Similar studies of estimation problems followed. Can one generalize from estimation problems? Based on Fischhoff's (1976) findings, the answer is yes. More important, forecasting studies have been conducted, as shown below.

Thorndike (1938) asked 1,200 subjects to predict 30 events. The average individual was incorrect for 38.1% of the forecasts. Combined forecasts from groups of four to six individuals were incorrect 35.6% of the time.

In the M2-Competition, the accuracy of five experts was compared with that for a combined forecast (Makridakis et al. 1993). The forecasting experts had no procedure for using domain knowledge. The data consisted of 23 monthly series with real-time forecasts made for one to 14 months ahead during 1988 and 1989 (from their Exhibit 3). Combining forecasts by the five experts produced a MAPE of 13.4%, compared with 16.5% for the average expert.

Annual earnings for firms are difficult to forecast, a factor that would favor combining. However, financial analysts draw upon similar information and they are often aware of other analysts' forecasts, factors that would reduce the benefits of combining. As a result, the correlations among analysts' forecasts are high; Richards and Fraser (1977) found an average correlation of .92 among nine analysts for earnings forecasts for 213 corporations in 1973. The average analyst=s MAPE for each of the firms was 24.7%. When Richards and Fraser calculated a combined forecast (the number of analysts was typically five), the MAPE was 22.7%.

Kaplan, Skogstad and Girshick (1950) asked 26 judges to forecast events in the social and natural sciences, obtaining over 3,000 forecasts. The average percentage of incorrect predictions for the judges was 47%. The combined forecasts were incorrect on 34% of the forecasts.

Zarnowitz (1984) examined forecasts by 79 professional forecasters for six variables for the U.S. economy. The forecasts, covering 1968-79, were collected by mail in the middle month of each quarter and covered a four-quarter horizon. This yielded 288 forecasts. Averaging across the six variables, Zarnowitz obtained a combined Root Mean Square Error ten percent lower than that for the typical individual=s errors.

Extrapolations

Newbold and Granger (1974) examined forecasts for 80 monthly and 26 quarterly time series. They used three extrapolation methods (Holt-Winters, Box-Jenkins, and stepwise autoregression). Although they did not assess the magnitudes of the gains, they concluded that combinations of forecasts from any two of the methods were superior to the individual forecasts most of the time.

Sanders and Ritzman (1989) examined one-day-ahead daily forecasts of shipments to and from a public warehouse. Their validation covered 260 forecasts over a one-year period. They used two different schemes for combining forecasts from three methods. On average, combining reduced the MAPE from 74.7% to 63.4% (calculated from data in their Table 1). The combined forecast was substantially more accurate than the best method.

Makridakis and Winkler (1983), using the 1,001 series from the M-competition, summarized the typical errors (line one of their Table 3) then showed what happens as up to ten methods were combined. Unfortunately, much of this analysis mixes annual, quarterly, and monthly time series. Nevertheless, the findings show that combining improves accuracy. For example, in the 18-ahead forecasts, where the sample consisted of 617 monthly series, combining two forecasts reduced the MAPE by 10.1%, while combining five forecasts reduced it by 24.2%.

In the M2-Competition (Makridakis et al. 1993), the accuracies of three exponential smoothing methods were compared with that for combined forecasts. The data consisted of 23 monthly series with real-time forecasts made for one- to 14-months ahead during 1988 and 1989 (their Exhibit 3). Combining three exponential smoothing methods produced a MAPE of 11.7%, compared with 12.2% for the average component.

Schnaars (1986) examined forecasts of annual unit sales for 103 products. He made forecasts for one to five-year horizons. He then used successive updating until the last one-year-ahead forecast. This provided a total of 1,412 forecasts by each of seven extrapolations. These series were difficult to forecast, and it was uncertain which method would be best. The average error from the seven methods was 23.5% (which I calculated from data in Schnaars' Exhibit 2). Using all seven methods, the errors from the equal-weights combined forecasts averaged 18.8%.

Econometric Forecasts

Landefeld and Seskin (1986) examined one-year-ahead forecasts for plant and equipment expenditures. They obtained forecasts from two econometric models that DRI and Wharton developed for seven years, through 1980. I calculated combined forecasts from their table. The MAE for the combined forecast was 21% less than that for the individual forecasts.

Clemen and Winkler (1986) examined the forecasts of GNP provided by four econometric models. Forecasts of Nominal GNP and Real GNP were made for horizons from one to four quarters for 1971 through 1982. This yielded about 45 forecasts for each variable from each of the econometric models. Using results from their Table 3, I calculated the typical Mean Absolute Deviation for each of the four methods. For nominal GNP, the equally weighted combined forecast (using all four methods) was 3.2% more accurate than the typical forecast. For real GNP, the combined forecast was 3.5% more accurate. For each variable and each forecast horizon, the combined forecast was nearly as accurate as the best of the component forecasts.

Shamseldin, O=Connor and Liang (1997) developed forecasts from five econometric models to predict the annual peak rainfall runoff in 11 areas in eight countries. They calibrated models using five to eight years of data and tested them on two years. The data covered the period from 1955 to 1980. A simple combined forecast reduced the MAPE from 37.1% to 33.6%.

Comparisons Across Methods

As shown earlier in this paper, combining is most useful when the component forecasts come from methods and data that differ substantially. When making comparisons across studies, however, combining across methods seemed no more effective than combining components based on the same method. This is probably due to the many sources of variation in making comparisons across studies.

Blattberg and Hoch (1990) forecasted catalog sales for clothing at two companies and customers= coupon redemption rates at three companies. They combined forecasts from an econometric model with forecasts made by a single buyer with the number of forecasts ranged from 100 to 1,008 across the five companies. The two methods were of roughly equal accuracy when used alone. In all five companies, the combined forecasts were better than the average and better than the best of the components. Unfortunately, because Blattberg and Hoch used R^2, I could not assess the magnitude of the gain.

Lobo (1992) examined the short-term annual earnings for 205 firms over eight years. He compared forecasts by analysts (an average of five professional analysts= published forecasts) with forecasts from three extrapolation models. He prepared three combined models, each using the financial analysts= forecast and an extrapolation forecast. For each of the four forecast horizons, the combined forecast was more accurate than the *best* of the components. The differences were always statistically significant. Compared with the average of the components, the MAPE dropped by about 11%.

Sanders and Ritzman (1992) used three years of 22 daily time series from a national public warehouse. Some series were fairly stable, while others fluctuated widely. Experts= judgmental forecasts were prepared by the warehouse=s supervisor in consultation with the warehouse manager. In addition, judgmental forecasts were obtained from 81 undergraduates. The students, randomly assigned to four of the 22 series, each produced 65 one-day-ahead forecasts and received feedback after each period. Sanders and Ritzman then made statistical forecasts based on an equally weighted combination of forecasts from three commonly used methods (single exponential smoothing, Holt=s two-parameter smoothing model, and an adaptive estimation procedure). When appropriate, Sanders and Ritzman made seasonal adjustments. Successive updating was used and the level was set equal to the last observation. All three forecasting methods were more accurate than the naive (no change) forecast. The combined forecast had a MAPE of 63.0% as compared with the 74.6% average for the components. In addition, the combined forecast was more accurate than the best component.

Earlier in this paper, I showed that combining the forecasts from three intentions studies described by Landefeld and Seskin (1986) reduced errors for one-year-ahead forecasts of plant and equipment expenditures. Combining two econometric forecasts also reduced errors substantially in this study. Now, what if we take the combined intentions forecasts and the combined econometric forecasts and combine them? When I did this for the seven years, 1974 to 1980, the MAE was reduced by an additional 11.5 %.

Lawrence, Edmundson and O=Connor (1986) asked 136 subjects to extrapolate 68 monthly series. Each subject made forecasts for horizons from one to 18 months. They had tables and graphs showing the data but they had no domain knowledge. The researchers also prepared extrapolations using deseasonalized exponential smoothing. A combination of the three forecasts (judgment based on the tables, judgment based on the graphs, and exponential smoothing) reduced the MAPE from 17.5% to 15.6%.

Vandome (1963) made forecasts for ten U.K. macroeconomic variables for the first two quarters of 1961, obtaining 20 ex ante forecasts. An econometric model had a MAPE of 6.25% and an extrapolation model was off by 5.05%, for an average component error of 5.65%. I calculated the combined forecast to have a MAPE of 5.08%.

In a study of the international camera market, I developed a combined forecast from an extrapolation and an econometric forecast (Armstrong 1985, p. 291). I made six-year backcasts for sales in 17 countries. The combined forecast had a MAPE of 31.6% versus the average component=s error of 33%.

In a study of forecasts for attendance at performing arts events, Weinberg (1986, Table 2) used an econometric model and managers= judgment to obtain ex ante forecasts for 15 events during 1977 and 1978. The MAPE for a simple combination of the manager and model was 28.9%, which was more accurate than either component and which reduced the MAPE of the components by 12.5%.

When market prices are involved, such as with the price of commodities, it is unlikely that any method will be as accurate as the market=s futures prices. However, combining can reduce the damage from bad forecasts. Brandt and Bessler (1983) used six methods to make one-quarter-ahead forecasts for U.S. hog prices. The methods included expert judgment, econometric models, and extrapolation. They examined forecasts for 24 quarters from 1976 through 1981. The combined forecast, with a MAPE of 7.3% was more accurate than the best component, whose MAPE was 9.5%. The combined forecast was also better than the best of the components (an extrapolation model). Bessler and Brandt (1981) used the same six methods to forecast prices for cattle and broiler chickens over the same time period. The combined forecasts based on three of those methods reduced the RMSE by 4.8% for cattle and 22.3% for broiler chickens. However, Brandt and Bessler found that instead of forecasting prices, farmers should have used current market prices.

Fildes (1991) examined construction forecasts in the UK. Forecasts were available from three sources: a panel of experts on construction, a naive extrapolation, and an econometric model. Annual forecasts were made for eight years for three sectors (private housing, industrial, and commercial construction) for a lead time of up to three years. This provided a total of 72 forecasts. On average, in comparison with the typical components, the equally weighted combined forecast reduced the MAE by 8.0 percent.

IMPLICATIONS FOR PRACTITIONERS

Organizations often call on the best expert they can find to make important forecasts. They should avoid this practice, and instead combine forecasts from a number of experts.

Sometimes important forecasts are made in traditional group meetings. This also should be avoided because it does not use information efficiently. A structured approach for combining independent forecasts is invariably more accurate.

For combining to be effective, one should have independent forecasts that are systematically combined. When this is not the case, combining is expected to have little value. For example, consider the problem of selecting the best from a number of job applicants when it is difficult to forecast their long-term success. To improve accuracy, some organizations use panels rather than a single person to interview a candidate. In this case, the information does not differ and the forecasts are typically based on unstructured discussions. As a result, one would not expect the panel to have an advantage over a single interviewer. In fact, Huffcutt and Woehr (1999) found the panel interview to be *less* accurate. (They speculate that the harm might be due to the stress induced by having the candidate face a group of interviewers.) You can gain the benefits of combining in this situation by conducting a series of individual interviews and then combining the interviewers' individual predictions.

Use combined forecasts when more than one reasonable method is available and when there is uncertainty about the situation and the selection of the method. Draw upon forecasts that make use of different information, such as forecasts from a heterogeneous group of experts. Use methods that analyze data in different ways.

If you have five or more forecasts, trim the mean, for example, by dropping the highest and lowest forecasts. Equal weighting provides a good starting point, but use differential weights if prior research findings provide guidance or if various methods have reliable track records in your situation. To the extent that these sources of evidence are strong, one can improve accuracy with larger departures from equal weights.

Combining is especially relevant when there is uncertainty about the method or situation and when it is important to avoid large errors. This implies that combining should be useful for inventory control. Chan, Kingsman and Wong (1999) used combining for forecasts of the monthly demand for ten printed forms used by a bank in Hong Kong. In comparison with forecasts by Holt's exponential smoothing, a combination based on four extrapolation methods allowed for a reduction of ten percent of the safety stock with no loss in service.

Bretschneider et al. (1989) summarized evidence from three surveys of forecasters working for U.S. state governments. Those that claimed to use combinations of forecasts had more accurate revenue forecasts than those that did not. This result is consistent with the finding that combining improves accuracy.

Because they encompass more information, combined forecasts are likely to have credibility among managers. This is speculative as I was unable to find studies on this topic.

IMPLICATIONS FOR RESEARCHERS

This review drew upon 57 studies that have contributed to principles for combining forecasts. In addition, there is a large literature that might have provided insights for these papers. For example, the paper by Bates and Granger (1969) was influential and stimulated research on combining.

Comparative empirical studies have been useful in providing evidence on the principles. As can be seen, however, the amount of evidence for each of the principles is typically limited. For example, to what extent should trimming be used? Further research would help to better define the procedures for combining and the conditions under which combining is most useful. In particular, how can domain knowledge be used in assessing weights?

It is hard to draw conclusions when looking across studies, as there are many aspects of combining. Exhibit 1 lists seven aspects of studies and this is only a partial list. For example, it is difficult to determine how the length of the forecast horizon is related to the gain from combining, as fewer than half of the studies examined anything other than a one-period ahead forecast. Furthermore, the effect of the forecast horizon length might be different for one-month-ahead as for one-year-ahead forecasts. Studies that would directly test these conditions would be especially useful. In other words, instead of assessing the effects of a variable across a number of studies, they would be tested within a study. This was done effectively, for example, by Batchelor and Dua (1995), who showed that combining was more effective when the data and methods differed substantially.

CONCLUSIONS

Combining is useful to the extent that each forecast contains different yet valid information. The key principles for combining forecasts are to use

- $ different methods or data or both,
- $ forecasts from at least five methods when possible,
- $ formal procedures for combining,
- $ equal weights when facing high uncertainty,
- $ trimmed means,
- $ weights based on evidence of prior accuracy,
- $ weights based on track records, if the evidence is strong, and
- X weights based on good domain knowledge.

Combining is most useful when there is

- $ uncertainty as to the selection of the most accurate forecasting method,
- $ uncertainty associated with the forecasting situation, and
- $ a high cost for large forecast errors.

Compared to the typical component forecast, the combined forecast is never less accurate. Usually it is much more accurate, with error reductions in the MAPE running over 12 percent for the 30 comparisons reviewed. Under ideal conditions (high uncertainty and combining many valid forecasts), the error reductions sometimes exceeded 20%. Also under ideal conditions, the combined forecasts were often more accurate than the best of the components. In short, the combined forecast can be better than the best but no worse than the average. That is useful for forecasters.

REFERENCES

D Agnew, C. (1985), "Bayesian consensus forecasts of macroeconomic variables," *Journal of Forecasting*, 4, 363-376.

D Armstrong, J. S. (1983), "Relative accuracy of judgmental and extrapolative methods in forecasting annual earnings," *Journal of Forecasting*, 2, 437-447. Full text at hops.wharton.upen.edu/forecast.

D Armstrong, J. S. (1985), *Long-Range Forecasting: From Crystal Ball to Computer.* New York: John Wiley. Full text at hops.wharton.upenn.edu./forecast.

D Armstrong, J. S., M. Adya & F. Collopy (2001), "Rule-based forecasting: Using judgment in time-series extrapolation," in J. S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA: Kluwer Academic Publishers.

D Armstrong, J. S., V. G. Morwitz & V. Kumar (2000), "Sales forecasts for existing consumer products and services: Do purchase intentions contribute to accuracy?" *International Journal of Forecasting* (forthcoming).

D Ashton, A. H. & R. H. Ashton (1985), "Aggregating subjective forecasts: Some empirical results," *Management Scienc*e, 31,1499-1508.

N Baker, E. J., S. G. West, D. J. Moss & J. M. Weyant (1980), "Impact of offshore nuclear power plants: Forecasting visits to nearby beaches," *Environment and Behavior*, 12, 367-407.

D Batchelor, R. & P. Dua (1990), "All forecasters are equal," *Journal of Business and Economic Statistics*, 8, 143-144.

D Batchelor, R. & P. Dua (1992), "Conservatism and consensus-seeking among economic forecasters," *Journal of Forecasting*, 11, 169-181.

D Batchelor, R. & P. Dua (1995), "Forecaster diversity and the benefits of combining forecasts," *Management Science*, 41, 68-75.

I Bates, J. M. & C. W. J. Granger (1969), "The combination of forecasts," *Operational Research Quarterly*, 20, 451-468.

D Bessler, D. A. & J. A. Brandt (1981), "Forecasting livestock prices with individual and composite methods," *Applied Economics*, 13, 513-522.

D Blattberg, R. C. & S. J. Hoch (1990), "Database models and managerial intuition: 50% model and 50% manager," *Management Science*, 36, 887-899.

D Brandt, J. A. & D. A. Bessler (1983), "Price forecasting and evaluation: An application in agriculture," *Journal of Forecasting,* 2, 237-248.

D Bretschneider, S., W. P. Gorr, G. Grizzle & E. Klay (1989), "Political and organizational influences on the accuracy of forecasting state government revenues," *International Journal of Forecasting*, 5, 307-319.

D Chan, C. K., B. G. Kingsman & H. Wong (1999), "The value of combining forecasts in inventory management – a case study in banking," *European Journal of Operational Research*, 117, 199-210.

D Clemen, R. T. (1989), "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559-583.

D Clemen, R. T. & R. L. Winkler (1986), "Combining economic forecasts," *Journal of Business and Economic Statistics*, 4, 39-46.

I Cronbach, L. J. & P. E. Meehl (1955), "Construct validity in psychological tests," *Psychological Bulletin*, 52, 281-302.

I Dalrymple, D. J. (1987), "Sales forecasting practices: Results from a United States survey," *International Journal of Forecasting*, 3, 379-391.

D Fildes, R. (1991), "Efficient use of information in the formation of subjective industry forecasts," *Journal of Forecasting*, 10, 597-617.

D Fischer, I. & N. Harvey (1999), "Combining forecasts: What information do judges need to outperform the simple average?" *International Journal of Forecasting* 15, 227-246.

I Fischhoff, B. (1976), "The effect of temporal setting on likelihood estimates," *Organizational Behavior and Human Performance*, 15, 180-184.

I Galton, F. (1878), "Composite portraits," *Journal of the Anthropological Institute of Great Britain and Ireland*, 8, 132-142.

D Gartner, W. B. & R. J. Thomas (1993), "Factors affecting new product forecasting accuracy in new firms," *Journal of Product Innovation Management*, 10, 35-52.

D Gordon, K. (1924), "Group judgments in the field of lifted weights," *Journal of Experimental Psychology*, 7, 398-400.

D Hogarth, R. (1978), "A note on aggregating opinions," *Organizational Behavior and Human Performance*, 21, 40-46.

D Huffcutt, A. I. & D. J. Woehr (1999), "Further analysis of employment interview validity: A quantitative evaluation of interview-related structuring methods," *Journal of Organizational Behavior*, 20, 549-560.

D Kaplan, A., A. L. Skogstad & M. A. Girshick (1950), "The prediction of social and technological events," *Public Opinion Quarterly*, 14 (Spring) 93-110.

D Klugman, S. F. (1945), "Group judgments for familiar and unfamiliar materials," *Journal of General Psychology*, 32, 103-110.

D Krishnamurti, T. N. et al. (1999), "Improved weather and seasonal climate forecasts from multimodal ensemble," *Science*, 285 (Sept. 3), 1548-1550.

D Landefeld, J. S. & E. P. Seskin (1986), "A comparison of anticipatory surveys and econometric models in forecasting U.S. business investment," *Journal of Economic and Social Measurement*, 14, 77-85.

I Langlois, J. H. & L. A. Roggman (1990), "Attractive faces are only average," *Psychological Science,* 1 (March), 115-121.

D Larreche, J. & R. Moinpour (1983), "Managerial judgment in marketing: The concept of expertise," *Journal of Marketing Research*, 20, 110-121.

D Lawrence, M., R. H. Edmundson & M. J. O'Connor (1986), "The accuracy of combining judgmental and statistical forecasts," *Management Science,* 32, 1521-1532.

D Levine, R. A. (1960), "Capital expenditures forecasts by individual firms," in National Bureau of Economic Research, *The Quality and Significance of Anticipations Data*. Princeton, N.J. pp. 351-366.

I Levins, R. (1966), "The strategy of model building in population biology," *American Scientist*, 54, 421-431.

D Libby, R. & R. K. Blashfield (1978), "Performance of a composite as a function of the number of judges," *Organizational Behavior and Human Performance*, 21, 121-129.

D Lobo, G. J. (1991), "Alternative methods of combining security analysts' and statistical forecasts of annual corporate earnings," *International Journal of Forecasting*, 7, 57-63.

D Lobo, G. J. (1992), "Analysis and comparison of financial analysts', times series, and combined forecasts of annual earnings," *Journal of Business Research*, 24, 269-280.

D Lobo, G. J. & R. D. Nair (1990), "Combining judgmental and statistical forecasts: An application of earnings forecasts," *Decision Sciences*, 21, 446-460.

N Lott, J. (1998), *More Guns, Less Crime*. Chicago: U. of Chicago Press.

D MacLaughlin, R. L. (1973), "The forecasters' batting averages," *Business Economics*, 3 (May), 58-59.

D Makridakis, S. (1990), "Sliding simulation: A new approach to time series forecasting," *Management Science,* 36, 505-512.

D Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord & L. F. Simons (1993), "The M2-competition: A real-time judgmentally based forecasting study," *International Journal of Forecasting*, 9, 5-22.

D Makridakis, S. & R. Winkler (1983), "Averages of forecasts: Some empirical results," *Management Science*, 29, 987-996.

D McNees, S. K. (1992), "The uses and abuses of consensus' forecasts," *Journal of Forecasting*, 11, 703-710

D Meade, N. & T. Islam (1998), "Technological forecasting - model selection, model stability, and combining models," *Management Science*, 44, 1115-1130.

N National Industrial Conference Board (1963), *Forecasting Sales.* Studies in Business Policy, No. 106. New York.

D Newbold, P. & C. W. J. Granger (1974), "Experience with forecasting univariate time series and the combination of forecasts," *Journal of the Royal Statistical Society*, Series A, Part 2, 131-165.

D Okun, A. E. (1960), "The value of anticipations data in forecasting national product," in National Bureau of Economic Research, *The Quality and Economic Significance of Anticipations Data*. Princeton, N.J.

D Pencavel, J. H. (1971), "A note on the predictive performance of wage inflation models of the British economy," *Economic Journal*, 81, 113-119.

D Plous, S. (1995), "A comparison of strategies for reducing interval overconfidence in group judgments," *Journal of Applied Psychology*, 80, 443-454.

N PoKempner, S. J. & E. Bailey (1970), *Sales Forecasting Practices.* New York: The Conference Board.

D Richards, R. M. & D. R. Fraser (1977), "Further evidence on the accuracy of analysts' earnings forecasts: A comparison among analysts," *Journal of Economics and Business,* 29 (3), 193-197.

D Rowse, G. L., D. H. Gustafson & R. L. Ludke (1974), "Comparison of rules for aggregating subjective likelihood ratios," *Organizational Behavior and Human Performance*, 12, 274-285.

D Sanders, N. R. & K. B. Manrodt (1994), "Forecasting practices in U.S. corporations: Survey results," *Interfaces* 24 (March-April), 92-100.

D Sanders, N. R. & L. P. Ritzman (1989), "Some empirical findings on short-term forecasting: Technique complexity and combinations," *Decision Sciences*, 20, 635-640.

D Sanders, N. R. & L. P. Ritzman (1992), "The need for contextual and technical knowledge in judgmental forecasting," J*ournal of Behavioral Decision Making*, 5, 39B52

D Schnaars, S. (1986), "A comparison of extrapolation models on yearly sales forecasts," *International Journal of Forecasting*, 2, 71-85.

D Shamseldin, A. Y., K. M. O'Connor & G. C. Liang (1997), "Methods for combining the outputs of different rainfall-runoff models," *Journal of Hydrology*, 197, 203-229.

D Stroop, J. R. (1932), "Is the judgment of the group better than the average member of the group," *Journal of Experimental Psychology*, 15, 550-560.

D Thorndike, R. L. (1938), "The effect of discussion upon the correctness of group decisions when the factor of a majority influence is allowed for," *Journal of Social Psychology*, 9, 343-362.

D Vandome, P. (1963), "Econometric forecasting for the United Kingdom," *Bulletin of the Oxford University Institute of Economics and Statistics*, 25, 239-281.

I Walker, H. E. (1970), "The value of human judgment," *Industrial Marketing Research Association Journal,* 6 (May), 71-74.

D Weinberg, C. B. (1986), "Arts plan: Implementation, evolution, and usage," *Marketing Science*, 143-158.

D Winkler, R. L. (1967), "Probabilistic prediction: Some experimental results," *Journal of the American Statistical Association,* 66, 675-685.

D Winkler, R. L. & S. Makridakis (1983), "The combination of forecasts," *Journal of the Royal Statistical Society*, Series A, 146, Part 2, 150-157.

D Winkler, R. L. & R. M. Poses (1993), "Evaluating and combining physicians' probabilities of survival in an intensive care unit," *Management Science*, 39, 1526-1543.

N Wolfe, H. D. (1966), *Business Forecasting Methods*. New York: Holt, Rinehart and Winston.

D Zarnowitz, V. (1984), "The accuracy of individual and group forecasts from business outlook surveys," *Journal of Forecasting*, 3, 11-26.

Acknowledgments: Dennis Ahlburg, P. Geoffrey Allen, David A. Bessler, Robert T. Clemen, William R. Ferrell, Robert Fildes, Nigel Harvey, Gerald J. Lobo, Stephen K. McNees, Nada R. Sanders, Robert L. Winkler, Thomas J. Yokum and Victor Zarnowitz provided helpful comments on early drafts. Editorial revisions were made by Raphael Austin, Ling Qiu and Mariam Rafi.