MANAGEMENT SCIENCE
Vol. 37, No. 5, May 1991
Printed in U.S.A.

INTERACTION OF JUDGEMENTAL AND STATISTICAL FORECASTING METHODS: ISSUES & ANALYSIS*

DEREK BUNN AND GEORGE WRIGHT
London Business School, Sussex Place, Regent's Park, London NW1 4SA *England*
Strathclyde Business School, Glasgow G4 OGE, United Kingdom

This paper reviews several of the current controversies in the relative value of judgemental and statistical forecasting methods. Where expert, informed judgemental forecasts are being used, a critical analysis of the evidence suggests that their quality is higher than many researchers have previously asserted, and circumstances favourable to this are identified. The issue of the interaction of judgemental and statistical methods is, however, identified as a more worthwhile line of inquiry, and research in this area is reviewed, differentiating approaches aimed at synthesising both of these inputs.

(JUDGEMENTAL FORECASTING; QUALITY OF JUDGEMENT; COMBINATION OF FORECASTS; BOOTSTRAPPING; CALIBRATION)

The purpose of this paper is to review the research on the quality of judgement in forecasting and the possible structures that are available for facilitating interaction with conventional statistical models. Fischhoff (1988) and Belsley (1988) have provided recent analyses of the role of judgement in quantitative modelling, and the observation that all serious forecasts require the exercise of some judgement is not, by itself, controversial. What is a major issue in current research is the *extent* to which judgement should be used in particular situations and *how* that process should be structured. With references mainly to the psychological research of the early 1970s which tended to describe the many ways that human judgement could be fallible (e.g. Tversky and Kahneman 1974), several researchers still argue against the general use of explicit judgement when statistical methods are available (e.g. Armstrong 1985, Makridakis 1988), except in special circumstances (e.g. major environmental or organisational change). Alternatively, empirical evidence has persuaded other workers, such as Lawrence et al. (1985) and Edmundson et al. (1988), that a well-structured judgemental process can consistently outperform a statistical model-based extrapolation. Part of this controversy can be explained by the differing nature and conditions of judgement used in the various comparative studies, and it is important to examine the features of the judgemental tasks when reviewing the evidence. More fundamentally, in the psychological research of the 1980s, a re-evaluation of the "heuristics and biases" research has taken place, and this suggests that we should qualify generalisations of judgemental fallibility into the forecasting context. This paper is organised around these issues.

The next two sections, therefore, examine the recent literature on the validity of judgemental forecasting and its comparison with statistical models. The conclusion is that studies in real-world settings do show the validity of judgement. §3 presents a critical analysis of the widely-held viewpoint, namely that judgement is flawed. Our focus is on the conditions that have produced variable evidence of the quality of judgement. Finally, §4 analyses the available structures for the interaction of judgement and statistical models. We suggest guidelines for further synthesis in order to mitigate much of the current controversy.

* Accepted by Vijay Mahajan; received May 24, 1988. This paper has been with the authors 14 months for 3 revisions. The ordering of authors is alphabetical.

0025-1909/91/3705/0501$01.25

1. Forecasting with Quantitative Methods: The Role and Validity of Judgement

Surveys of corporate forecasting practices show that most important forecasts involve judgement. Rothe (1978) surveyed 52 manufacturing firms and found 50 of them used judgemental methods in one form or another. Klein and Linneman (1984) surveyed 500 of the world's largest corporations and found that the overwhelming majority of corporate planners identified severe limitations in using purely statistical techniques. In the context of sales forecasting, Cerullo and Avila (1975) surveyed 110 companies drawn from the Fortune 500 firms, finding that 89% used judgemental forecasting alone or combined with other methods. These researchers defined judgement as opinion rather than judgemental adjustment to statistical models. Similarly, the survey evidence quoted in Dalrymple (1988) indicates the widespread use of judgement in sales forecasting, although in this case the judgemental process is defined as either a salesforce composite or jury of executive opinion. Thus, such surveys of usage do show a general inclination towards judgemental methods of forecasting, but there is clearly no uniformity in the actual processes of judgement being used.

As with survey data, published reports from practitioners have tended to endorse the use of judgement. For example, Lawson (1981) discussed traffic usage forecasting at Bell Telephone, observing that judgemental adjustments were commonly made to extrapolative methods to achieve "acceptable" forecasts. In many instances such adjustments were made on the basis of an "eyeball" analysis of time series plots. In the context of sales forecasts, Soergel (1983) and Jenks (1983) pointed out that only judgement can anticipate one-time events such as extraordinary competitive developments. Similarly, the outputs from large econometric models are routinely subject to judgemental adjustments (e.g. Malley 1975; Irland, Colgon and Lawton 1984; Glendinning 1975; Wallis et al. 1984–88; Turner 1990). Turner (1990) has examined the recent adjustment process in the major UK econometric forecasts. Although only a minority of the variables in the models are subject to residual adjustment, the effect of this is very significant on the forecasts. For example, the subjective residual adjustment on the London Business School 1988 forecast of consumer expenditure growth took the values from a model-based +3% to an adjusted −2%. This was footnoted as being justified by a change in the savings ratio in the late 1980s which had not, at that time, been effectively specified in the model. From the discussions in McNees and Perna (1981), Corker et al. (1986) and Turner (1990), it is apparent that most judgemental adjustments to econometric models are made for one of two reasons:

(1) *Specification Error*: The model has not been performing adequately recently and it is more expedient to perform an ad hoc adjustment to the output rather than re-specify the offending component of the model. This may be due to an omitted variable, change in a coefficient not yet statistically estimatable or unsatisfactory modelling of the dynamics of new policy changes.

(2) *Structural Change*: Some external factor or change in background assumption, not incorporated into the model is expected to influence future events. This is essentially a subjective intervention in time-series analysis terms. (In the psychological literature, this is often called a "broken-leg cue" (Meehl 1957) from analogy to the way that you would certainly adjust a statistical model of a person's mobility if you have just learnt that the person has a broken leg!)

Indeed it is because of the incorporation of such extra-model information, "broken-leg cues" in fact, that judgemental forecasts have been generally valued (Pankoff and Roberts 1968, Armstrong 1981, Brown 1988). However, despite such *surveys* and published professional *advice*, *comparative research* on the effectiveness of judgemental forecasting has been rather more mixed. As with the survey and advice literature, much of this work is difficult to consolidate into general conclusions because of many confounding factors which make individual studies difficult to compare. In particular, we shall see that it is important to establish:

1. **The nature of the comparison**
   (a) Statistical model compared to a judgemental adjustment
   (b) Statistical model compared to a separate judgemental forecast
2. **The nature of the experiment**
   (a) Were the subjects experts or students?
   (b) Were the tasks laboratory or real-world?
   (c) Were the tasks "informed" or "abstract"? (i.e. Did the subjects know anything about the time-series being forecast?)
   (d) Were the tasks repetitive or one-off?
   (e) Did the subjects receive feedback on their performance?
   (f) For experts in real applications, were the judgemental forecasts targets, budgets or best estimates?
3. **The nature of the methods**
   (a) Was "best practice" used in each case?
      (i) Was the statistical model well-specified with good diagnostics?
      (ii) Was the judgemental forecast coherent and defensible? Did the subjects document and provide an "audit trail" on the judgemental reasoning?
   (b) Was the judgement individual or an aggregation?
   (c) Was the judgement holistic or from a decomposition?
   (d) Were interactive graphics used to aid judgement?

We comment on these characteristics of the experiments and forecasting methods utilised in the context of several comparative studies below.

Effectiveness of Judgemental Adjustment: Experimental Evidence

In general, the extensive survey and professional advice on judgemental adjustment has been endorsed by many case studies and experiments. Using intervention analysis, Reinmuth and Guerts (1972) demonstrated that when an unusual event, like a promotion, occurs, time-series sales forecasts combined with judgement will substantially increase forecasting accuracy for the atypical period. Using a real company setting over a period of time, Mathews and Diamantopoulos (1989) showed the value of using extra-model information to adjust time-series forecasts of sales. Huss (1986) presented results to show, in the context of electricity sales, that judgemental adjustment of trends by company experts performed better than econometric methods and Wolfe and Flores (1990) have experimental results showing that an ARIMA approach to forecasting earnings can be improved with a structured judgemental adjustment process. In the econometric context, the annual studies of Wallis et al. (1984–88) on the performance of UK econometric models show that the forecasts would have achieved less accuracy if they had not been subject to their documented adjustments. All of this would seem to suggest that when experts are used in their familiar real-world forecasting context, then a judgemental adjustment process which is formalised as "best practice" in some or all of the ways indicated above and/or contains some formal structure will enhance the value of a good statistical model. With this in mind, it is interesting to re-evaluate two studies which have shown poor results from adjustment.

1. In an intervention study, Guerts and Kelly (1986) utilised extrapolation methods modified with two variables which "retailers often express" as having a major impact on sales: number of shopping days between Thanksgiving and Christmas and the number of weekends in the month, and found that these adjustments had negligible impact on the accuracy of the sales forecasts. Clearly, untested "conventional wisdom" can be at best irrelevant and at worst misleading. Judgemental inputs need to be defensible and ideally supported by evidence.

2. Carbone et al. (1983) reported an experiment using 25 time series and small teams of MBA students who prepared forecasts using the statistical methods of Box-Jenkins, Holt-Winters and Carbone-Longini. In addition, the teams had later to prepare judgemental forecasts on the basis of the "information available to them," which consisted of their previously made statistical forecasts. These investigators concluded that judgemental adjustment by the students did not improve accuracy, but they noted that their student sample did not have expert knowledge in the series they examined.

Effectiveness of Separate Judgemental Forecasts: *Experimental Evidence*

When judgemental methods are kept separate from statistical models, there is plenty of evidence to support the use of expert judgement. Basu and Schroeder (1977) found that a Delphi method using the "experienced judgement of 23 key corporate individuals" was more accurate than both regression and exponential smoothing in forecasting sales potential for a manufacturer of construction equipment. The output of the Delphi technique was also considered more "credible" by top management. In a recent study of the *Business Week* industry outlook surveys, Schnaars and Mohr (1988) investigated the validity of this "grass-root, judgemental approach to forecasting based on industry expertise." They noted that the magazine's claims for predictive accuracy ran counter to most of the advice contained in the academic literature. However, the experts in these surveys provided good forecasts in an absolute sense as well as in comparison to a simple benchmark model. Edmundson et al. (1988), in a business case-study context, have shown how the extra product information that experts possess can explain how a sales force composite judgemental forecast can outperform a time series model, and Brown (1978) has documented the superior performance of experts in financial earnings forecasting, again largely attributable to extra-model information (see also Brown 1988).

Many studies have shown little difference in comparing judgemental and statistical approaches. Armstrong (1981) investigated annual share earnings forecasts, observing that previous research on the accuracy of judgement and extrapolation in this context was mixed. Of eight studies reviewed by Armstrong, two concluded judgement was superior, one concluded extrapolation was superior and five were inconclusive. Earlier, Armstrong (1978), in a similar type of "head-count" review, again found no significant difference between econometric and purely judgemental forecasts in four separate studies. Similarly, Bessler (1980) found aggregate expert judgements of agricultural yields comparable in accuracy to those of an ARIMA model, as did Bunn and Seigal (1983) for certain aspects of daily electricity load forecasting from expert judgement and a regression model. It is possible that many of the comparisons are confounded by the reward structures associated with the judgemental forecasts and the extent to which they may be serving more as targets or budgets than best estimates. This seems to be the case in a recent study of IBM monthly shipments of a particular product where both Box-Jenkins and Holt-Winters seasonal models outperformed the expert forecasts (Wu et al. 1991).

There has also been an active theme of research in the laboratory setting, using students and largely "abstract" forecasting tasks, i.e. time-series consisting of nonattributed numbers. Whilst this research cannot address some of the real-world, expert-use issues, it has allowed some insights to be developed from the more controlled experiments. For example, Angus-Leppan and Fatseas (1986) found that, in experiments using accounting students to forecast short-term interest rates, judgemental extrapolation of graphical presentations was comparable to the best statistical method employed, a power curve, and that whether the students were "informed" about the series, or it remained an "abstract" task, made little difference. In a series of many experiments with students performing abstract "eyeball" extrapolations (Lawrence (1983); Lawrence, Edmundson and O'Connor (1985); Edmundson (1990)), judgemental forecasting was found to outperform standard time series techniques. Although contradictory results were found by Carbone and Gorr (1985), Edmundson (1990) has observed that Lawrence et al. (1985) introduced more decomposition into the judgemental task and his later work (Edmundson 1990) strongly argues for the value of decomposition in providing the basis for superior judgemental forecasts. Decomposition in this context refers to separate judgemental elicitation of level, trends, seasonals, etc., in a time series, rather than a holistic extrapolation.

Thus, the value of judgemental forecasts seems to be endorsed by both the researcher and professional forecaster. Where involved experts make forecasts in their areas of expertise, and if the elicitation process ideally has some formal structure (such as decomposition) and evidential support (avoiding untested "conventional wisdom"), then it is hard for the critics to argue against judgemental forecasts on performance grounds. Performance (in the accuracy sense) is not everything, of course, and statistical models may be much cheaper (see Mabert 1976) and far more amenable to consistent policy analysis and simulations (see later sections). We do, however, not wish to take a polemical position on this subject and will argue for a synthetic approach to incorporate the benefits of both methodologies. Nevertheless, it is important to understand why, in the face of the type of published evidence above, many writers on forecasting have still argued against the use of judgement when statistical models are available. Two areas of research have been influential in substantiating the more critical view of judgemental forecasting:

1. *The "Bootstrapping" Research.* Studies have shown that simple statistical models have incremental validity over experts in judgemental prediction tasks.

2. *The "Quality of Judgement" Research.* Psychological laboratory-based studies have shown that people are subject to many sorts of judgemental heuristics that can lead to biases.

We discuss the results of these two research themes in the next sections and show how more recent research is causing a re-evaluation of their generalisability to real-world forecasting practice.

2. The "Bootstrapping" Research

This research has looked at how effectively judgemental rules, used in intuitive prediction, can be represented by a statistical model of the judge. The model is an averaging model and so, it is argued, is less prone to random error. The method is simply to model, via linear regression, the relationship between what predictions or judgements a person makes and the information upon which the judgements are based. Early research studies (e.g. Bowman 1963, Meehl 1959, Goldberg 1965, Dawes and Corrigan 1974) tended to show that statistical models of judgement had more predictive accuracy than the judge upon whom the model was based, but later studies are now more equivocal about the superiority of the statistical model. Kunreuther (1969) was one of the first to observe that "an automatic rule may prove valuable as a guide to the manager but it should not be blindly followed" (p. 431) and speculated that models will do relatively better in situations where the process is stable and regular, but worse in irregular processes where peripheral information may have positive value. Ebert and Kruse (1978) made a similar point that judgemental forecasters should improve over the statistical models of themselves when they have specific information about the future which cannot be extrapolated from the past.

Many recent studies, conducted in real-world settings, have looked at bankruptcy prediction. Libby (1975), in a major study, had 43 experienced loan officers make predictions for 60 real, but disguised companies, half of which had failed. These predictions were made on the basis of the limited financial information contained in five financial ratios. Other information such as absolute amount of income, notes to the accounts, etc., was excluded from the experimental study. Nevertheless, the mean predictive accuracy of the loan officers' judgements was high, at 74%. However, in this artificially limited study only 9 of the 43 judges did better than a simple ratio of assets to liabilities (see Dawes 1979 for an insightful discussion of the issues).

Whitred and Zimmer (1985) point out that, in principle, loan officers may outperform a linear model by the valid use of nonlinear relationships between ratios and (non) bankruptcy. However, the robustness of the models to violations of nonlinearity will make this potential advantage of man over model practically immaterial. For loan officers to systematically outperform the model, they must have access to information unavailable to the model, information which may be prevalent in real life rather than in laboratory situations. In fact, Shepanski (1983) reported an experiment to test a linear representation and various nonlinear representations of information processing behaviour in the task of credit evaluations. Participants in the experiment were presented with sets of information describing prospective business borrowers in terms of payment record, financial condition and quality of the company's management. Shepanski argued that the credit judgement task is best represented by a nonlinear model. Additionally, in real-life credit valuations the composition and size of the information employed will change. Information gathering is costly and, for example, applications for a large loan will entail a much more comprehensive credit investigation than a small loan application. Such flexibility in information search cannot be captured by statistical modelling that is better suited to repetitive forecasts with a static number of predictor variables. However, as Dawes, Faust and Meehl (1989) have pointed out, the small number of studies that have provided clinicians with access to preferred sources of information have generally shown the superiority of the statistical model. As these authors note, human judgement can theoretically improve on statistical modelling by recognising events that are not included in the model's formula and that countervail the actuarial conclusion. Dawes et al. argue that such events are rare but, as we have already shown in §1, this is the exact situation where forecasting practitioners advocate the need for judgement. Indeed, more recent studies by Johnson (1988) and Blattberg and Hoch (1989) provide evidence of the quality of human judgement compared to statistical model when "broken leg" cues are part of the information available for decision making.

To illustrate, Chalos (1985) investigated the ability of an outcome-based credit-scoring model to assess financial distress and compared the performance of the model with that of loan review committees and individual loan officers. The major finding was that the loan review committees significantly outperformed the model and the individual officers. The model was a stepwise discriminant model built, using eight financial ratios as cue variables. The loan review officers/committees had **additional information** for each judgement in the previous three years' financial statements. Chalos' results indicated that loan committees may be beneficial, and the additional time required may be more than offset by the reduction in loan cost errors. In a related study, Casey and Selling (1986) noted that if a firm's specific financial data do not provide a clear-cut signal of its financial viability, then subjects would be expected to incorporate available prior probability information into their judgement processes. Such additional information is, of course, likely to be available in the work-a-day situations of loan officers.

Although early studies of linear modelling in clinical settings showed evidence that the model of the judge outperforms the judge on whom the model was based, the evidence of poor performance in experimental environments which are artificial in terms of the information available, in principle, to the forecaster, does provide a seriously restricted evaluation of the quality of experienced judgement. In the real-world studies, where the forecasting process has been less stable and less routine, with a need to incorporate special peripheral information, then the experts have outperformed bootstrapping models.

3. The Quality of Judgement Research

The studies of human judgement in real-world settings using experts—rather than college students—perhaps provide the greatest potential for the demonstration of the validity of human judgement, since no artificial ceiling is put upon human performance. Certainly these are the sort of situations where practitioner reports argue for the need to "adjust" statistical forecasts. In a recent paper, Beach, Christensen-Szalanski and Barnes (1987) argued that issues regarding the quality of human judgement are not settled and that a commonly held belief that human judgement is poor is not based on convincing data.

Phillips (1987) has argued that interest should now focus on **what people can do under favourable conditions** whereas the research literature has tended to be dominated by reports of what people actually do without help, guidance or training. He makes the point that conditions need to be appropriate for the generation of precise reliable and accurate assessments of probability and lists eight conditions which need to be satisfied. These include training in probabilistic thinking if the assessor is unfamiliar with probability concepts and the use of experts with substantive expertise in the area where judgements are required. In reviewing recent evidence, O'Connor (1989) has concluded that with well-trained, well-motivated experts, judgemental performance seems to be more reliable in practice than the previous literature has suggested.

Careful evaluation of this psychological research on the quality of judgement is very important, since it is often used as the basis for generalisation into forecasting practice. For example, perhaps the most commonly quoted article in the forecasting literature which expresses doubt about the capabilities of human judgement is that by Hogarth and Makridakis (1981). These authors argue that: "many of the numerous information processing limitations and biases revealed in the literature apply to tasks performed in forecasting and planning" (p. 115). However, in this case, we must observe that the biases quoted had mostly, at that time, been identified in undergraduate students' answers to simple paper and pencil tasks completed in the psychological laboratory. Further, most of the tasks had to do with judgement, per se, rather than judgement in forecasting. Hogarth and Makridakis, rather than the authors who subsequently cite their work, did recognise that these biases may not be generalisable outside the laboratory. Similarly, when these authors concluded that forecasters have a ". . . mistaken confidence in judgement" (p. 127), they were using the early references on the "calibration" of subjective probability judgements reviewed by Lichtenstein, Fischhoff and Phillips (1977). Calibration is one measure of the validity of subject probability assessment such that for a person to be perfectly calibrated, assessed probability should equal percentage correct where repetitive assessments are being used (see also Lichtenstein, Fischhoff and Phillips 1982). However, the studies of calibration that Lichtenstein et al. review have almost exclusively used general knowledge items in the form of dichotomous questions such as, 'Which canal is longer? (a) Suez Canal (b) Panama Canal.' In answering these questions, subjects are required to indicate their degree of belief in its correctness. General knowledge questions have been extensively used in studies of calibration because subjects' answers can be immediately and conveniently evaluated by the experimenter. This research has documented the generality of "overconfidence" in probability assessment.

More recently, it has been observed that probability assessments for future events involve different cognitive processes than those involved in putting a probability to the veracity of one's own memory. In a series of studies, Wright (1982), Wright and Wisudha (1982), Wright and Ayton (1986, 1988, 1989) have shown differences in calibration and related measures for sets of questions where the answer is already known (general knowledge verification) and where the answer is not known at the time of the probability assessment (judgemental probability forecasting). In general, people do not use as many certainty assessments in judgemental probability forecasting, and the forecasts tend to be better calibrated. One instance where judgemental probability forecasts are routinely generated is weather forecasting. The official forecasts issued by the National Weather Service in the United States are subjective probability forecasts. Murphy and Brown (1985) have evaluated these subjective forecasts and found that, for certain categories of weather, they were more accurate than the available objective statistical techniques. In this case, the forecasters have a very large amount of information available, including the output from statistical techniques. They also receive detailed feedback and have the opportunity to gain experience of making forecasts under a wide range of meteorological conditions. Furthermore, they have considerable practice in quantifying their internal state of uncertainty. These circumstances may well be ideal for the relatively successful application of judgemental, as compared with purely quantitative, forecasting.

Additionally, good calibration has been demonstrated in several real-world forecasting situations apart from weather forecasting. These situations include horseracing (Hoerl and Fallin 1974), prediction of the future interest rates by bankers (Kabus 1976), and prediction of the success of R & D projects (Balthasar, Boschi and Menke 1978). In Wallsten and Budescu's (1983) terms, there is an "existence demonstration" of valid judgemental probability forecasting. It appears that performance-demonstrated expertise in probability judgements is underpinned by practice and regular performance feedback. As Einhorn and Hogarth (1978) have argued, most judgements are made without the benefit of accurate feedback. Einhorn and Hogarth traced these difficulties to three main factors. The first is a lack of search for and use of disconfirming evidence and the second is the use of unaided memory for coding, sorting, and retrieving outcome information. Finally when people take an action based on a forecast in order to facilitate or avoid possible futures, they can often only observe feedback associated with the action taken and not the action not taken. This final factor is, of course, immaterial in contexts such as weather forecasting where actions cannot be taken to increase or reduce the likelihood of the forecast event. Unconfounded feedback in such circumstances is likely to prove more useful for the improvement of forecasting ability.

Murphy and Brown (1985) have argued that the presence of actual or potential users of judgemental forecasts provides the forecasters with a strong motivation for conducting the forecasting process in an efficient and more effective manner. Moreover, feedback from users of forecasts frequently contains information regarding possible improvements. The use of judgement in real-world forecasting thus contrasts strongly with the study of judgement in the psychological laboratory. Fischhoff (1988) notes that creating the conditions needed for learning judgemental forecasting as a **skill** may produce the same quality of forecasting performance in other domains as has been demonstrated with the meteorologists. The documented effectiveness of judgemental adjustments in the major UK econometric models (Wallis et al. 1984–88) supports this view.

To date, most of the research on the quality of human judgement in forecasting has followed the paradigm of comparing holistic judgement against a variety of statistical models. However, what is at issue is not that judgement is better or worse than models but that there are advantages and disadvantages in each approach which are best resolved by allowing **structured** interaction of judgement and statistical forecasting methods. Studies of the incremental validity of judgement on model outputs, although well-documented by anecdotes, have not been investigated in a systematic way. Similarly, investigations of the validity of judgement on model inputs are sparse. Given the evidence that we have reviewed for the quality of holistic judgemental forecasts, and given the multiple ways in which judgement can be incorporated with statistical models, it follows that these additional structural dimensions for interaction now need close investigation.

4. Model Structures for the Interaction of Judgement

What is apparent from the above survey is that much of the interaction of judgement with statistical models is of a casual, informal nature. The process is generally one of ex post adjustment to make the forecast acceptable. This is a suspect method, not because it may diminish accuracy (as we have seen, it generally improves accuracy), but because it is hard to justify to others and may undermine the credibility of the whole process. Glantz (1977) shows how vulnerable this procedure can be to adversarial challenge. The

U.S. Government was sued by a group of farmers for forecasting and issuing the associated directives for a drought which did not occur. The forecast was based upon judgemental adjustment of a quantitative model, which had it been used alone would have been more accurate in not instigating drought directives. Whilst these revisions were unquestionably done in good faith, the suit claimed that it represented unprofessional practice. Armstrong (1985) quotes this example, and issues a categorical "guideline" that forecasters should not adjust model predictions. Similarly, Geistauts and Eschenbach (1987) have argued that decision makers who rely on judgemental forecasts undertake greater personal risks than do those relying on 'objective' quantitative forecasts. These authors pointed out that statistical methods draw upon quantified data and follow a specific procedure, thus leaving what they called an 'audit trail' for examination and replication. By comparison, the steps and data used in the judgemental processes advocated by practitioners are inherently more difficult to describe, replicate and defend. Methodology in forecasting cannot guarantee accuracy in a particular instance, but it can establish credibility and defensibility, and if it improves the coherence of the approach, this should be reflected in improved average performance. One way to improve the credibility of the interaction is to give it more apparent "structure." In the Glantz (1977) example, above, had there been an explicit structure for the incorporation of judgement, or a 'model' of the judgemental influence, then it may have been more defensible. Essentially, what Armstrong (1985) and Geistauts and Eschenbach (1987) are warning against is the unstructured interaction of judgement and statistical model.

If there is to be a formal interaction of judgement and model, then there must be techniques of modelling judgement. We have seen the influences of the "bootstrapping" model in the previous section and have more generally witnessed the success of decision analysis in providing explicit models of judgemental uncertainty. More elaborate structures have also evolved over the past decade or so to explicitly structure an essentially subjective forecast. We can include the use of hierarchical inference (Barclay 1977), influence diagrams (Howard and Matheson 1984), scenario decomposition (Moskowitz and Sarin 1983), systems dynamics (Morecroft 1984) and possibly some aspects of expert systems in this. The extent that these structures can facilitate the interaction with statistical models depends upon the *level* of interaction between judgement and statistical model, the key *issues* associated with each level and the extent to which these issues provide, or require, "gateways" for the incorporation of judgement. We will discuss these issues in terms of two broad levels of interaction, viz. within a single model and across several models.

Judgement within a Single Model

Building a statistical model requires the input of many aspects of judgement. The various classes of models such as ARIMA, state-space, decomposition, regression, exponential smoothing, all present different judgemental problems and incorporate extra subjective information with varying degrees of facility. We can identify four general issues common within all classes of statistical model-building which open gateways for incorporating judgement.

(a) *Variable Selection*. We have seen that one of the results of the "bootstrapping" studies has been the observation that the usefulness of experts is often more in the set of variables to which they refer, rather than how they actually use them to make forecasts (Dawes 1975, Armstrong 1985). Indeed, even in the statistical and econometric literature, conventional wisdom is that variable selection should be essentially judgemental (with insightful use of diagnostic statistics) and not automated (recall the scepticism which "stepwise regression" usually fosters). The judgemental challenge is therefore the creative one of eliciting key variables from experts. Techniques of process tracing and knowledge engineering are clearly of value here (see Wright, Ayton and Whalley 1987).

The incorporation of intervention variables, or "broken-leg cues," has been seen in previous sections to be instrumental in making the use of judgement effective. Both the

Bayesian approach to this (e.g. West and Harrison 1989) and the more conventional structural modelling (e.g. Harvey and Durbin 1986) facilitate this with the aid of interactive graphic software, and this must therefore be a major ingredient in a more formalised approach to the use of extra-model information. Additionally, on this theme of using extra-model information to adjust a statistical model, the use of Saaty's "Analytic Hierarchy Process" by Wolfe and Flores (1990) is an interesting and apparently useful attempt to structure more formally the factors (and their relative weightings) used in the adjustment. We must assume that in general not all extra-model information will be specific enough to be encodable as an intervention and that some final adjustment will be required. The provision of structure at this level is important, as we have demonstrated, in order to provide defensibility and evidential support. At the very least, an "audit trail" (Armstrong 1985) should be recommended.

(b) *Model Specification.* In terms of specifying the relationships between variables (e.g. linear, nonlinear, lagged, etc.), the current attractions of "structural modelling" in econometrics and "state-space models" in forecasting are largely based around a desire for a more transparent understanding of the statistical model (Harvey 1984). This in turn is aimed at promoting greater judgemental interaction. These models embody generally well-understood factors such as trends, step-changes, level changes and seasonal effects and are manipulated in a less obscure way than, for example, in the Box-Jenkins ARIMA methodology. The judgemental tasks in traditional Box-Jenkins model identification have been difficult for the expert and confusing to the layman. Thus, it is easy to see why the trend in software for state-space forecasting has been towards the more interactive, inviting more judgemental input, whereas that for Box-Jenkins has been towards reliable automation, **removing** the judgemental problems. The interactive judgemental aid developed by Edmundson (1990) has performed particularly well in facilitating a purely judgemental structural decomposition for forecasting, and in a manner analogous to the more statistically-based structural modelling. There is clearly scope for a promising synthesis in these two areas of research.

Fischhoff (1988) notes that in modelling the decision processes of expert forecasters into automated expert forecasting systems new judgemental challenges are presented. Technical feasibility issues aside, how should one expert system be chosen from a range of alternatives on offer to the consumer? Belsley (1988) is particularly critical of attempts to automate the specification of statistical models. Within his framework for reliable model-building, quality is achieved through the attainment of *fit*, *meaning* and *diagnostics*, all of which, but especially that of being meaningful, requires the exercise of critical judgement.

We must await the incorporation of the current judgemental software for influence diagrams, hierarchical inference and "soft" systems modelling on to basic time-series methods to see interactive software helping both judgemental and statistical analysis. Certainly, the impact of influence diagrams, and their inherent connections to state-space modelling provide a useful modelling synergy in this respect (Chow and Oliver 1988).

(c) *Parameter Estimation.* Following the innovations in statistical decision theory of the 1960s (e.g. Raiffa and Schlaifer 1961), and the early enthusiasm for encoding subjective belief on parameters of models via Bayesian prior distributions, it would seem that a practical structure for incorporating judgement on parameter uncertainty only awaited further psychological research on effective elicitation procedures. Thus, Zellner (1971) showed how subjective prior distributions could be incorporated into regression models and, likewise, Harrison and Stevens (1976) developed a Bayesian Forecasting procedure around the use of prior distributions on a state-space formulation. Unfortunately, whilst experts may have considerable insights into the determinants of the variable, encoding belief on the parameters of an associated statistical model (regression coefficients, transition probabilities, etc.) appears to have been elusive. We must infer from the lack of applied case studies explicitly using encoded parameter distributions, and the recent tendency even for the advocates of "Bayesian Forecasting" to rely upon heuristic default values for the parameter estimates (Ameen and Harrison 1984), that this has been difficult to implement. Furthermore, attempts to impute parameter uncertainty from expressions of the uncertainty in the actual forecast variable (e.g. Pepper 1973 in an ARIMA context) have also met with little practical success. In a similar vein, the recent work of Cholette (1982) and Pankratz (1989) has attempted to use constraints (predictions) on future outcomes in the estimation of ARMA parameters as a way of incorporating extra-model information. Although, apparently, a very useful approach in this context, we must await more studies to assess its practicability.

Thus, it would seem that what was a promising theoretical approach for the rational incorporation of judgement on model parameters has failed to make a major impact in forecasting, and that parameter estimation is still primarily a statistical process. Perhaps, with the more comprehensible features of some structural models, judgement may be more readily incorporated. Certainly, the "gateway" is there. It may, however, be the case that most users of forecasts do not actually have much subjective knowledge to bring to the parameters, per se, and that imposing constraints, rather than a full measure of uncertainty, over their ranges is the best that can be done.

(d) *Data Analysis*. Data Analysis parallels the Variable Selection issue, in that it is still a primarily judgemental aspect of model-building. How far back to go in a time-series involves judgements of structural stability. How much to correct data, ex post, for special events is a judgemental decision upon how conditional the forecasts should become with respect to unusual variation. Sometimes, if unusual events can be better predicted outside the basic model (e.g. television effects on electricity demand, Bunn and Seigal 1983), then judgemental inputs are required both to forecast these events and take out their effects from the historical time-series. The more explicitly this can be done, the more defensible is the practice. Statistical methods to help identify step changes, outliers and influential observations have improved considerably in recent years (e.g. Belsley et al. 1980), and software has become more conscious of data analytic aspects, but the tasks still remain essentially judgemental and are open challenges for judgemental researchers into forecasting practice.

Thus, in terms of incorporating judgement at the level of a single model, we suggest:

(a) that the **variable selection** and **data analysis** aspects are primarily judgemental, looking for more structure from judgemental researchers and more support from statistical diagnostics.

(b) that **model specification** is currently the most promising issue for the development of integrated judgemental / time-series methods as structural modeling techniques in both areas (e.g. state-space, decomposition and influence diagrams) are becoming quite complementary.

(c) that, conversely, the scope for judgemental interaction on **parameter estimation** seems to have lost its early Bayesian promise.

Judgement across Several Models

Given the results of several forecasting models, judgement needs to be exercised on the issues of:

(a) which to select and which to reject;

(b) how to produce an effective combination of those selected;

(c) whether to combine a separate judgemental forecast on the outcome variable with the models, adjust the models, or incorporate judgement within the models.

Very little research has appeared on the judgemental selection and rejection of forecasting models. Furthermore, whilst there are established statistical tests for selecting the "best" model, there are no established guidelines for selecting the best subset of models for combination. That may not be so important when computational capabilities allow a search of all subsets and the combination is to be done along statistical lines, but if the data base does not allow such an empirical approach, then it becomes an important open problem. The only guidelines we have are the evident advantages of using models based upon very different assumptions and/or data bases and the need to avoid positive correlations between forecast errors (Bunn 1987). The former is easy to see, a priori; the latter may be difficult to assess.

Almost all of the work in combining quantitative models has taken a statistical approach to the combination, which involves very little judgement (Bates and Granger 1969; Newbold and Granger 1974; Winkler and Makridakis 1983). The method of assigning judgemental "outperformance" probabilities by the forecaster across the set of models has shown useful, if limited, applicability (Bunn 1975, 1985; Bessler and Chamberlain 1987; Gupta and Wilton 1987). This limited use of judgement in encoding belief across quantitative models (rather than on events) is to be contrasted with the large body of research in combining expert forecasts using subjectively assigned weights on the experts (see Lock 1987, for a recent review).

In dealing with multiple forecasting models by the use of a statistical method of combination, the choice of the appropriate method of combination itself is now an open issue of judgement. From the dozen or so heuristic variations on the original Bates and Granger (1969) approach, through robust Bayesian variations (e.g. Bunn 1977, Agnew 1986), to the variety of regression-based methods (Granger and Ramanathan 1984, Diebold and Pauly 1987), there is a considerable selection of estimation techniques within the class of linear variance-minimising combinations. Furthermore, outside this class, the multiobjective formulation of Reeves and Lawrence (1982) offers scope for combinations subject to several criteria, beyond just error variance, and possibly of a more subjective nature. Again, there are very few guidelines available here to aid judgement. Where the forecast errors are likely to be positively correlated, efficiency in practice requires a robust method of combination, and equal-weighting is often hard to beat in this situation, providing the individual models are in themselves good predictors. Differential weighting, on the other hand, is often required to devalue the impact of "poor" models. Where differential weighting schemes are being used, there is a strong requirement that they be adaptive in order to deal with the nonstationary nature of relative performance in the forecasting models.

5. Conclusions

We have documented an extensive review of the role and validity of judgement in statistical forecasting. In terms of the judgemental component, our review suggests that when **experts are used in their real world context and the judgemental process is made explicit through a form of decomposition or audit trail,** empirical studies and surveys of practice give a general endorsement of its value. A re-evaluation of the basic psychological research on *bootstrapping*, *judgemental biases* and *calibration* in this context has provided further support for the quality of "best practice" judgement in forecasting.

In terms of facilitating interaction with statistical models, several levels need to be considered.

Within a single model, we have suggested that **variable selection** and **data analysis** are primarily judgemental aspects, with some help from diagnostics, and look mainly to judgmental researchers for more explicit support; that **model specification** is possibly the most promising area for integrated structural modelling between statistical and judgemental methods and that, surprisingly, **parameter estimation** may now have less to offer in judgemental interaction, remaining largely statistical. Extra-model information can be effectively used to adjust the output from a model **if there is some explicit structure** to the process.

Across several models, the use of outperformance probabilities seems to offer a synthetic way of combining explicit judgement and data to effect a combination.

The practice of ex post adjustments to quantitative models is common and generally effective, but, we argue, it is too informal to be defensible, and undermines the communicability of the forecast. In terms of introducing a more formal structure:

(a) At the very least, a well-documented **audit trail** should explain verbally the reasoning behind the adjustment process. Turner (1990) makes a strong appeal for this in order to help users of econometric forecasts understand the full assumptions that were incorporated in forecast. Producers of forecasts which will be used by others should actually go further than this and provide some sensitivity analysis of the effect of such judgemental adjustments. Furthermore, the absence of an audit trail weakens the value of feedback on judgemental interventions in seeking to improve forecasts.

(b) More formally, an **adjustment decomposition structure** such as an application of the Analytical Hierarchy Process (as in Wolfe and Flores 1990), a scenario or causal map (as in Nilsson 1989), an influence diagram (as in Chow and Oliver 1988) or a variation of some of the "soft systems" representations (Rosenhead 1989) can provide a more explicit and systematic representation of the rationale behind the adjustment. Whilst this gives more structure to the adjustment, it still represents an ad hoc treatment of the output from the basic statistical model without any real interaction with the model and is mostly qualitative.

(c) Ideally, an **interactive decomposition structure,** whereby the specification of the subjective evidence and the statistical model are both part of an overall coherent modelling effort, would provide the most defensible and credible approach. We only really have clues to the future here. State-space methods are amenable to both qualitative (e.g. "influence diagrams") and statistical (e.g. "structural models") specifications and would seem to be the most promising. The purely subjective decompositional software of Edmundson (1990) which allows judgement to be incorporated on trend, seasonal and residual components has so much structural similarity to common statistical decomposition methods that some synthesis must be attractive. The ideal is to have interaction of judgement and quantitative modelling at the level of components in an overall model, rather than at the level of ad hoc adjustment of statistical model outputs.

When expert knowledge is sufficiently substantial to provide an alternative forecast, then an interesting issue is whether a combination should be used or whether this knowledge should be formalised as an interactive adjustment, as above. To a certain extent, this will depend upon context and the criteria under which the forecasting is to be evaluated.

(a) If **accuracy** is the main objective, a combination may be the most efficient. Indeed, the literature on combining is now so supportive that the baseline method should now be taken as a simple combination. Discussion should then proceed as to why the combining weights should be different.

(b) If the forecasting model is to be the basis of **policy simulation,** then coherent relationships of input to output variables are of essential importance (such relationships can get obscured, or even biased, in simple combinations of models based only upon outputs). This seems to suggest a case for an interactive decompositional structure, based upon formal adjustments as discussed above. Certainly, the ideal of an overall coherent model renders the forecast more **defensible to adversarial criticism,** or more easily **communicated** to others than is often the case with a combined forecast, which can sometimes appear to be a pragmatic mixture of different perspectives.

Research on interaction models is still a long way behind the knowledge that we have separately upon judgemental and statistical models, yet this would appear to be the more relevant theme. Fildes (1989), in looking at construction industry forecasts, found that modelling the errors of industry expert forecasts via a multivariate statistical approach was more effective than forming a separate combination. Alternatively, Clemen and Winkler (1987) in a weather forecasting context found a simple combination of subjective and objective models to be more accurate than using the objective data to reduce the residuals of the subjective forecasts. Perhaps the key to choosing whether to combine forecasts or use one model to reduce the residuals of the other depends upon whether both models are efficient in their own right (**leading to a combination**) or if the information in one of the models only has incremental, but not equivalent, value to the other (**leading to an adjustment**). Despite the extensive literature on this subject, our conclusions for practice are very speculative and the research need for extensive applied research on comparative methods for facilitating an interaction of judgemental and statistical is very apparent.[1]

[1] The authors wish to acknowledge the constructive advice of the Departmental Editor, Vijay Mahajan, an Associate Editor and four reviewers through several revisions of this paper.

References

ABDEL-KHALIK, A. R. AND K. EL-SHESHAI, "Information Choice and Utilization in an Experiment of Default Prediction," *J. Accounting Res.*, (Autumn 1980), 325–342.

AGNEW, C. E., "Bayesian Consensus Forecasts of Macroeconomic Variables," *J. Forecasting*, 4, 4 (1986), pp. 363–376.

AMEEN, J. AND P. HARRISON, "Discount Weighted Estimation," *J. Forecasting*, 3, 3 (1984), 285–296.

ANGUS-LEPPAN, P. AND V. FATSEAS, "The Forecasting Accuracy of Trainee Accountants Using Judgemental and Statistical Techniques," *Accounting and Business Res.*, (Summer 1986), 179–188.

ARMSTRONG, J. S., "Forecasting with Econometric Methods: Folklore versus Fact," *J. Business*, 51 (1978), 549–564.

———, "What to Ask about Managements' Forecasts," *Directors and Boards*, 6 (1981), 20–26.

———, *Long Range Forecasting*, John Wiley, New York, 1985.

———, "Forecasting Methods for Conflict Situations," in Wright, G. and Ayton, P. (Eds.), *Judgemental Forecasting*, Wiley, Chichester, 1987.

ASHTON, A. H. AND R. H. ASHTON, "Aggregating Subjective Forecasts: Some Empirical Results," *Management Sci.*, 31 (1985), 1499–1508.

BALTHASAR, H. U., R. A. A. BOSCHI AND M. M. MENKE, "Calling the shots in R and D," *Harvard Business Rev.*, (May–June 1978), 151–160.

BARCLAY, S. ET AL., *Handbook for Decision Analysis*, Decisions and Designs Inc., Suite 600, 8400 Westpark Drive, McLean, VA 22101, USA, 1977.

BASU, S. AND R. G. SCHROEDER, "Incorporating judgements in sales forecasts: Application of the Delphi method at American Hoist and Derrick," *Interfaces*, 7 (1977), 18–27.

BATES, J. M. AND C. W. J. GRANGER, "The Combination of Forecasts," *Oper. Res. Quart.*, 20 (1969), 451–468.

BEACH, L. R., J. CHRISTENSEN-SZALANSKI AND V. BARNES, "Assessing Human Judgement: Has It Been Done, Can It Be Done, Should It Be Done?" in Wright, G. and Ayton, P. (Eds.), *Judgemental Forecasting*, Wiley, Chichester, 1987.

BELSLEY, D. A., "Modelling and Forecasting Reliability," *Internat. J. Forecasting*, 4, 3 (1988), 427–447.

———, E. KUH AND R. E. WELSH, *Regression Diagnostics*, Wiley, New York, 1980.

BESSLER, D. W., "Aggregated Personalistic Beliefs on Yields of Selected Crops Estimated using ARIMA Processes," *Amer. J. Agricultural Economics*, 62, 4 (1980), 666–674.

——— AND P. J. CHAMBERLAIN, "On Bayesian Composite Forecasting," *Omega*, 15 (1987), 43–48.

BISCHOFF, C., "The Combination of Macroeconomic Forecasts," *J. Forecasting*, 8, 3 (1989), 293–314.

BLATTBERG, R. C. AND S. J. HOCH, "Database Models and Managerial Intuition: 50% Model and 50% Manager," Report from the Center for Decision Research, Graduate School of Business, University of Chicago, (May 1989).

BOPP, A. E., "On Combining Forecasts: Some Extensions and Results," *Management Sci.*, 31 (1985), 1492–1498.

BOWMAN, E. H., "Consistency and Optimality in Managerial Decision-Making," *Management Sci.*, 9 (1963), 310–321.

BROWN, L. D., "Comparing Judgemental to Extrapolative Forecasts: It's Time to Ask Why and When," *Internat. J. Forecasting*, 4, 2 (1988), 171–173.

——— AND M. ROZEFF, "The Superiority of Analysts Forecasts on Measures of Expectation: Evidence from Earnings," *J. Finance*, 43 (1978), 1–16.

BUNN, D. W., "A Bayesian Approach to the Linear Combination of Forecasts," *Oper. Res. Quart.*, 26 (1975), 325–329.

———, "Comparative Evaluation of the Minimum Variance and Outperformance Methods for the Linear Combination of Forecasts," *Oper. Res. Quart.*, 28 (1977), 653–663.

———, "Statistical Efficiency in the Linear Combination of Forecasts," *Internat. J. Forecasting*, 1 (1985), 81–193.

———, "Expert Use of Forecasts: Bootstrapping and Linear Models," in *Judgemental Forecasting*, Wright, G. and Ayton, P. (Eds.), Wiley, Chichester, 1987.

——— AND J. P. SEIGAL, "Forecasting the Effects of Television Programming Upon Electricity Loads," *J. Oper. Res. Soc.*, 34 (1983), 17–21.

CARBONE, R., A. ANDERSON, Y. CORRIVEAU AND P. P. CORSON, "Comparing for Different Time Series Methods the Value of Technical Expertise, Individualized Analysis, and Judgemental Adjustment," *Management Sci.*, 79 (1983), 559–566.

——— AND W. L. GORR, "Accuracy of Judgemental Forecasting of Time Series," *Decision Sci.*, 16 (1985), 153–160.

CASEY, C. AND T. I. SELLING, "The Effect of Task Predictability and Prior Probability Disclosure on Judgement Quality and Confidence," *Accounting Rev.*, 61 (1986), 302–317.

CERULLO, M. J. AND A. AVILA, "Sales Forecasting Practices: A Survey," *Managerial Planning*, 24 (1975), 33–39.

CHALOS, P., "The Superior Performance of Loan Review Committee." *J. of Commercial Bank Lending*, 68 (1985), 60–66.

CHOLETTE, P., "Prior Information and ARIMA Forecasting," *J. Forecasting*, 1 (1982), 375–384.

CHOW, T. AND R. M. OLIVER, "Predicting Nuclear Accidents," *J. Forecasting*, 7, 1 (1988), 49–62.

CHRISTENSEN-SZALANSKI, J. J. J., "Improving the Practical Utility of Judgment Research," in Brehmer, B., Jungerman, H., Lowens, P., and Sevon, G. (Eds.), *New Directions in Research on Decision Making*, North-Holland, New York, 1986.

CLEMEN, R. T. AND R. L. WINKLER, "Calibrating and Combining Precipitation Probability Forecasts," in Viertl, R. (Ed.), *Probability and Bayesian Statistics*, Plenum, New York, 1987.

CONNOLLY, T. AND A. L. PORTER, "Discretionary Databases in Forecasting," *J. Forecasting*, 9, 1 (1990).

CORKER, R. J., S. HOLLY AND R. G. ELLIS, "Uncertainty and Forecast Precision," *Internat. J. Forecasting*, 2, 1 (1986), 53–70.

DALRYMPLE, D. J., "Sales Forecasting Practices," *Internat. J. Forecasting*, 3, 3 (1988), 379–391.

DAWES, R. M., "Graduate Admission Variables and Future Success," *Science*, 187 (1975), 721–743.

———, "The Robust Beauty of Improper Linear Models," *Amer. Psychologist*, 34 (1979), 571–582.

——— AND B. CORRIGAN, "Linear Models in Decision-Making," *Psychological Bulletin*, 81 (1974), 95–106.

———, D. FAUST AND P. MEEHL, "Clinical versus Actuarial Judgement," *Science*, 243 (1989), 1668–1673.

DENNIS, R. L., "Forecasts and Mediation: Colorado and the Clean Air Act," *Internat. J. Forecasting*, 1, 3 (1985), 297–308.

DIEBOLD, F. AND P. PAULY, "Structural Change and the Combination of Forecasts," *J. Forecasting*, 6, 1 (1987), 21–40.

EBERT, R. J. AND T. E. KRUSE, "Bootstrapping the Security Analyst," *J. Appl. Psychology*, 63 (1978), 110–119.

EDMUNDSON, R. H., "Decomposition: A Strategy for Judgemental Forecasts," *J. Forecasting*, 9, 4, (1990), 305–314.

———, M. LAWRENCE AND M. O'CONNOR, "The Use of Non-Time-Series Information in Sales Forecasting: A Case Study," *J. Forecasting*, 7, 3 (1988), 201–212.

EINHORN, H. J. AND R. HOGARTH, "Overconfidence in Judgment: Persistence of the Illusion of Validity," *Psychological Rev.*, 85 (1978), 395–476.

FILDES, R., "Efficient Use of Information in the Formation of Subjective Industry Forecasts," Working Paper, Manchester Business School, England, (1989).

FISCHHOFF, B., "Judgemental Aspects of Forecasting: Needs and Possible Trends," *Internat. J. Forecasting*, 4 (1988), 331–339.

——— AND D. MCGREGOR, "Subjective Confidence in Forecasts," *J. Forecasting*, 1, 2 (1982), 155–172.

GEISTAUTS, G. A. AND T. G. EISCHENBACH, "Bridging the Gap between Forecasting and Action," in Wright, G., and Ayton, P. (Eds.), *Judgmental Forecasting*, Wiley, Chichester, 1987.

GLANTZ, M. H., "Consequences and Responsibilities in Drought Forecasting," *Water Resources Res.*, 18 (1977), 3–13.

GLENDINNING, R., "Economic Forecasting," *Management Accounting*, 11 (1975), 409–411.

GOLDBERG, L. R., "Diagnosticians versus Diagnostic Signs: The Diagnosis of Psychosis versus Neurosis from the MMPI," *Psychological Monographs*, 79 (1965), 602–643.

GRANGER, C. W. J. AND R. RAMANATHAN, "Improved Methods of Combining Forecasts," *J. Forecasting*, 3 (1984), 197-204.

GUERTS, M. D. AND J. P. KELLY, "Forecasting retail sales using alternative models," *Internat. J. Forecasting*, 2 (1986), 261-272.

GUPTA, S. AND P. C. WILTON, "Combination of Forecasts: An Extension," *Management Sci.*, 33 (1987), 356-372.

HARRISON, P. J. AND C. F. STEVENS, "Bayesian Forecasting," *J. Roy. Statist. Soc. Ser. B*, 38 (1976), 205-247.

HARVEY, A., "A Unified View of Statistical Forecasting," *J. Forecasting*, 3, 3 (1984), 245-276.

——— AND J. DURBIN, "The Effect of Seat-belt Legislation on British Road Casualties: A Case Study in Structural Time-Series Modelling," *J. Roy. Statist. Soc. Ser. A*, 149 (1986), 187-227.

HOERL, A. AND H. K. FALLIN, "Reliability of subjective evaluation in a high incentive situation," *J. Roy. Statist. Soc.*, 137 (1974), 227-230.

HOGARTH, R. M. AND S. MAKRIDAKIS, "Forecasting and Planning: An Evaluation," *Management Sci.*, 227 (1981), 115-138.

HOWARD, R. AND J. MATHESON, "Principles and Applications of Decision Analysis," Strategic Decisions Group, Menlo Park, CA, (1984).

HUSS, W. R., "Comparative Analysis of Company Forecasts and Advanced Time-Series Techniques Using Annual Electric Utility Energy Sales Data," *Internat. J. Forecasting*, 1, 3 (1985), 217-239.

IRLAND, L. C., C. S. COLGON AND C. T. LAWTON, "Forecasting a State's Economy: Maine's Experience," *The Northeast J. Business*, 11 (1984), 7-19.

JENKS, J. M., "Non-Computer Forecasts to Use Right Now," *Business Marketing*, 68 (1983), 82-84.

JOHNSON, E. J., "Expertise and Decision under Uncertainty: Performance and Process," in Chi, M. T. H., Glaser, R. and Farr, M. J. (Eds.), *The Nature of Expertise*, Erlbaum, Hillsdale, NJ, 1988.

KABUS, I., "You Can Bank on Uncertainty," *Harvard Business Rev.* (May-June 1976), 95-105.

KANG, H., "Unstable Weights in the Combination of Forecasts," *Management Sci.*, 32 (1986), 683-695.

KEEN, P. G. AND T. SCOTT-MORTON, *Decision Support Systems: An Organisational Perspective*, Addison-Wesley, Reading, MA, 1978.

KLEIN, H. E. AND R. E. LINNEMAN, "Environmental Assessment: An International Study of Corporate Practice," *J. Business Strategy*, 5 (1984), 66-84.

KUNREUTHER, H., "Extensions of Bowman's Theory on Managerial Decision-Making," *Management Sci.*, 15 (1969), 415-439.

LAWRENCE, M. J., "The Exploration of Some Practical Issues in the Use of Quantitative Forecasting," *J. Forecasting*, 2 (1983), 169-179.

———, R. H. EDMUNDSON AND M. J. O'CONNOR, "An Examination of the Accuracy of Judgemental Extrapolation of Time Series," *Internat. J. Forecasting*, (May 1985), 14-25.

———, ——— AND ———, "The Accuracy of Combining Judgemental and Statistical Forecasts," *Management Sci.*, 32 (1986), 1521-1532.

LAWSON, R. W., "Traffic Usage Forecasting; Is It an Art or a Science?" *Telephony*, (February 1981), 19-24.

LIBBY, R., "Accounting Ratios and the Prediction of Failure: Some Behavioral Evidence," *J. Accounting Res.*, (Spring 1975), 150-161.

LICHTENSTEIN, S., B. FISCHOFF AND L. D. PHILLIPS, "Calibration of Probabilities: The State of the Art," in Jungermann, H. and de Zeeuw, G. (Eds.), *Decision Making and Change in Human Affairs*, D. Reidel, Amsterdam, 1977.

———, ——— AND ———, "Calibration of Probabilities: The State of the Art to 1980," in Kahreman, D., Slovic, P. and Tversky, A. (Eds.), *Judgement under uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, 1982.

LINDLEY, D. V., *Making Decisions*, Wiley, Chichester, 1971.

LOCK, A., "Integrating Group Judgements in Subjective Forecasts," in Wright, G. and Ayton, P. (Eds.), *Judgemental Forecasting*, Wiley, Chichester, 1987.

LOPES, L. L., "Procedural Debiasing," *Acta Psychologica*, 64 (1987), 167-185.

MABERT, V., "Statistical versus Sales-Force Opinion Short-Range Forecasts," *Decision Sci.*, 7 (1976), 310-318.

MAKRIDAKIS, S., "Metaforecasting," *Internat. J. Forecasting*, 4, 3 (1988), 467-491.

———, A. ANDERSON, R. CARBONE, R. FILDES, M. HIBON, R. LEWANDOWSKI, J. NEWTON, E. PANZEN AND R. WINKLER, "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition," *J. Forecasting*, 1 (1982), 11-153.

MALLEY, D. D., "Lawrence Klein and His Forecasting Machine," *Fortune*, (March 1975), 152-157.

MATHEWS, B. P. AND A. DIAMANTOPOULOS, "Judgemental Revision of Sales Forecasts: A Longitudinal Extension," *J. Forecasting*, 8, 2 (1989), 129-140.

McNees, S. K. and N. S. Perna, "Forecasting Macroeconomic Variables: An Eclectic Approach," *New England Economic J.*, (May/June 1981), 15-30.

Meehl, P. E., "When Shall We Use Our Heads Instead of the Formula?," *J. Counseling Psychology*, 4 (1957), 268-273.

———, "A Comparison of Clinicians with Five Statistical Methods of Identifying Psychotic MMPI Profiles," *J. Counselling Psychology*, 6 (1959), 102-122.

Morecroft, J., "Strategy Support Models," *Strategic Management J.*, 5, 3 (1984), 715-729.

Moskowitz, H. and R. Sarin, "Improving Consistency of Conditional Probability Estimates for Forecasting and Decision Making," *Management Sci.*, 29 (1983), 735-749.

Murphy, A. H. and B. G. Brown, "A Comparative Evaluation of Objective and Subjective Weather Forecasts in the United States," in Wright, G. (Ed.), *Behavioural Decision Making*, Plenum, New York, 1985.

Newbold, P. and C. W. J. Granger, "Experience with Forecasting Univariate Time Series and the Combination of Forecasts," *J. Roy. Statist. Soc. Ser. A*, 137 (1974), 131-165.

Nilsson, G., "The Credibility of Inflation-related Scenarios of Different Lengths," in H. Montgomery and O. Svenson (Eds.), *Process and Structure in Human Decision Making*, John Wiley, Chichester, 1989.

O'Connor, M., "Models of Human Behaviour and Confidence in Judgement," *Internat. J. Forecasting*, (1989) 159-170.

Pankoff, L. D. and H. V. Roberts, "Bayesian synthesis of clinical and statistical prediction," *Psychological Bulletin*, 70 (1968), 762-773.

Pankratz, A., "Time Series Forecasts and Extra-Model Information," *J. Forecasting*, 8, 2 (1989), 75-84.

Pepper, M. P. G., "Comment on a Paper by Chatfield and Prothero," *J. Roy. Statist. Soc. Ser. A*, 136 (1973), 326-329.

Pereira, B., R. Coqueiro and A. Perrota, "Experience in Combining Subjective and Quantitative Forecasts of Open Market Rates," *J. Forecasting*, 8, 3 (1989), 331-342.

Phillips, L., "Requisite Decision Modelling," *J. Oper. Res. Soc.*, 33 (1982), 303-312.

———, "On the Adequacy of Judgmental Forecasts," in Wright, G., and Ayton, P. (Eds.), *Judgmental Forecasting*, Wiley, Chichester, 1987.

——— and W. Edwards, "Conservatism in a Simple Probability Inference Task," *J. Experimental Psychology*, 72 (1966), 346-357.

Pitz, G. F., "Decision Making and Cognition," in Jungerman, H., and de Zeew, G. (Eds.), *Decision Making and Change in Human Affairs*, D. Reidel, Amsterdam, 1977.

Raiffa, H. and R. Schlaifer, *Applied Statistical Decision Theory*, Harvard University Press, Cambridge, MA, 1961.

Reeves, G. R. and K. D. Lawrence, "Combining Multiple Forecasts Given Multiple Objectives," *J. Forecasting*, 1 (1983), 271-280.

Reinmuth, J. E. and M. D. Guerts, "A Bayesian Approach to Forecasting Efforts of Atypical Situations," *J. Marketing Res.*, (August 1972).

Rosenhead, J. (Ed.), *Rational Analysis for a Problematic World*, John Wiley, Chichester, 1989.

Rothe, J. T., "Effectiveness of Sales Forecasting Methods," *Industrial Marketing Management*, (April 1978), 114-118.

Schnaars, S. P. and I. Hohr, "The Accuracy of *Business Week's* Industry Outlook Survey," *Interfaces*, 18 (1988), 31-38.

Shepanski, A., "Tests of Theories of Information Processing Behaviour in Credit Judgement," *Accounting Rev.*, 58 (1983), 581-599.

Simon, H. A., *The New Science of Management Decision*, Harper & Row, New York, 1960.

Soergel, R. F., "Probing the Past for the Future," *Sales and Marketing Management*, 130 (1983), 39-43.

Turner, D., "The Role of Judgement in Macroeconomic Forecasting," *J. Forecasting*, 10 (1990).

Tversky, A. and D. Kahneman, "Belief in the Law of Small Numbers," *Psychological Bulletin*, 76 (1971), 105-110.

——— and ———, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185 (1974), 1124-1131.

Wallis, K. F. et al., *Models of the UK Economy: Reviews 1-5*, Oxford University Press, Oxford, 1984-88.

Wallsten, T. S. and D. V. Budescu, "Encoding Subjective Probabilities: A Psychological and Psychometric Review," *Management Sci.*, 29 (1983), 151-173.

West, M. and J. Harrison, "Subjective Intervention in Formal Models," *J. Forecasting*, 8, 1 (1989), 33-54.

Whitred, G. and I. Zimmer, "The Implications of Distress Prediction Models for Corporate Lending," *Accounting and Finance*, 25 (1985), 1-13.

Willems, E. P., "An Ecological Orientation in Psychology," *Merrill-Palmer Quart. Behavior and Development*, 11 (1965), 317-343.

Winkler, R. L. and S. Makridakis, "The Combination of Forecasts," *J. Roy. Statist. Soc. Ser. A*, 146 (1983), 150-157.

——— AND A. M. MURPHY, "Experiments in the Laboratory and the Real-World," *Organisational Behaviour and Human Performance*, 10 (1973), 252–270.

WOLFE, C. AND B. FLORES, "Judgemental Adjustment of Earnings Forecasts," *J. Forecasting*, 9, 4, (1990), 389–406.

WRIGHT, G., "Changes in the Realism and Distribution of Probability Assessment as a Function of Question Type," *Acta Psychologica*, 52 (1982), 165–174.

——— AND P. AYTON, "Subjective Confidence in Forecasts: A Reply to Fischhoff and McGregor," *J. Forecasting*, 5 (1986), 117–123.

——— AND ———, "Tasks Influences on Judgemental Forecasting," *Scandinavian J. Psychology*, 28 (1987), 115–127.

——— AND ———, "Immediate and Short-Term Judgemental Forecasting: Personologism, Situationism, or Interactionism?," *Personality and Individual Differences*, 9 (1988), 109–120.

——— AND ———, "Judgmental Probability Forecasts for Personal and Impersonal Events," *Internat. J. Forecasting*, (1989).

———, ——— AND P. WHALLEY, "Eliciting and Modelling Expert Knowledge," *Decision Support Systems*, 3 (1987), 13–26.

——— AND A. WISUDHA, "Distribution of Probability Assessments for Almanac and Future Event Questions," *Scandinavian J. Psychology*, 23 (1982), 219–224.

WU, L. S-Y., N. RAVISHANKER AND J. R. M. HOSKING, "Forecasting for Business Planning: A Case Study of IBM Product Sales," *J. Forecasting*, 11 (1991), in press.

ZELLNER, A., *An Introduction to Bayesian Inference in Econometrics*, John Wiley, New York, 1971.

ZIMMER, I., "A Lens Study of the Prediction of Corporate Failure by Bank Loan Officers," *J. Accounting Res.*, (Autumn 1980), 629–636.