ELSEVIER

International Journal of Forecasting 12 (1996) 91–118

international journal of forecasting

Judgemental and statistical time series forecasting: a review of the literature

Richard Webby*, Marcus O'Connor

School of Information Systems, University of New South Wales, Kensington, N.S.W. 2052, Australia

Abstract

This paper reviews the literature on the contributions of judgemental methods to the forecasting process. Using a contingent approach, it first reviews the empirical studies comparing the performance of judgemental and statistical methods and finds emphasis for the importance of judgement in providing contextual information for the final forecasts. It then examines four methods of integrating contextual information with the output of statistical models. Although judgemental adjustment of statistical forecasts is a viable alternative, simple combination of forecasts may offer superior benefits. Promising developments can also be gained from the use of decomposition principles in the integration process.

Keywords: Time series forecasting; Judgemental forecasting; Statistical forecasting; Forecast combination; Judgemental adjustment; Judgemental decomposition

1. Introduction

In the past decade, opinion has been divided as to the role of judgement in time series forecasting. Some (e.g. Makridakis, 1988) suggested that judgement cannot be trusted. Others (e.g. Dalrymple, 1987) provided evidence that judgement was strongly preferred as the most important method of practical sales forecasting. In addition, studies of judgemental forecast accuracy have shown the credibility of simple eyeballing (e.g. Lawrence et al., 1985) and the need for judgement in situations involving contextual information (e.g. Edmundson et al., 1988).

This paper seeks to focus the debate by reviewing the empirical evidence regarding the role of judgement in time series forecasting. We examine studies that compared judgemental and statistical forecast accuracy, as well as studies where judgement was integrated with statistical or structured forecasting methods. To do this in a unified manner, we develop a contingent framework within which studies can be compared and contrasted along a number of task, human and environmental dimensions.

Section 2 presents the studies that have compared judgemental and statistical forecast accuracy. We begin by establishing the practical validity of this comparison (i.e. we ascertain that both judgement and statistical methods are used in practice) and then proceed with a factor-by-factor review of the literature along the contingent dimensions referred to above.

* Corresponding author.

SSDI 0169-2070(95)00644-3

Section 3 then extends the results from Section 2 by examining statistics and judgement as complementary approaches rather than competing approaches. We suggest that statistical (objective) and judgemental (subjective) methods could perhaps be synthesised to gain the dual benefits of objective mechanical precision and subjective human interpretative abilities. Four main approaches to integrating subjective and objective forecasting are classified: model specification, forecast combination, judgemental adjustment and judgemental decomposition. The literature is reviewed within those four categories and the pros and cons of the approaches are debated with particular reference to their accuracy, efficiency, practicality and validity.

2. Subjective versus objective forecasting methods

This section begins by briefly discussing the survey research documenting the practical use of subjective forecasting methods compared to objective forecasting methods. The aim of this exercise is to confirm that both subjective and objective methods are prevalent in practice, and hence establish the validity of their empirical comparison. The section proceeds by using a contingent methodology to assess the accuracy of subjective versus objective time series forecasting in the empirical literature. This evaluation attempts to determine whether the effectiveness of subjective versus objective methods is dependent upon task, human and environmental differences (such as the availability or otherwise of contextual information).

2.1. Practical use of subjective versus objective methods

Many surveys have examined the relative use of subjective and objective methods in practice (Cerullo and Avila, 1975; Lawrence, 1983; Mentzer and Cox, 1984; Sparkes and McHugh, 1984; Dalrymple, 1987; Taranto, 1989; Batchelor and Dua, 1990; Sanders and Manrodt, 1994). The evidence documented in those surveys indicates a remarkedly consistent finding–that subjective techniques generally represent about 40–50% of the total techniques used in time series forecasting. When comparing early and recent surveys, there appears to have been little diminution in the role of judgement, contrary to what one might expect given the emergence of desk-top computing and forecasting software. In considering this evidence, it is important to note that the majority of the respondents in those surveys were either forecasting experts or came from large organisations. Given that subjective methods are used more frequently in small firms (Dalrymple, 1987; Taranto, 1989), it seems the general prevalence of judgement in forecasting may even be somewhat greater than indicated by the above surveys.

It seems therefore that subjective and objective methods are *both* commonly used in practice, and we can proceed with a comparison of the empirical accuracy of subjective and objective forecasting methods, knowing that it is practically relevant as well as academically interesting.

2.2. Contingent analysis methodology

The evaluation of the empirical studies comparing the accuracy of subjective and objective methods of time series forecasting uses a contingent framework. The strategy adopted is to attempt to provide a comprehensive multi-factor review of the literature, rather than focus on a particular factor in isolation, since comparisons made across studies cannot control for multiple factors. Although the provision of contextual information is considered to be an important factor, it would be deficient to analyse its effects across a variety of empirical studies without considering the effects of other mitigating factors.

The factors considered are as follows. First, the effects of time series characteristics and presentation on the relative accuracy of the human judge are examined. Characteristics of the time *series* include the *periodicity* of the data, *trend*, *seasonality*, *noise*, *instability*, number of

historicals and number of *forecasts* required. Aspects of the presentation of the time series include whether the series were presented in *graphical* or tabular form and the nature of the *feedback* provided to the judges. Second, the effects of subject/environment characteristics are considered. The situational characteristics of the *judges* that are deemed important to the task include the *experience* of the forecasters, the extent to which the judges made the forecasts with an understanding of the *context* or nature of the series and an assessment of the *motivation* of the judges in the task.

The following sections discuss the empirical evidence for each of those contingent factors. To supplement and support this discussion, a set of reference tables is included in an Appendix to the paper, which provides further detail about the studies. Details of the empirical studies discussed in the following sections can be found in the reference tables.

2.3. Empirical evidence

2.3.1. Series/task characteristics

2.3.1.1. Trend

In extrapolating trends, the human judge may revert to the mean, continue or enhance the trend, or apply an intermediate strategy and dampen the trend. The tendency to continue a trend may be enhanced by its length and consistency (Andreassen and Kraus, 1990) and by the presence of confirming news (De Bondt, 1990). Empirical work on line fitting by Mosteller et al. (1981) indicates that judgemental fitting is highly accurate with both upward and downward series, compared with linear regression. In that study, subjects may have tended to overstate slopes slightly, but this was not significant, and noise had no influence on trend identification. Judgemental modelling of trend in *real* series is also highly accurate (Edmundson, 1987; Winton and Edmundson, 1993), and it seems that judges may dampen these trends. Studies of the extrapolation of exponential growth (Wagenaar and Sagaria, 1975; Timmers and Wagenaar, 1977) have shown that people greatly underestimate exponential processes (although note the criticism of their normative model by Jones, 1984), but exponential growth is rare in economic series because many are affected by multiple causal forces (see Collopy and Armstrong, 1993).

Thus, based on the related literature, judgement can be expected to be reasonably accurate with trending series in comparison to statistical methods. However, the outcome of the comparison will be dependent on the type of statistical technique used to model trend–some techniques are highly accurate, others are likely to be particularly weak in handling trend.

Studies of subjective versus objective time series forecasting, as opposed to line fitting, show that there is a distinct tendency for subjects to dampen upward and downward trends (Eggleton, 1982; Lawrence and Makridakis, 1989; Sanders, 1992). Lawrence and Makridakis (1989) found that forecasts of downward series were dampened significantly more than forecasts of upward series. O'Connor (1993) also observed the forecasting accuracy of downward sloping series was significantly inferior to upward sloping series. Sanders (1992) showed that judgemental forecasts of upward sloping series were relatively less accurate than forecasts of flat series, although statistical forecasts of trending series were also worse.

Thus, there may be a weak association between the presence of trend and comparative judgemental inaccuracy, but all the above studies were conducted using artificial or manipulated data, and may not apply in real life situations which often require trends to be dampened (Gardner and McKenzie, 1985). The assumption of constancy made by researchers may not be valid, and the supposed 'bias' of trend dampening may actually be appropriate behaviour.

2.3.1.2. Seasonality

A substantial body of research suggests that humans possess good pattern recognition skills, in fields such as music, object perception and speech recognition (e.g. Simon and Sumner, 1968). A time series pattern is arguably much less complex than the patterns in those tasks, but

nevertheless Edmundson (1987) found that humans could perform as well as automatic mechanisms in seasonal identification.

The empirical evidence suggests that, in studies involving artificial series, judgemental performance deteriorates with high seasonality (Adam and Ebert, 1976; Sanders, 1992), and high signal (Eggleton, 1982; Lawrence and O'Connor, 1992). However, the accuracy of the statistical techniques also declines, and it is difficult to tell from these studies whether or not judgement is comparatively worse when the series contain high seasonality.

Lawrence et al. (1985) showed that, with real time series, judgemental forecasts of *seasonal* series were *more* accurate than judgemental forecasts of non-seasonal series. However, there may have been differences other than seasonality involved, e.g. trend, noise and instability. The study compared judgemental forecasts with statistical forecasts, finding that the best statistical method was generally more accurate than judgement, except in the case of tabular seasonal series forecasted by the authors.

Edmundson (1987) conducted a discriminant analysis in order to determine the time series characteristics that allow us to differentiate between judgemental and statistical forecasts in monthly series. He found that a purely seasonal metric was not statistically important, but a 'signal to noise' metric was significant in discriminating between judgemental and deseasonalised single exponential smoothing (DSE) forecasts over short (1–6 months) horizons. Edmundson concluded that judgement had relatively greater difficulty with a strong stable signal, counter to expectation. However, the 'signal to noise' metric was not a good discriminant between judgemental forecasts and Box–Jenkins forecasts. This implies that this metric may be associated more with DSE accuracy than judgemental *in*accuracy. In other words, it may simply identify series where DSE performance differs, and hence those series where there is room for judgement to be better.

Thus, the presence of seasonality is associated with higher forecast error for judgemental and statistical methods. It is relatively difficult to determine whether, with different levels of seasonality, subjective methods are comparatively better than objective methods.

2.3.1.3. Noise

Noise refers to the random or white noise in a series, as distinct from instability which is related to the presence of discontinuities. The findings from studies in related fields suggest that high random noise leads to comparatively poor human performance (Beach and Peterson, 1967; Laestadius, 1970).

The evidence in the time series forecasting literature tends to confirm that random noise has a decremental effect on the forecast accuracy of judgemental techniques (Adam and Ebert, 1976; Eggleton, 1982; Sanders, 1992; O'Connor et al., 1993). However, statistical methods also fare poorly. Determining which method is best for series with different noise levels is again not straightforward. Lawrence et al. (1985) showed that students with graphs were more accurate than statistical methods for 'macro' series, which would be expected to have relatively low noise, but not for 'micro' series, which would probably have high noise. However, the micro series may have also exhibited higher instability (see next section). The forecasts of the researchers themselves in that study were more accurate with both types of series when they were presented in graphical form. Perhaps one reason for the problem that novices had with noisy series was because they may have read too much signal in the random perturbations (O'Connor et al., 1993).

Thus, it is unclear whether subjective methods perform better or worse than objective methods when random noise is high.

2.3.1.4. Instability

Instability refers to the presence of discontinuities or temporal disturbances in the series. The literature on 'bootstrapping' suggests that models outperform man, provided the data are relatively stable (Kleinmuntz, 1990). However, when 'broken-leg cues'[1] are available (which are

[1] A broken-leg cue refers to an unusual important piece of information whose presence would dramatically alter judgement compared to a model of that judgement (Kleinmuntz, 1990).

not able to be incorporated into the models), human judgement is most effective. This implies that judges will outperform models when they have contextual information to help them comprehend discontinuities in series. In the absence of these explanatory cues, human pattern recognition skills may still allow the judge to notice discontinuities in the series, and hence discount them when making the forecast.

In forecasting artificial series containing discontinuities, judges generally perform *worse* than statistical methods (Sanders, 1992; O'Connor et al., 1993). Only in one case was judgement superior–when, in Sanders (1992), subjects forecast a low noise series that contained an immediate level change ('step') mid-way through the series history. However, Sanders did not test judgement against the adaptive response rate (ARR), which was found to be consistently better than single exponential smoothing (SES), and judgement, in O'Connor et al. (1993).

One problem with both these studies is that the forecasts were made in the absence of contextual information. In a realistic setting, such information should help judges identify, explain and anticipate discontinuities. For example, Sanders and Ritzman (1992) determined that expert *contextual* judgement was most effective for the highly variable series–they found a significant interaction effect between contextual knowledge and data variability. Edmundson et al. (1988) also found that for forecasts for products with sporadic promotions in supermarkets (high instability), contextual knowledge was vitally important.

Thus, it appears that judgement generally does not deal well with instability, unless contextual information is available to explain discontinuities.

2.3.1.5. Number of historical data points

Human information processing theory (Schroder et al., 1967) suggests that there is an inverted bell-shaped relationship between information load and decision quality. The empirical work (e.g. Miller, 1956; Phelps and Shanteau, 1978) on information overload suggests that anywhere between three and ten pieces of information (cues) represent the limits of human information processing capacities. The source of this variation lies not only in task and subject differences, but also in the different definitions of what constitutes a 'cue' (see Wood, 1986, for a definition). For example, a time series may be considered to be a single 'configural' cue (Bruner et al., 1956), a set of cues indicative of its observable characteristics, or it may contain cues for every historical data point.

There is some indication that as little as 40 data points may lead to overload (Lawrence and O'Connor, 1992), although this result was interpreted more as a failure in the assumption of constancy and hence its implication is unclear. It has been suggested that complex statistical techniques such as Box–Jenkins may require as many as 50 observations in the modelling process (Box and Jenkins, 1970). Parsimonious techniques require significantly fewer observations[2] (e.g. the naive requires just one historical value), but it is possible that judgement (especially *contextual* judgement) would be better than even these parsimonious methods when there is little historical data (e.g. new product forecasting, Wind et al., 1981).

2.3.1.6. Length of forecast horizon

The empirical evidence suggests that short-term forecasts are generally more accurate than long-term forecasts, in line with expectations (Lawrence et al., 1985; Brown et al., 1987; Lawrence and Makridakis, 1989; Hopwood and McKeown, 1990). This is true for both subjective and objective methods, but again it seems unclear which method has the greater advantage for different horizon lengths.

However, recent research by Lawrence et al. (1994) suggests that judgemental forecast accuracy sometimes deteriorates with shorter forecast horizons, despite the availability of more time series history and improved contextual information. This contrasts with the findings in the earnings forecasting literature that analyst timing advantage (i.e. proximity to the earnings announcement) plays an important role in the

[2] See Makridakis et al. (1983, p. 556) for guidelines as to the minimum data requirements of various objective methods.

superiority of subjective forecasts over objective methods (e.g. Brown et al., 1987).

Thus, the majority of the evidence indicates that judgement performs worse than objective methods over long forecast horizons.

2.3.1.7. Feedback

In general, feedback has been found to have a beneficial effect on task performance in most estimation tasks (O'Connor, 1989). The type of feedback also appears to have an important influence on task performance (Balzer et al., 1992). Outcome feedback may be less effective (Fischer, 1982) than task performance feedback (Benson and Önkal, 1992) in probability learning tasks (i.e. presenting an actual value may be less effective than presenting a summative error measure). In time series extrapolation, there has been relatively little work done on this issue. One study, by Mackinnon and Wearing (1991), suggested that simple outcome feedback may be just as effective. However, the deterministic nature of their series might have made outcome feedback appear more effective than it would be in a normal probabilistic forecasting task (Goodwin and Wright, 1993).

It is difficult to report on the effect of feedback on judgemental time series forecasting. None of the studies we found systematically varied the feedback dimension, and it was also difficult to determine the quality of feedback in these studies. There is a clear need for research on the effect of different types of feedback on judgemental forecasting performance.

2.3.1.8. Graphical versus tabular data presentation

There is mixed evidence in the related literature about the relative effectiveness of graphical versus tabular presentation. Some representative studies are mentioned here (see also DeSanctis, 1984). Benbasat and Dexter (1985) found no performance differences between graphical and tabular reports, while DeSanctis and Jarvenpaa (1989) concluded that the value of graphics only emerged after practice. Remus (1984, 1987) found that tables were superior to graphics in environments of low complexity, but graphics were superior with intermediate complexity (Remus, 1987). Dickson et al. (1986) also found a task-related effect. They concluded that "for a task activity that involves seeing time dependent patterns in a large amount of data, graphs are a good choice of format." (p. 46).

It is difficult to tell whether a graphical or tabular presentation mode is best in time series forecasting. Lawrence et al. (1985) found no significant difference between the accuracy of judgemental forecasts made with graphical versus tabular displays, although they suggested that forecasts made by the table method were more 'robust' (i.e. had lower standard deviation). There was also some weak support for the hypothesis that graphs were better with short horizons and tables were better in the longer term.

2.3.2. Judge/environmental characteristics

2.3.2.1. Experience

This factor involves an assessment of the experience possessed by the judges in the time series forecasting task. This is an attempt to codify the overall competence and knowledge of the human judge. There appear to be two major aspects to forecaster experience:

- *Technical knowledge* (Sanders and Ritzman, 1992), which is knowledge about data analysis and formal forecasting procedures, and
- *Causal knowledge*, which pertains to an understanding of the cause–effect relationships involved (i.e. the structure, as opposed to the content), and is typically gained from general industry forecasting experience (Edmundson et al., 1988). This is one aspect of 'non-time series information', which also includes 'product knowledge', which is explained under the heading "Context", in the next section.

The above two aspects are not analysed individually in this review because it is generally impossible to separate them in the studies (cf. Sanders and Ritzman, 1992).

Armstrong (1985) provides a review of the value of expertise (pp. 91–96). The evidence from a variety of fields (e.g. psychology, finance, medicine) generally indicates that expertise

beyond a certain minimal level has little incremental value (cf. the value of the expert weather forecaster–Murphy and Brown, 1984). Perhaps this also holds true in time series forecasting–maybe total novices (e.g. undergraduates) will be inept at the task, but subjects with a limited amount of training (e.g. graduates) will be just as good as experts, e.g. managers (Remus, 1990).

The evidence from the time series forecasting literature is, unfortunately, rather difficult to interpret because of the confounding effects of different contextual and motivational levels in the studies. It is difficult to say whether experience is related to the forecast accuracy of subjects compared with judges from this overview table. At the intra-study level, it appears that experience (industry or technical) has no effect on forecast accuracy (Edmundson et al., 1988; Sanders and Ritzman, 1992). However those two studies varied experience within a time series only task.

2.3.2.2. Context

The 'context' factor represents an assessment of the 'contextual information' available to the forecaster. Several terms have been used in the literature that are similar to the meaning intended here by 'contextual information', e.g. contextual knowledge (Sanders and Ritzman, 1992), product knowledge (Edmundson et al., 1988) and extra-model knowledge (Pankratz, 1989). The existing definitions of forecasting 'context' unfortunately do not distinguish clearly between contextual information and forecaster experience. Thus 'contextual information' is defined here as *information, other than the time series and general experience, which helps in the explanation, interpretation and anticipation of time series behaviour*. The exclusion of 'general experience' from this definition relates to the forecaster's knowledge of causal relationships among variables–this knowledge was defined as part of the experience dimension. Note that this definition of contextual information encompasses the labels used to describe series (cf. Goodwin and Wright's, 1993, definition), as labels provide some context within which to interpret time series behaviour. Whether or not this context is appropriate or valid is another issue.

The definition of contextual information used in this analysis can be related to similar notions in literatures other than time series forecasting:

- *Broken-leg cues* (Meehl, 1957). As mentioned, this term is used in the psychology and bootstrapping literatures (Kleinmuntz, 1990) to describe any dramatic changes or events which cannot be captured in a formal model. The broken-leg analogy was made to emphasise the type of unusual event that invalidates the output of a clinical model.
- *Soft information* (Mintzberg, 1975). In his studies of managerial work, Mintzberg observed that much of the information processed by managers is of an informal, verbal, qualitative or 'soft' nature. Soft information is often in the form of rumour and hearsay and is particularly cherished by managers over the 'hard' data contained in MIS reports. Managers even strive to encourage the flow of soft information to the extent of leaving their door open to invite interruptions. Mintzberg's results have been replicated and confirmed by Kurke and Aldrich (1983).

According to the definition given for 'contextual information', not all of it will consist of 'soft' information but an important portion of it will.

Thus, on the basis of the prior related literature, subjective methods are expected to perform significantly better than objective methods when contextual information is available.

Table 1, which examines when objective and subjective methods are best across different levels of contextual knowledge, provides strong support for the above expectation. It appears that subjective methods generally perform better in the presence of contextual information. However, as Appendix reference Table A1(a)–(e) indicates, it is also possible that the improved performance is related to other factors, such as experience, motivation or differences in group versus individual forecasting.

Edmundson et al. (1988), and later Sanders and Ritzman (1992), confirmed that contextual information is critical in sales forecasting, much more so than expertise. Edmundson et al.

Table 1
Observations showing best method by context

Context	Best method	
	Objective	Subjective
None	10	1
Low	2	
Moderate	2	
High		4
Very High		2

showed that forecasts made with product knowledge were significantly more accurate (around 10% in terms of MAPE) than the techniques of DSE and judgemental extrapolations made by both novices and experts. Sanders and Ritzman (1992) also concluded that "contextual knowledge is particularly important in making good judgemental forecasts, while technical knowledge has little value" (p. 39).

In conclusion, one possible advantage of contextual information is that it explains past discontinuities and anticipates future discontinuities in time series. Without it, forecasters generally do not deal competently with rapid level changes (Sanders, 1992; O'Connor et al., 1993). Both Edmundson et al. (1988) and Sanders and Ritzman (1992) found that the effectiveness of contextual information was contingent upon the instability in the series (although different motivational levels may again play a role).

2.3.2.3. Motivation

Motivation refers to the environmental characteristics that induce rigour in the application of a forecaster's selected strategy. Beach et al. (1986a) described four motivating characteristics of the forecasting environment:

- the extrinsic *benefit* of making an accurate forecast,
- whether the forecast will be *revisable* in the light of further information before accuracy is assessed,
- whether the forecast is perceived to be within the judge's area of *competence*, i.e. whether the judge's reputation is on the line, and
- the adequacy, reliability and understandability of the *information*–motivation will be higher with quality information, because there will be "no obvious excuse for inaccuracy".

There have been a number of empirical studies that generally support a link between incentives and performance. A recent study (Henry and Sniezek, 1993) related to forecasting performance found that judgements of future performance in an 'almanac questions' task were the most accurate when individuals perceived high levels of internal control and when monetary rewards were present.

High levels of motivation are expected to have a beneficial effect on human performance, particularly when contextual information is available. Unfortunately in the empirical research we reviewed, it is practically impossible to separate motivation from contextual information–the situations where contextual information is available are also the situations where managers are making the forecasts as an important part of their job.

Edmundson et al. (1988) compared the forecasts made for key versus non-key products (managers would probably be more motivated for the important, key products) but again other factors interfere in the analysis such as greater instability and perhaps more contextual information with the key products, from the increased number of supermarket promotions.

2.4. Summary

Table 2 summarises the findings made for each of the contingent factors. It shows:

- the expected relationships, based on a priori reasoning and the related literature, between high levels of each factor and judgemental accuracy in comparison to the accuracy of objective forecasting methods;
- the overall findings made for each relationship, and whether they were in agreement with the hypotheses; and
- the effect of the interactions between factors that were investigated in the studies.

In summary, therefore, contextual information appears to be the prime determinant of judgemental superiority over statistical models. When time series are unstable, contextual in-

Table 2
Summary of the comparative accuracy of subjective methods for high levels of each contingent factor

Category	Factor	Hypothesis	Finding	Agreement
Task	Trend	=	◇▾	✗
	Seasonality	◇	?	-
	Noise	▾	?	-
	Instability	=	◇▾	✗
	Historicals	▾	◇▾	✓
	Forecast horizon	▾	◇▾	✓
	Feedback	▲	?	-
	Graphical	◇	?	-
Subject/ Environment	Experience	◇	?	-
	Contextual information	▲	▲	✓
	Motivation	▲	◇	✓
Interactions	High context & high instability	▲	▲	✓
	High seasonality & low instability	◇	◇▾	✗
	Tabular presentation & high experience	?	?	-

[a] In the absence of contextual information.
▲, strong positive; ◇, weak positive; ▾, strong negative; ◇▾, weak negative.

formation is particularly advantageous, presumably because of the greater number of discontinuities that can be explained by human judgement. However, contextual information is by no means the only factor to consider, as the above table shows. Factors such as trend, instability, historicals, the length of the forecast horizon and high seasonality in the presence of low instability may all have detrimental effects on judgemental accuracy, but perhaps only in the absence of contextual information. There is a need for research to investigate the effects of different time series characteristics when contextual information is available.

The next section examines the role of judgement *in conjunction* with objective methods of forecasting, rather than *in competition* with such methods. As we have seen, judgemental approaches are useful in interpreting contextual information, yet lack the mechanical precision and rigour of statistical approaches. *A composite approach may enable the advantages of judgemental interpretation of 'soft' contextual information to be coupled with the benefits of mechanical approaches, such as defensibility, objectivity and precision.* In the next section, we explore the alternatives to integrating objective and subjective forecasting. We attempt to determine whether the effectiveness of the different approaches to integrating subjective and objective forecasting is dependent on contingent factors, such as contextual information. For example, judgemental adjustment may or may not be better than forecast combination when contextual information is high. We build upon the literature reviewed earlier in this paper by using the same contingent methodology to review the approaches to integrating subjective and objective forecasting.

3. Integrating objective and subjective forecasting

This section explores four approaches which may facilitate the interaction of judgement with structured forecasting methods: (a) model building, (b) forecast combination, (c) judgemental adjustment and (d) judgemental decomposition. The literature in each category is reviewed, and then the paper concludes with a debate of the pros and cons of the four approaches.

Fig. 1 presents a diagrammatic classification of the four approaches. It summarises how the subjective and objective processes may interact in each approach, and in so doing, provides a background for the discussion that follows. In the diagram, note that the input of *soft* 'contextual factors' has been separated from the *hard* quantitative information (represented by the time series Y_t and the causal variables X_1 to X_n).

Model building is an approach in which judgement is used to select variables, specify model structure and set parameters based on contextual factors. From this, an econometric or statistical model is constructed which uses a time series, and perhaps other causal variables, to produce the forecast (Y_{t+k}).

In the *combination* approach, an 'objective' forecast is combined with a 'subjective' forecast. The subjective forecast may have been produced from the time series only, or with the added information provided by contextual factors and

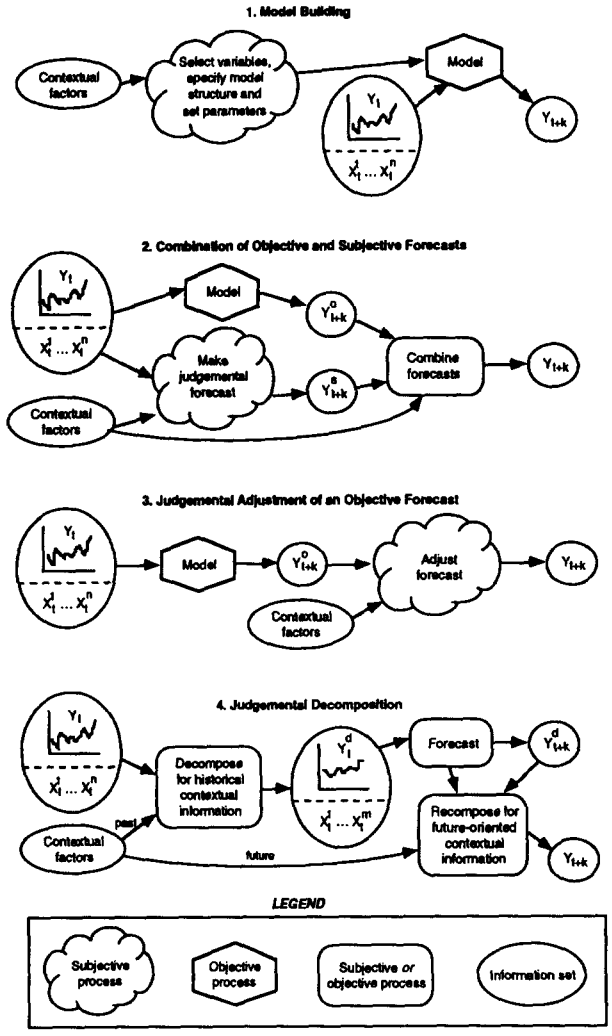

Fig. 1. Integrating subjective and objective forecasting.

other causal variables. The combination process, represented by the rounded box, may be either objective (e.g. simple averaging), or subjective (perhaps supplemented by further contextual knowledge).

Judgemental adjustment simply takes the objective forecast and adjusts it for any contextual factors to produce the forecast.

Judgemental decomposition is a three-step process; judgement and statistical methods may operate at any of these steps. First, the time series is decomposed for any historical data (soft or hard), then the decomposed time series is forecast, and the forecast is recomposed with any future-oriented information.

Note also that the four approaches are by no means mutually exclusive–they can and do interact. For example, model building is precursory to forecast combination and judgemental adjustment.

3.1. Model building

The formulation of a statistical forecast requires the input of many aspects of judgement (Bunn and Wright, 1991). At the very least, a human judge is required in the first instance to select a forecasting model (or a set of models for combination, or from which the most accurate can be selected). The purpose of this section is to identify the areas where judgement is important in model building. It does not seek to describe a step-by-step approach to model building, nor does it provide a treatise on different forms of objective models.

Bunn and Wright (1991) identified four areas where judgement may play a role in statistical model building: variable selection, model specification, parameter estimation and data analysis. These will be discussed with particular reference to the role of judgement in interpreting 'soft' contextual information.

Variable selection and data analysis. There are often many variables that may be causally related to the forecast series. Statistical diagnostics (data analysis) can help the forecaster find important causal variables, but, as Bunn and Wright state, "conventional wisdom is that variable selection should be essentially judgemental" (p. 509). The tasks of model specification and parameter estimation (described below) can be automated using techniques such as multiple regression or 'bootstrapping' (Dawes et al., 1989), but judgement is still required to identify the variables. Bootstrapping is successful when the relationships among variables are linear and when the variables are well-identified, but is vulnerable when there are unaccounted 'broken-leg' cues affecting the data (Kleinmuntz, 1990). Recent research by Lawrence and O'Connor (1992, 1993) has shown that bootstrapping the

time series forecaster was effective for stable artificial series, but not for real-world series.

Model specification. When causal information is *quantitative*, it can be incorporated fairly easily into a formal model, either through the use of regression as indicated above, or through building an econometric model. Causal variables can also be incorporated into time series analysis, using the technique of multivariate ARIMA. For example, the influence of advertising expenditure on a sales series could be accounted for by 'prewhitening'[3] the advertising series and applying this transformation to the sales series before making the forecast (Makridakis et al., 1983, Chapter 10).

Empirical research, summarised by Armstrong (1985, pp. 404–412), indicates that causal and extrapolative methods are roughly equivalent in accuracy, but that causal methods are more successful than extrapolative methods when the level of environmental change is high. This suggests that causal methods would work best for long-range forecasts. However, recent research by Clemen and Guerard (1989) indicates otherwise–the gains from econometric forecasts tend to decrease rapidly as the forecast horizon increased.

Turner (1990) and Donihue (1993) indicate that, in practice, judgemental adjustments are frequently and successfully made to econometric models to incorporate extra-model information. However, instead of adjusting a model's *output* to incorporate the extra-model information, the model itself may be changed. There appear to be two main ways of incorporating soft information into the model. First, the information may be subjectively encoded in the form of quantitative variables in a detailed econometric modelling approach (e.g. Abramson and Finizza's, 1991, application of belief networks in forecasting oil prices). However, this approach is costly, and may only be warranted for very important series. Moreover, it may still not accommodate all the 'broken-leg' cues. The second approach, model intervention (see for example, West and Harrison, 1989) allows the forecaster to *identify* 'broken-leg' cues, without needing to specify the *influence* that the cues may have had on the time series. The influence is determined within the modelling process. Harvey and Durbin (1986) illustrated how model intervention could be used to incorporate the effect of a once-only event (e.g. the introduction of special legislation) into a structural time series model.

Parameter estimation. Harvey and Durbin (1986) also showed how the parameters associated with a special event could be set so that the effect of the event diminished over time. However, setting parameters may be a cognitively difficult task. Model intervention allows judges to identify special events and it automatically assesses the effects of these events on the series, but forecasters find it difficult to encode their beliefs as to the temporal effects of these events. Parameter estimation may be difficult because it requires understanding not only the influence of the event, but also the underlying model.

The following sections examine alternative ways of introducing judgemental interpretation of contextual information into time series forecasting. These approaches may not possess the objectivity and credibility of pure model-building, yet they appear simpler and (arguably) more practical.

3.2. Combination of objective and subjective forecasts

The combination of a judgemental forecast with a statistical forecast is a pragmatic approach to introducing subjective interpretation of contextual factors. There is a great deal of evidence that combining two or more independent forecasts leads to significant improvements in accuracy (Clemen, 1989). The past literature has focused predominantly on the combination of objective forecasts, whereas this section explicitly examines the empirical evidence on combination of objective *and* subjective forecasts.

Only a few studies have investigated the influence of situational differences on the effective-

[3] Prewhitening involves the use of ARIMA techniques to remove the signal from a series, so that only white noise remains.

ness of combining objective and subjective forecasts. Lawrence et al. (1986) examined the effectiveness of combination for series having different levels of forecasting difficulty and seasonality. Contrary to expectation, they found that combination was more effective for the low MAPE series. Seasonality, on the other hand, was found to have no influence on the benefit gained from combination. The effect of forecast horizon was also somewhat surprising, as the greatest improvement in accuracy was found in the short run (cf. Conroy and Harris, 1987).

Sanders and Ritzman (1990) further investigated the effect of series difficulty, operationalising it in terms of a coefficient of variation rather than MAPE. They also found that combination was best for 'easier' series. For the 'harder' series, where judgemental forecasts were significantly more accurate than statistical forecasts, the combined forecast was less accurate than the judgemental forecast. This finding is consistent with the argument that combination would be ineffective when one forecast is *substantially* superior to another, since combining a very good forecast with a very poor forecast would be generally expected to reduce accuracy (in comparison to the very good forecast).

The evidence presented in Appendix reference Table A2(a)–(c) confirms the result of Section 2 that judgement was particularly accurate when contextual information is available. However it also indicates that combination was usually *effective* in these circumstances (cf. Conroy and Harris, 1987; Bohara et al., 1987). Hence, with context, judgemental forecasts were superior to statistical forecasts, but the combination of the two improved accuracy even further.

That finding can be related to the notion of 'conditional efficiency' (Granger and Newbold, 1973). A forecast is conditionally efficient if its error is not significantly greater than the composite forecast. In a study of the conditional efficiency of judgemental forecasts, Moriarty and Adams (1984) compared management forecasts with those based on the Box–Jenkins method. They found that the management forecasts were conditionally efficient with respect to the statistical forecasts, even though the composite forecast was slightly (but not significantly) more accurate than the management forecast. The management forecasts could not be significantly improved by combining them with statistical forecasts.

However, most of the studies show greater improvement from combination than Moriarty and Adams (1984). Although tests of conditional efficiency were not made in these studies, it appears that, in many cases, judgemental forecasters may not use information efficiently. Judges have an information advantage, but they are not fully exploiting it.

Several studies have focused on the influence of different combination schemes on accuracy. There is general support for the notion that the greater the number of forecasts, the better (Lawrence et al., 1986; Edmundson et al., 1988; Guerard and Beidleman, 1987). There is also some indication that mechanical combination is better than subjective combination (Angus-Leppan and Fatseas, 1986; Lawrence et al., 1986). There is recent evidence that a rule-based combination strategy employing judgemental assessment of 'net causal forces' can aid in generating the weights for combining objective forecasts (Collopy and Armstrong, 1992b). This method is still mechanical, but uses judgemental input based on 'eyeballing' the series. Hence, it combines objectivity and subjectivity at the combination weighting level, rather than at the level of the forecasts.

There are mixed results regarding the best type of mechanical combination. The related literature on combining objective methods suggests that simple averaging is often just as good as weighted averages based on ex ante correlations (e.g., Newbold and Granger, 1974). Some studies of the combination of objective and subjective forecasts support this finding (Conroy and Harris, 1987; Bohara et al., 1987; Blattberg and Hoch, 1990), while others indicate that regression-based weighting is more accurate (Guerard and Beidleman, 1987; Newbold et al., 1987; Lobo, 1991). The use of the contingent methodology is needed to facilitate the interpretation of these sort of mixed results, but, unfortunately, most papers do not report the necessary information about task differences. For

example, instability in series may cause problems for regression-based weighting[4]. Also somewhat disappointing among the empirical research is the lack of testing for statistical significance (cf. Blattberg and Hoch, 1990). This makes it difficult to make any conclusions with a great deal of confidence, apart from just generally stating that forecast combination improves accuracy. Clearly more contingency-based research is needed in this area, especially to determine the role of contextual information in the combination of objective and subjective forecasts.

3.3. *Judgemental adjustment of objective forecasts*

Judgemental adjustment of an objective forecast appears to be quite a common practice (see Turner, 1990, for instance). In terms of use, it seems to be the major competing alternative to combination for integrating subjective and objective forecasting. Yet, strangely, there appears to be a lack of studies comparing these two common approaches (cf. Lim, 1993, who found adjustment to be less accurate than combination, albeit in an extrapolation task without contextual information).

Based on the literature on anchoring and adjustment (Tversky and Kahneman, 1974), judgemental adjustment might be expected to have a detrimental effect on accuracy. However, the presence of contextual information, which we saw in Section 2 to be important in judgemental accuracy, may have a mediating role. With context, one would expect, a priori, that adjustment may be beneficial because the judge would bring information to the forecast that the model could not. However, without contextual information, judgemental adjustment might worsen accuracy because of the action of the anchoring and adjustment heuristic. The effectiveness of adjustment may also depend on the initial accuracy of the base statistical forecast; this may in turn be related to series characteristics (see also Section 2).

Studies of non-contextual adjustment. This form of adjustment may be attempted because the judge feels that the statistical forecast is inaccurate, even though the judge cannot associate any contextual information with the series. There is mixed evidence regarding the effectiveness of non-contextual adjustment. It appears it may be beneficial when there is 'excess error', or room for improvement, in the statistical forecast (Willemain, 1989, 1991; Lim and O'Connor, 1993), but not when the statistical forecast is accurate. Judgemental adjustment may also be more effective with seasonal series (Willemain, 1989), especially those with low noise (Sanders, 1992), but again this is related to the quality of the statistical forecast (Lim and O'Connor, 1993; Carbone et al., 1983).

Contextual adjustment. When contextual information is available, the judge may wish to adjust the model-based forecast to incorporate the effects of this extra knowledge. This type of adjustment is generally effective (cf. Mathews and Diamantopolous, 1986). However, as in Section 2, it is again difficult to attribute the accuracy improvement solely to context–experience and motivation may also play a part. There is a lack of studies which isolate these factors, as Edmundson et al. (1988) and Sanders and Ritzman (1992) did in comparing objective and subjective methods.

Structured adjustment. The practice of judgemental adjustment has been criticised because of its informal, ad hoc nature (Bunn and Wright, 1991). Four studies applied structured methods to generate judgemental adjustments. Reinmuth and Guerts (1972) used Bayesian techniques to incorporate customer survey information into the adjustment made by the marketing staff. Cook et al. (1984) and Wolfe and Flores (1990) used the analytic hierarchy process (AHP) to generate adjustments. In Wolfe and Flores' study, subjects were given contextual information in the form of a brief economic commentary, accounting information, and historical EPS for one real undisguised company. The resultant adjustments significantly

[4] Kang (1986) found, in combining objective forecasts, that simple averaging is best with unstable series, because instability tends to invalidate the ex ante weights derived from past series behaviour.

improved the accuracy of ARIMA-generated forecasts, especially when the ARIMA forecasts were initially inaccurate due to series volatility. Wolfe and Flores also examined the effect of expertise and forecast horizon, but neither affected accuracy. The external validity of this research may be limited because it appears that subjects did not see the base forecasts, or the resultant adjustments. The AHP is also quite an involved approach, although the complexity problem was partially addressed in a later study by Flores et al. (1992).

Bunn and Wright (1991) point out that even the use of a structured adjustment process is still ad hoc, lacks interaction with the model and is "mostly qualitative". They proposed the need for an "interactive decomposition structure"–interactive in the sense that judgement should be part of an overall coherent modelling effort. The topic of judgemental decomposition is examined in the next section.

3.4. Judgemental decomposition

The principle behind decomposition is the notion of 'divide and conquer' (Raiffa, 1968). Task decomposition encourages a person to concentrate selectively on one element at a time by breaking a task into its component parts.

The basic difference between judgemental adjustment and judgemental decomposition of time series is that adjustment tries to account for past and future contextual factors and model misspecifications *after* the forecast is made, while decomposition tries to remove the effects of past contextual factors from the series history *before* extrapolating and accounting for future-oriented factors in the forecast. In essence, *decomposition considers the assessment of current status to be separate from the assessment of change* (Armstrong, 1985).

This section considers the empirical studies that have investigated decomposition *in general*, and then focuses on the use of decomposition in time series forecasting.

Armstrong et al. (1975) asked subjects a set of general knowledge questions, and found that people made better judgements when using a decomposed approach. Decomposition was especially useful when the subject knew relatively little about the topic. Phelps and Shanteau (1978) found that expert livestock judges considered fewer than three pieces of information in picture form in comparison to 9–11 cues when the information was decomposed to a verbal form. They interpreted this result as implying that there are inter-correlations present in real stimuli that tend to reduce the number of dimensions processed, but it may also imply that decomposition enhanced the amount of information processed and hence decision quality (or that decomposition simply made hard-to-see cues obvious). In work related to decomposition, Fischhoff and Bar-Hillel (1984) found that focusing techniques such as isolation analysis can improve judgement in inferential problem solving. Isolation analysis requires subjects to consider what judgement they would make if each cue was the only one available, prior to making a summary judgement based on all the information.

Lyness and Cornelius (1982) compared holistic and decomposed judgement in a performance rating experiment. Students were required to evaluate hypothetical instructors based on written information. Lyness and Cornelius found that an algorithmic decomposition strategy (where the ratings were derived from dimensional evaluations and importance weights) was generally more effective than holistic judgement and clinical decomposition (where dimensions were considered separately, but the subjects then made overall evaluations). Unfortunately because of the nature of the task, their analysis considered only reliability, convergence and agreement measures; it did not compare the predicted values against actual or optimal values. Moreover, the result varied according to the measure used–with mean absolute deviation between raters, decomposition was superior; with correlations, holistic and decomposed judgement were equally effective. Lyness and Cornelius speculated that decomposition would be most advantageous in complex circumstances,

and tested the effect of information load (three, six or nine performance dimensions) on the effectiveness of decomposition. The hypothesised interaction between judgement strategy (holistic, decomposition) and information load was not found for the reliability and convergence scores. High inter-rater agreement was found with algorithmic decomposition at high information levels. However, the significance of this result was not statistically tested. Hence, the impact of decomposition on relationship between environmental complexity (information level) and decision quality has not been fully investigated.

Three recent papers by MacGregor and associates have further explored the role of decomposition in answering almanac-type questions. MacGregor et al. (1988) showed that accuracy increased with greater structure in the decision aid. Subjects performed best when provided with the full algorithmic decomposition strategy–they were given the complete algorithm, asked to estimate its components, and then shown how to combine the parts.

MacGregor and Lichtenstein (1991) showed that extending algorithmic decomposition to include alternative problem solving approaches led to further improvement in performance. However, they found that providing training on decomposition, so that people could create their own algorithms, was of little benefit. They also cautioned that "erroneous knowledge incorporated into an estimation aid may lead one astray without a warning that it is doing so" (p. 115)–subjects with *misinformation* tended to perform particularly badly.

One issue raised in the MacGregor and Lichtenstein paper was that decomposition seemed to work best with estimates of large numerical values. MacGregor and Armstrong (1993) confirmed that decomposition improved accuracy for problems involving extreme and uncertain values, but was *less* accurate for problems with target values that were not extreme and uncertain. It was not clear whether the contingency of decomposition performance was due to the large magnitude of the target values, or their uncertainty, or both. However, these papers indicate that decomposition may be a risky strategy in certain circumstances and should be applied with caution. This issue will be discussed further.

In the context of time series forecasting, Edmundson (1990) developed a decision aid, called GRAFFECT, that encouraged judgemental forecasting along the lines of classical time series analysis. Users were able to judgementally extract the trend and seasonality from a time series, leaving a residual component. Edmundson found that the extrapolative accuracy of both novice and expert forecasters using GRAFFECT were better than unaided forecasters and deseasonalised single exponential smoothing for the shorter forecast horizons. Considering that the subjects were not provided with a contextual information advantage, this seems to be a somewhat surprising finding. Why should decomposed judgement be superior to one of the most accurate statistical techniques (see Makridakis et al., 1982) in a purely numerical task? One possible explanation is that the subjects were able to recognise and discount discontinuities in the series that were naively incorporated by the statistical model. The results of O'Connor et al. (1993) and Sanders and Ritzman (1992), to a lesser extent, suggest otherwise, but their series were artificially generated and their subjects were given different instructions and lacked Edmundson's decision aid. It is, of course, open to speculation as to how well the subjects in both those studies might have performed if the events causing the discontinuities had been identified.

Collopy and Armstrong (1993) showed that decomposition of series by 'causal forces' can reduce forecast error. The effect of traffic volume was removed from three series (highway deaths, injuries and accidents) to create component series that were forecast separately and then recomposed with the forecast for traffic volume. All forecasts were made using Holt's exponential smoothing. Collopy and Armstrong found that decomposition was effective for the deaths and injuries series, but not the accidents series. They cautioned that multiplicative recomposition of decomposed components can be risky because it has the potential for explosive errors, echoing the sentiments of MacGregor and Lichtenstein (1991) and MacGregor and Armstrong (1993). They also stressed that re-

search attention should focus on the conditions where decomposition is effective.

Goodwin and Wright (1993, p. 154) suggested that decomposition would be *in*effective where:

- the decomposition is mechanical and/or the judge is sceptical about the decomposition technique that is being employed;
- there is unfamiliarity with the technique used for decomposition and the type of judgements required by it;
- the judgements required by decomposition are actually more complex psychologically than holistic judgements;
- the judge experiences boredom or fatigue because the duration of the task is lengthened and the number of judgements increased.

Goodwin and Wright (1993) also emphasised the need for decomposition research to focus on "analysing and integrating judgements based on time series information with contextual information" (p. 155). Very little work has examined judgemental decomposition of contextual factors from time series. There are software packages that facilitate additive decomposition of special events (e.g. FORSYS by Lewandowski, 1982) and multiplicative decomposition of events (e.g. FUTURCAST by Carbone and Makridakis[5]), but it seems that only anecdotal evidence (e.g. Gorr, 1986a,b) has been reported about the effect of event decomposition on forecast accuracy. Lee et al. (1990) developed a prototype expert system for forecasting the demand of oil products for a single company. The system was designed to detect historical events, relate analogous events, and to decompose events from the time series. However, the system was not empirically tested–the authors simply reported on its use in an oil company.

3.5. Summary

Four approaches to integrating subjective and objective forecasting have been examined. All have shown the potential to reduce forecast error, but they may differ in ease-of-use, credibility, cost, and other factors identified as important by practitioners (see Collopy and Armstrong, 1992a).

The incorporation of contextual factors in *model building* involves either quantifying those factors in an extensive econometric model or using the technique of model intervention. Both approaches would produce credible forecasts, but they suffer from high set-up costs. Model intervention is a complex approach, requiring an understanding of the underlying model in order to codify beliefs about contextual factors by parameter setting.

Forecast combination is a relatively practical and simple approach which generally improves accuracy. However, it suffers from duplication of effort, redundancy of information (see Clemen and Winkler, 1985), loses resolution when the dependent variable (e.g. sales) is unstable (see Moriarty, 1990), and does not facilitate backward inference[6] (see Einhorn and Hogarth, 1982). Some believe that combination should be replaced by an all 'encompassing' model (Diebold, 1989). The counter argument is that combining is more cost-effective than building a single comprehensive model (see Bunn, 1989).

Judgemental adjustment of a statistical forecast is perhaps the most easy-to-use and cost-effective of the four approaches, but it introduces the possibility of bias from the use of the anchor and adjustment heuristic. As Bunn and Wright (1991) emphasised, judgemental adjustment is ad hoc, informal and open to adversarial criticism.

Judgemental decomposition is arguably more complex than combination or adjustment, but brings structure, defensibility and backward inference to the task of forecasting with contextual information. It is possible that decomposition can reduce cognitive load, but it may also be risky and ineffective in certain circumstances (Goodwin and Wright, 1993). Research is needed on the accuracy of judgement decompo-

[5] Reported by Gorr (1986a).

[6] Backward inference refers to understanding causality from past data. For example, it may be possible to infer that a high sales value was because of an advertising campaign. This would have benefits not only in forecasting, but also in assessing the effectiveness of marketing.

sition of special events. Such an approach may facilitate the interpretation of 'soft' event information by the judgemental forecaster, and have broader implications for the integration of soft and hard information in general decision making.

4. Conclusion

This paper has reviewed the literature on the importance of human-interaction factors in a comparison of judgemental and statistical approaches to time series forecasting. It demonstrated the crucial role of the knowledge of the context of the time series to accuracy. The accuracy of judgemental eyeballing may be roughly equivalent to the statistical methods, but the major contribution of judgemental approaches lies in the ability to integrate this non-time series information into the forecasts. Accordingly, Section 3 examined four ways of integrating judgemental and statistical forecasts–model building, combination, judgemental adjustment and judgemental decomposition. Simple combination provides considerable advantages, while judgemental adjustment may be subject to bias. Promising advantages can be made from an improvement of the process of decomposition in any combination process.

Appendix. Reference tables

This appendix presents the tables that were constructed to provide the detailed support for the conclusions made in the main body of the paper. They are included here because we feel they are a useful reference source for future research in judgemental forecasting. It should be possible for a researcher doing work on the influence of seasonality, for example, to scan the table and quickly find the studies that examined seasonality.

The tables do however require some interpretive effort on behalf of the reader. A graphical presentation format was chosen to display the results succinctly and attractively. Fig. A1 shows

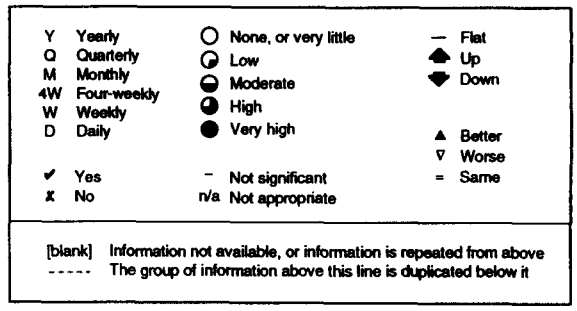

Fig. A1. Key to abbreviations and symbols used in Appendix reference tables.

a key to the symbols used in the reference tables. It is hoped that future empirical studies would provide greater information about their data, methods and environmental characteristics because in many instances it was difficult to adequately classify past studies within this contingent framework. We also hope that future literature reviews could utilise the framework and build and improve upon this approach.

A brief comment about the classification rules for certain factors:

Trend–The trend factor reports the direction of slope in the series (up, flat, down). Ideally, it would be useful to present the mean gradient (or equivalent) in the series, or at least the proportion of ups versus downs, but these measures were rarely reported in the studies.

Seasonality–The seasonality factor indicates the presence of signal or pattern in the series. This measure should be interpreted in conjunction with the data presented in the "Period" column to determine the type of seasonality (e.g. monthly versus quarterly).

Historicals–This refers to the number of historical data points used. In the studies of 'real world' forecasting, it is often impossible to determine how many data points were used by the judges, so the table presents the number of data points used by the objective methods in those cases.

Forecasts–This refers to the number of forecasts made by the human judge. This measure should be interpreted in conjunction with the 'feedback' dimension described in the next sec-

tion. If feedback was not provided, then the number represents the length of the forecast horizon, and the error measures have been calculated across the entire horizon. Where the forecast is indicated as a specific number (e.g. #16) or a range (#16–#18) then the errors represent those specific step-ahead periods.

Feedback–The assessment of feedback is a binary measure showing whether feedback was provided to subjects. Unfortunately, it was difficult to determine whether outcome or performance feedback was provided in most cases.

Experience–The following decision rules were applied: ○ Undergraduate students with no experience in forecasting. ◔ Graduate students or equivalent, who may be good surrogates for managers (Remus, 1990), but generally have little experience in forecasting. ◒ Subjects with some training and evidence of some practical experience in forecasting. ◕ Subjects with considerable experience in forecasting. ● Forecasting experts or practitioners whose job involves the regular preparation of forecasts (it is not known if they are really experts, but it is beyond the scope of the review to assess this).

Context–The following rules were used in making assessments of context: ○ No contextual information supplied; only the time series. ◔ The series was labelled (e.g. "quarterly earnings"), but subjects were unable to place the series in a broader context, through lack of detail. ◒ Incomplete level of contextual information (e.g. a production department forecasting without knowledge of the plans of the sales department). ◕ External level of contextual information, e.g. security analysts. They may possess *some* inside information and have some influence on the outcome, but this is generally limited in comparison to managers (Brown, 1988). ● Internal level of contextual informa-

Table A1(a)
Objective versus subjective methods of time series forecasting (studies ordered by amount of contextual information)

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	BEST OBJECTIVE METHOD	Error	JUDGES	# per Series	Experience	Context	Motivation	Error	Accuracy	Significance
Lawrence, Edmundson & O'Connor (1985)	M-Competition [N=111]	20	Y						6	✘	✔	Holts	12.7	Students [N=202]	1	○	○	○	14.5	∇	
		23	Q						8		✔ ✘	DSE	18.5						16.4	▲	
		68	M						18		✔ ✘	DSE	16.5						22.2	∇	
																			17.6	∇	
																			20.5	∇	
														Authors [N=3]	1	●	○	◒	13.0	∇	
																			19.7	∇	
																			24.7	∇	
																			18.3	∇	
																			15.8	▲	
		60				✔				✘	✔ ✘	DSE	15.2	Students	1	○	○	○	15.5	∇	
		51				✘					✔ ✘	Box-Jenkins	18.3						18.8	∇	
																			23.7	∇	
																			25.7	∇	
														Authors	1	●	○	◒	17.1	∇	
																			14.9	▲	
																			18.2	∇	
																			19.7	∇	
		33					Micro			✘	✔ ✘	DSE	18.0	Students	1	○	○	○	19.9	∇	
		35					Macro				✔ ✘	Box-Jenkins	13.7						24.2	∇	
																			10.9	▲	
																			14.8	∇	
														Authors	1	●	○	◒	10.8	▲	
																			17.3	∇	
																			12.5	▲	
																			14.9	∇	
		68	M						#16-#18	✘	✔ ✘	DSE	26.4	Authors	1	●	○	◒	28.7	∇	
																			21.7	▲	
Andreassen [a] (1991)	Economic and sporting	4	Y, Q, W					30	30	✔	✘	Naive	8.1	Students [N=56]	~3?	○	○	○	9.9	∇	

[a] Andreassen investigated the effect of causal prediction versus extrapolation. Subjects were significantly worse at causal prediction ($P < 0.0001$), but only his results involving time series extrapolation are presented here.

tion, e.g. managers who not only have inside information, but also significant control over the outcome of the forecast variable, including the ability to change a firm's accounting practices to make the reported outcome consistent with the forecast (Brown, 1988).

Motivation–The following decision rules were utilised: ○ No incentives, accountability or reason for accurate performance. ◔ Rewards or course credit offered to students. ◒ The judges were reasonably well motivated, as forecasting was part of their job, but they were neither accountable nor of the belief that the series were critical. ◕ Forecasting represented a major responsibility of the job and judges were accountable for accuracy. ● Judges recognised that these series were very important and were accountable for forecast accuracy.

Accuracy–In the case of Appendix reference Table A2(a)–(c) (Combination), the "Accuracy" column compares the accuracy of the resultant forecast with the error of the combined forecast with the *best* single method used. For Appendix reference Table A3(a)–(c) (Adjustment), accuracy is compared to the original (base) forecast.

Table A1(b)
Objective versus subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	BEST OBJECTIVE METHOD	Error	JUDGES	# per Series	Experience	Context	Motivation	Error	Accuracy	Significance
Lawrence & Makridakis (1989)[b]	Artificial [N=18]	6	'Y'	▲	○		○	7	#3	✗	✔	OLS regression		MBA students [N=~350]	~19	◔	○	○	-5.0[c]		} .01
		6		—															1.6		
		6		▼															8.7		
									#8										-20.4		
																			4.4		
																			32.1		
		6				◔			#3										1.2		} -
		6				◒													2.5		
		6				◕													1.7		
									#8										4.9		
																			7.6		
																			4.7		
Lawrence & O'Connor (1992)	Stationary ARMA [N=40]	20	'Q'	—	signal ◔	◔	○		4	✗	✔	General model-of-judge	0.6	Students [N=?]	1?	◔	○	○	1.4	▽	
		20			◒								1.1						2.6	▽	
		20						20											1.9		
		20						40											2.2		
Lawrence & O'Connor (1993)	M-Competition[d] [N=111]	68	M						#1	✗	✔	Model-of-judge	13.1	Students [N=?]	1?	○	○	○	11.3	▲	–
									#2				16.2						10.2	▲	.015
									#3				15.9						11.6	▲	–
									#4				15.3						11.1	▲	–
Adam & Ebert (1976)	Generated from a demand function [N=?]		'M'	▲	○		○	60	12[e]	✔	✔	SES		Grad. students [error = MAD]	20	◔	◑	○	17.7	▽	.02
					◒														18.8	▽	.001
					◕														25.1	▽	.001
						◔						Winter's 3 factor							12.4	▽	.001
						◒													28.6	▽	.001

[b] The effect of a vertical repositioning of the graph was also investigated. With a low position, judgemental performance was worse than regression ($P < 0.01$).

[c] Error measure is only the deviation from the regression forecast (i.e. judgemental estimate–least squares). Hence, effect on accuracy cannot be reported.

[d] Only the results for monthly data are presented here. The results for annual and quarterly series and between seasonal and non-seasonal series also supported the finding that the bootstrapping model was less accurate than the human judge.

[e] Their analyses were based on each forecaster's last 12 forecasts. They found that "terminal performance was significantly better than initial performance ($P < 0.001$)".

[f] Adam and Ebert also investigated the effect of telling subjects that the series represented demand for medical supplies, but found no differences in accuracy.

Table A1(c)
Objective versus subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	BEST OBJECTIVE METHOD	Error	JUDGES	# per Series	Experience	Context	Motivation	Error	Accuracy	Significance
Sanders (1992)	Artificial [N=10]	1	'M'	—	○	◔	○	48	12	✗	✔,✗	SES	3.0	Undergraduates [N=38]	~7-8	○	○	○	3.5	▽	.05
		1				◔							16.4						22.6	▽	.05
		1		▲		◔						Winters	1.5						4.0	▽	.05
		1				◔							8.7						13.9	▽	.05
		1		—	◕	◔							5.4						5.7	▽	–
		1				◔							14.8						16.8	▽	.05
		1		▲		◔							3.2						5.5	▽	.05
		1				◔							19.9						23.7	▽	.05
		1		—	○	◔	◔					SES	7.7						7.1	▲	.05
		1				◔							21.2						26.4	▽	.05
O'Connor, Remus & Griggs (1993)	Artificial [N=10]	1	'Q'	—	○	◔	○	20	28	✔	✔	SES	4.9	Graduate students [N=17]	17	◔	◔	○	7.0	▽	.05
		1				◔							18.3						21.1	▽	–
		1		▲		◔	◔					ARR	6.2						8.2	▽	.05
		1				◔							13.5						18.0	▽	–
		1				◔	●						6.4						7.8	▽	–
		1				◔							15.7						19.4	▽	.05
		1		▼		◔	◔						6.3						9.7	▽	–
		1				◔							17.5						24.1	▽	–
		1				◔	●						7.7						10.5	▽	–
		1				◔							18.7						25.3	▽	–
Carbone & Gorr (1985)	M-Competition [N=10]	2	Y					30-140	6 8 18	✗	✔	Holt-Winters [MdAPE]	3.9	MBA students [g] (7 teams of 2)	7	◔	◔	○	7.2	▽	✔[h]
		2	Q																		
		6	M																		

[g] This data represents their initial 'eyeball' extrapolation. A second group of students were involved, but they did not make an eyeball extrapolation.
[h] The statistical testing of these results has been criticised (Andreassen, 1991).

Table A1(d)
Objective versus subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	BEST OBJECTIVE METHOD	Error	JUDGES	# per Series	Experience	Context	Motivation	Error	Accuracy	Significance
Edmundson, Lawrence & O'Connor (1988)	Sales [N=18]	18	M			◔	◔	48	6			DSE	31.1	Grad. students	6	◔	○	○	35.2	▽	–
	– key	10				◔	●						28.8						36.7	▽	
	– non-key	8				●	◔						33.9						33.3	▲	
														Managers	6	●	○	○	33.3	▽	–
																			32.9	▽	
																			33.8	▲	
														Management group	1	●	●	◕	22.7	▲	.001
																		●	19.0	▲	
																		◑	27.3	▲	
Sanders & Ritzman (1992)	Demand for public warehouse	22	D			◔		[i] ~750	65	✔	✗	Average of SES, Holts, AEP	22.8	'Non-technical' students [N=32]	4-6	○	○	◕	30.6	▽	
						◔							28.2						35.6	▽	
						◔							74.3						75.2	▽	
						◔							132.7						135.3	▽	
														'Technical' students [N=49]	4-6	◕	○	◕	28.2	▽	
																			34.7	▽	
																			73.2	▲	
																			137.3	▽	
														Planning group	1	●	●	◕	27.5	▽	.05
																			28.9	▽	–
																			51.5	▲	.05
																			65.2	▲	.05
Walker & McClelland (1991)	Sales of consumer products	2	Q	●				32-40	12	✔		ARIMA	6.6	Sales dept. [j]	1	◕	◕	◕	14.4	▽	.025
														Production dept.	1				9.3	▽	—
														Finance dept.	1				10.1	▽	—
Makridakis et al. (1993)	M2-Competition	23	M						14	✔		Combination of SES, Holt, Dampen, Long	11.4	Combination of [k] experts	5	●	◕	◕	13.1	▽	

[i] Students were only provided with 3 months of daily actuals and 2 years of monthly actuals (both in table form).
[j] The forecasts by the sales department were reported to be subject to political and organisational biases.
[k] *Most*, but not all, experts were using objective methods (some judgementally adjusted) to make their forecasts. *Most* also used contextual information.

Table A1(e)
Objective versus subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	BEST OBJECTIVE METHOD	Error	JUDGES	# per Series	Experience	Context	Motivation	Error	Accuracy	Significance
Armstrong (1983) *[averages of 14 studies]*	Company earnings	1609	Y									Extrapolative	28.4	Managers and analysts			*generally high*		21.0	11 ▲ 3 ▽ 1 =	*most*
Brown, Hagerman, Griffin & Zmijewski (1987)	Company earnings	233	Q						60-83	#1 #2 #3		Brown-Rozeff Brown-Rozeff Brown-Rozeff	27.3 30.7 33.1	Analysts (Value Line)		●	◕	◕	20.7 26.3 28.7	▲ ▲ ▲	.01 .01 .01
Hopwood & McKeown (1990)	Company earnings	258	Q						68	#1 #2 #3 #4		Brown-Rozeff Brown-Rozeff Brown-Rozeff Hopwood-McKeown 2	34.0 39.1 42.1 43.8	Security analysts (Value Line)		●	◕	◕	28.0 36.7 39.8 43.4	▲ ▲ ▲ ▲	
Schnaars & Mohr (1988)	Industry-level variables (e.g. sales, profit, market share)	84	Y							1	✗	Naive	27.0	*Business Week* editors		●	●	◕	19.6	▲	

Table A2(a)
The combination of objective and subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	JUDGES	# per Series	Experience	Context	Motivation	SINGLE METHODS (ranked)	Error	COMBINATION Method	Weighting	Error	Accuracy	Significance
Lawrence, [a] Edmundson & O'Connor (1986)	M-competition	68	M						18	✗	✔,✗	Students Authors	1 1	○ ●	○	○	1. DSE 2. Graph 3. Table	16.5 17.9 18.1	1&2 1&3 1,2&3	Equal	15.9 16.2 15.6	▲ ▲ ▲	
									6									11.0 12.1 14.2			10.4 11.6 10.9	▲ ▽ ▲	
									#13-#18									24.4 25.3 23.6			23.3 22.3 22.1	▲ ▲ ▲	
		17			Low MAPE				18									4.8 5.7 6.3			4.0 5.0 4.4	▲ ▽ ▲	
		17			High MAPE													38.1 34.8 35.6			34.6 34.5 33.2	▲ ▲ ▲	
Angus-Leppan & Fatseas (1986)	Interest rates	1	M					48	12	✗	✗	Undergraduate students	123	○	○	◔	1. Judgement 2. Selected from *Foresight*	12.7 5.5 to 31.4	1&2	Subjective	18.0	▽	
Edmundson, Lawrence & O'Connor (1988)	Sales	18	M					48	6	✔		Company personnel (IND) Students (JMT)	6 6	● ◔	○ ○	○ ○	1. DSE 2. IND 3. JMT	31.1 33.3 35.2	1&2 2&3 1,2,3	Equal	31.3 29.7 29.2	▽ ▲ ▲	
Guerard & Beidleman (1987)	EPS of manufacturing firms	35	Y					58	2			Security analysts (Standard and Poor's)		●	◕	●	1. Analysts 2. Box-Jenkins 3. Avg growth [error=MSFE]	10.4 15.3 18.8	1&2 1&2 1&2 1,2&3 1,2&3	Equal OLS Ridge OLS Ridge	9.5 8.4 9.4 8.1 9.8	▲ ▲ ▲ ▲ ▲	
Guerard (1987)	EPS	261	Y					21	1			Analysts (Standard & Poor's)		●	◕	●	1. Analysts 2. Univariate models [MSFE]	0.33 [b] 0.35	1&2	Several regression techniques	0.27 (best)	▲	

[a] The effect of seasonality was also investigated, and found not to influence the benefit of combining forecasts.
[b] Errors for 1981 sample. The results were similar, yet more marked in 1982.

Table A2(b)
The combination of objective and subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	JUDGES	# per Series	Experience	Context	Motivation	SINGLE METHODS (ranked)	Error	COMBINATION Method	Weighting	Error	Accuracy	Significance
Newbold, Zumwalt & Kannan (1987)	EPS for electric utilities	65 to 73	A[c]					5	7			Analysts (Value Line) [error=MSE]		●	◕	●	1. Value Line[d] 2. Dividend growth 3. Sharpe-Linter 4. Brown-Rozeff	15.6 20.8 21.2 24.6	All All	Equal Regression	19.6 15.3	▲ ▲	
Lobo & Nair (1991)	EPS	95	A[c]					15	1			Security analysts (I/B/E/S) [Percent[e] Better]		●	◕	●	1. Analyst 2. ARIMA 'F1' 3. ARIMA 'F2' 4. ARIMA 'F3' 5. Random walk with drift	n/a	1&2 1&3 1&4 1&5	Equal	52.9 55.9 58.0 48.1	▲ ▲ ▲ ▲	
Lobo (1991)	EPS	96	A[c]					15	1			Security analysts (Value Line and I/B/E/S)		●	◕	●	1. I/B/E/S 2. Value Line 3. Random walk with drift 4. ARIMA	55.6 54.7 64.1 64.7	1&4 1&3 2&4 2&3 1,2,3,4	Linear[f] model with no constant term and no constraints	52.0 50.4 50.1 50.8 50.4	▲ ▲ ▲ ▲ ▲	
Conroy & Harris (1987)	EPS	> 1500	A					20	#1			Financial analysts from brokerage firms	1 to 12+	●	◕	●	1. Mean analyst forecast 2. Random walk 3. Best of 5 methods[g]	24.0 42.1 47.8	1&2 1&3 1&2	Equal Minimum Variance Minimum Variance	29.5 30.0 36.3	▽ ▽ ▽	
									#5									37.2 34.9 38.1			33.7 36.1 45.0	▲ ▽ ▽	

[c] Quarterly data were used ion the extrapolative models, and aggregated.
[d] Other models were used but not combined.
[e] A measure of percentage error was also presented, but had to be winsorised to remove outliers.
[f] This was the best technique for generating combination weights in the study, superior to equal weighting and other linear models.
[g] The most accurate method (based on lowest APE for last year's forecast) was selected from random walk, single exponential smoothing, double exp. smoothing, triple exp. smoothing, and average EPS over the past 5 years.

Table A2(c)
The combination of objective and subjective methods of time series forecasting

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	JUDGES	# per Series	Experience	Context	Motivation	SINGLE METHODS (ranked)	Error	COMBINATION Method	Weighting	Error	Accuracy	Significance
Bessler & Brandt (1981)	Livestock prices[i]	3	Q					60	14			Expert economists [MSE]		●	◕	●	1. Expert 2. ARIMA 3. Econometric model	37.6 34.2 35.4	1,2&3	Equal	32.4	▲	
Bohara, McNown & Batts (1987)	Macro-economic[j]	4	Q					118	28			Economic forecasters [RMSE]	50 to 60	●	◕	●	1. Judgement 2. ARIMA	9.2 9.3	1&2	Equal[k] Regression	10.2 12.9	▽ ▽	
Moriarty & Adams (1984)	Sales of consumer durables	2	M					72	36[l]			Managers		●	●	●	1. Judgement 2. Box-Jenkins[m]	14.5 ?	1&2	Regression	14.0	▲	
Sanders & Ritzman (1990)	Demand for public warehouse	22	D				◔	~750	250			Expert planners		●	●	●	1. Judgement 2. Combination[n] of objective techniques	26.3 28.1 50.6 64.1 21.6 27.9 74.0 131.0	1&2	1/4 of #1 + 3/4 of #2	17.2 21.9 67.8 120.3	▲ ▲ ▽ ▽	
Blattberg & Hoch (1990)	Catalog fashion sales, coupon redemption rates	> 1500	M?						3-6?			Buyers and brand managers (5 firms) [mean R^2]		●	●	●	1. Judgement 2. Regression models	.552 .524	1&2	Equal	.652	▲	.0001

[i] Only the results for the cattle prices are presented here. Combination was also effective with hog and broiler prices.
[j] Only the results for one representative series (corporate profits) are presented here.
[k] The MSE for the first 20 points of the forecast horizon were regressed to determine weights. The remaining eight points were used in out-of-sample comparison. A 'restricted' regression combination was also presented.
[l] The weights were determined by fitting OLS regression to 24 of the forecast points; the remaining 12 points were used in ex ante comparison.
[m] Box–Jenkins was less accurate than judgement over the full 36 points (MAPE of 19.2 versus 11.1).
[n] Adaptive exponential smoothing (1/3), single exponential smoothing (1/6), linear regression (1/6), adaptive estimation procedure (1/6) and Holt's two parameter model (1/6).

Table A3(a)
Judgemental adjustments of statistical forecasts

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	# per Series	JUDGES	Experience	Context	Motivation	BASE FORECAST	Error	ADJUST-MENT	Improvement Error	Accuracy	Significance
Willemain [a] (1989)	Artificial ARIMA [N=36]	12	12	(5)	—	✘ (drift)		✘	30	10	✘	✔	MBA students & faculty [N=12]	1	◒	○	○	SmartForecasts [b]	0.1	Holistic		(0.1)	▽	
		12			✔	✔												['excess' error]	1.8			0.0	–	
		12																	3.5			0.7	▲	
Willemain (1991)	Selected from M-competition [N=24]	7	A, Q, M						6, 8, 18				Students [4] Executives [3] Faculty [3]	10	◒	○	○	Naive SmartForecasts		Holistic		(1.2)	▽	–
		10																				(6.4)	▽	–
		7																						
Lim & O'Connor (1993)	Selected from M-Competition	10	M						30	✔	✔		Grad. students [N=64]	6?	◔	○	◔	Naive [c]	21.7	Holistic	20.3	1.4	▲	.002
																		Damped exp sm.	11.8		15.8	(4.0)	▽	.0005
		5				✘																0.7	▲	
																						(1.9)	▽	
		5				✔																2.1	▲	
																						(6.1)	▽	
Sanders (1992)	Artificial [N=10]	1	'M'	—	○	◔	○	48	12	✘	✔,✘		Undergraduates [N=38]	-7-8	○	○	○	SES	3.0	Holistic	2.9	0.1	▲	–
		1				◒													16.4		21.8	(5.4)	▽	.05
		1		◆		◔												Winters	1.5		1.8	(0.3)	▽	.05
		1				◒													8.7		12.0	(3.3)	▽	.05
		1		—	◕	◔													5.4		4.0	1.4	▲	.05
		1				◒													14.8		17.4	(2.6)	▽	.05
		1		◆		◔													3.2		2.8	0.4	▲	.05
		1				◒													19.9		20.0	(0.1)	▽	–
		1		—	○	◔	◕											SES	7.7		6.0	1.7	▲	.05
		1				◒													21.2		28.1	(6.9)	▽	.05
Carbone, Andersen, Corriveau & Corson (1983)	Selected from the M-competition [N=25]	12	A, Q, M			✘			13-150	6, 8, 18	✘		MBA students (6 teams of 2)	6	◔	◔	◔	Holt-Winters		Holistic			▽	–
																		Carbone-Longini					▽	.05
																		Box-Jenkins					▽	–
		13				✔																	▽	.05
																							▽	.05
																							▲	–

[a] Willemain also investigated the effect of autocorrelation, finding that improvement was greatest for positively autocorrelated seasonal series.
[b] SmartForecasts II selected one of five parsimoneous extrapolative techniques.
[c] The objective forecast was provided after subjects had made a judgemental extrapolation.

Table A3(b)
Judgemental adjustments of statistical forecasts

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	# per Series	JUDGES	Experience	Context	Motivation	BASE FORECAST	Error	ADJUST-MENT	Improvement Error	Accuracy	Significance
Carbone & Gorr (1985)	Selected from M-competition [N=10]	2	A, Q, M						30-140	6-18	✘	✔	MBA students [d] (7 teams of 2) [median errors]	7	◔	◔	○	Holt-Winters	3.9	Holistic [e]	5.0	(1.1)	▽	
		2																Carbone-Longini	5.4			0.4	▲	
		6																Box-Jenkins	6.7			1.7	▲	
Angus-Leppan & Fatseas (1986)	Interest rates	1	M						48	12	✘	✘	Undergraduate students	123	○	◔ [f]	◔	Combination of subjective and objective extrapolation (see exhibit 2.2)	18.0	Holistic	12.4	5.6	▲	
Wolfe & Flores [g] (1990)	Earnings per share [N=28]	14	Q						>40	4	✘	✔	Corporate loan officers	1 [h]	◕	◒		ARIMA	102	Analytic Hierarchy Process (AHP)	96	7.2	▲	.03
		14											Executive MBA students	1 [h]	○				60		56	3.9	▲	.06
Flores, Olsen & Wolfe (1992)	◄ Same as above ►																	►		Centroid method	(similar to AHP in forecast accuracy gains)			

[d] Another group of students was used, but they were not required to revise their base forecasts.
[e] Revision was made after considering all three base forecasts.
[f] Students made eyeball extrapolations (graph and table), combined these with an objective forecast, and *then* made adjustments after being told the nature of the series.
[g] Wolfe and Flores also found that improvement was greatest for series where the ARIMA forecasts were inaccurate.
[h] Two subjects (and hence series) were excluded from the analysis in each case.

Table A3(c)
Judgemental adjustments of statistical forecasts

STUDY	SERIES	#	Period	Trend	Seasonality	Noise	Instability	Historicals	Forecasts	Feedback	Graphical	JUDGES	# per Series	Experience	Context	Motivation	BASE FORECAST	Error	ADJUST-MENT	Improvement Error	Accuracy	Significance
Reinmuth & Guerts (1972)	Sales of frozen food	1	M					-60	1			Marketing staff [i] Customers	? 9	●	●	●	Similar to Winters' exp. smoothing	10.8	Bayesian	0.8	10.0	▲	n/a
Mathews & Diamantopolous (1986) and Diamantopolous & Mathews (1989)	Sales of UK health care company	281	M					8	1			Product managers	1	●	●	●	Holt's exponential smoothing	97.7	Holistic	115.9	(18.2)	▽	-
Mathews & Diamantopolous (1989, 1990, 1992)	Sales of UK health care company	900[j]	M					8-14	#1 #6	✔		← Same as above → [error shown = |(A-F)/(A+F)|]					Holt's exponential smoothing	41.1 29.9	Holistic	27.2 21.1	13.9 8.8	▲[k] ▲	.000 .000
Cook, Falchi & Mariano (1984)	Household population & macroeconomic variables	8	Q					43-76	4			Experienced forecasters & economists		●	●	●	Urban population model with a Box-Jenkins component	0.9	Analytic Hierarchy Process (AHP)	0.7	0.2	▲	
Donihue (1993)	Macroeconomic	12	Q						4			Group of forecasters [error = RMSE]	1	●	●	●	Macroeconomic model (MQEM)		Holistic			▲	
McNees (1990)	Macroeconomic	21	Q						1-8			Groups of prominent forecasters [Percent Better]	4	●	●	●	Macroeconomic models		Unknown		57 to 62	▲	.10[l]

[i] The marketing staff made probabilistic assessments of the effect of promotions on the forecast, and these assessments were modified by Bayesian incorporation of sample information from the customers.
[j] Approximately "one third to one half" of the 900 forecasts were revised. M&D (1990) found that the selection of forecasts for revision was effective.
[k] Results varied by error measure and period, but adjustment generally improved accuracy.
[l] Improvement was only significant for the one-quarter ahead forecast.

References

Abramson, B. and A. Finizza, 1991, Using belief networks to forecast oil prices, *International Journal of Forecasting*, 7, 299–315.

Adam, E.E. and R.J. Ebert, 1976, A comparison of human and statistical forecasting, *AIIE Transactions*, 8(1), 120–127.

Andreassen, P.B., 1991, Causal prediction versus extrapolation: effects of information source on judgemental forecasting accuracy, *Working paper*, Sloan School of Management.

Andreassen, P.B. and S.J. Kraus, 1990, Judgemental extrapolation and the salience of change, *Journal of Forecasting*, 9, 347–372.

Angus-Leppan, P. and V. Fatseas, 1986, The forecasting accuracy of trainee accountants using judgemental and statistical techniques, *Accounting and Business Research*, 16, 179–188.

Armstrong, J.S., 1983, Relative accuracy of judgemental and extrapolative methods in forecasting annual earnings, *Journal of Forecasting*, 2, 437–447.

Armstrong, J.S., 1985, *Long-Range Forecasting: From Crystal Ball to Computer*, 2nd edn. (Wiley-Interscience, New York).

Armstrong, J.S. and F. Collopy, 1992, Error measures for generalising about forecasting methods: empirical comparisons, *International Journal of Forecasting*, 8, 69–80.

Armstrong, J.S., W.B. Denniston and M.M. Gordon, 1975, The use of the decomposition principle in making judgements, *Organisational Behaviour and Human Performance*, 14, 257–263.

Balzer, W.K., L.M. Sulsky, L.B. Hammer and K.E. Sumner, 1992, Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? *Organisational Behaviour and Human Decision Processes*, 53, 35–54.

Batchelor, R. and P. Dua, 1990, Forecaster ideology, forecasting technique, and the accuracy of economic forecasts, *International Journal of Forecasting*, 6, 3–10.

Beach, L.R. and C.R. Peterson, 1967, Man as an intuitive statistician, *Psychological Bulletin*, 68(7), 29–46.

Beach, L.R., V.E. Barnes and J.J.J. Christensen-Szalanski, 1986a, Beyond heuristics and biases: a contingency model

of judgemental forecasting, *Journal of Forecasting*, 4, 143–157.

Beach, L.R., J.J.J. Christensen-Szalanski and V.E. Barnes, 1986b, Assessing human judgement: has it been done, can it be done, should it be done?, in: G. Wright and P. Ayton, eds., *Judgemental Forecasting* (Wiley, New York), 49–62.

Benbasat, I. and A.S. Dexter, 1985, An experimental evaluation of graphical and colour-enhanced information presentation, *Management Science*, 31(11), 1348–1364.

Benson, P.G. and D. Önkal, 1992, The effects of feedback and training on the performance of probability forecasters, *International Journal of Forecasting*, 8, 559–573.

Bessler, D.A. and J.A. Brandt, 1981, Forecasting livestock prices with individual and composite methods, *Applied Economics*, 13, 513–522.

Blattberg, R.C. and S.J. Hoch, 1990, Database models and managerial intuition: 50% model + 50% manager, *Management Science*, 36(8), 887–899.

Bohara, A., R. McNown and J.T. Batts, 1987, A re-evaluation of the combination and adjustment of forecasts, *Applied Economics*, 19, 437–445.

Box. G.E.P. and G.M. Jenkins, 1970, *Time Series Analysis: Forecasting and Control* (Holden-Day).

Brown, L.D., 1988, Comparing judgmental to extrapolative forecasts: it's time to ask why and when, *International Journal of Forecasting*, 4, 171–173.

Brown, P., G. Foster and E. Noreen, 1985, *Security Analyst Multi-Year Earnings Forecasts and the Capital Market* (American Accounting Association).

Brown, L.D., R.L. Hagerman, P.A. Griffin and M.E. Zmijewski, 1987, Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings, *Journal of Accounting and Economics*, 9, 61–87.

Bruner, J.S., J.J. Goodnow and G.A. Austin, 1956, *A Study of Thinking* (Wiley, New York).

Bunn, D., 1989, Forecasting with more than one model, *Journal of Forecasting*, 8, 161–166.

Bunn, D. and G. Wright, 1991, Interaction of judgemental and statistical forecasting methods: issues and analysis, *Management Science*, 37(5), 501–518.

Carbone, R. and W.L. Gorr, 1985, Accuracy of judgemental forecasting of time series, *Decision Sciences*, 16, 153–160.

Carbone, R., A. Andersen, Y. Corriveau and P.P. Corson, 1983, Comparing for different time series methods the value of technical expertise, individualized analysis and judgemental adjustment, *Management Science*, 29(5), 559–566.

Cerullo, M.J. and A. Avila, 1975, Sales forecasting practices: a survey, *Managerial Planning*, September/October, 33–39.

Clemen, R.T., 1989, Combining forecasts: a review and annotated bibliography, *International Journal of Forecasting*, 5, 559–583.

Clemen, R.T. and J.B. Guerard, 1989, Econometric GNP forecasts: incremental information relative to naive extrapolation, *International Journal of Forecasting*, 5, 417–426.

Clemen, R.T. and R.L. Winkler, 1985, Limits for the prediction and value of information from dependent sources, *Operations Research*, 33(March–April), 427–442.

Collopy, F. and J.S. Armstrong, 1992a, Expert opinions about extrapolation and the mystery of the overlooked discontinuities, *International Journal of Forecasting*, 8, 575–582.

Collopy, F. and J.S. Armstrong, 1992b, Rule-based forecasting: development and validation of an expert systems approach to combining time series extrapolations, *Management Science*, 38, 1394–1414.

Collopy, F. and J.S. Armstrong, 1993, Decomposition by causal forces: using domain knowledge to forecast highway deaths, *Working paper*.

Conroy, R. and R. Harris, 1987, Consensus forecasts of corporate earnings: analysts' forecasts and time series methods, *Management Science*, 33(6), 725–738.

Cook, T., P. Falchi and R. Mariano, 1984, An urban allocation model combining time series and analytic hierarchical methods, *Management Science*, 20(2), 198–208.

Dalrymple, D.J., 1987, Sales forecasting practices: results from a United States survey, *International Journal of Forecasting*, 3, 379–392.

Dawes, R.M., D. Faust and P.E. Meehl, 1989, Clinical versus actuarial judgement, *Science*, 243(March), 1668–1674.

De Bondt, W.F.M., 1990, Risk and return in the stock market as most people (who don't know) see it, *Working paper*, University of Wisconsin–Madison.

DeSanctis, G., 1984, Computer graphics as decision aids: directions for research, *Decision Sciences*, 15(4), 463–487.

DeSanctis, G. and S.L. Jarvenpaa, 1989, Graphical presentation of accounting data for financial forecasting: an experimental investigation, *Accounting, Organisations and Society*, 14(5/6), 509–525.

Diamantopolous, A. and B.P. Mathews, 1989, Factors affecting the nature and effectiveness of subjective revision in sales forecasting: an empirical study, *Managerial and Decision Economics*, 10, 51–59.

Dickson, G.W., G. DeSanctis and D.J. McBride, 1986, Understanding the effectiveness of computer graphics for decision support: a cumulative experimental approach, *Communications of the ACM*, 29(1), 40–47.

Diebold, F.X., 1989, Forecast combination and encompassing: reconciling two divergent literatures, *International Journal of Forecasting*, 5, 589–592.

Donihue, M.R., 1993, Evaluating the role judgment plays in forecast accuracy, *Journal of Forecasting*, 12, 81–92.

Edmundson, R.H., 1987, A computer based decision aid for forecasting monthly time series, *Ph.D dissertation*, University of New South Wales.

Edmundson, R.H., 1990, Decomposition: a strategy for judgemental forecasting, *Journal of Forecasting*, 9, 301–314.

Edmundson, R.H., M.J. Lawrence and M.J. O'Connor, 1988, The use of non-time series information in sales forecasting: A case study, *Journal of Forecasting*, 7, 201–211.

Eggleton, I.R.C., 1982, Intuitive time-series extrapolation, *Journal of Accounting Research*, 20(1), 68–102.

Einhorn, H.J. and R.M. Hogarth, 1982, Prediction, diagnosis, and causal thinking in forecasting, *Journal of Forecasting*, 1, 23–36.

Fischer, G.W., 1982, Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting, *Organisational Behaviour and Human Performance*, 29, 352–369.

Fischhoff, B. and M. Bar-Hillel, 1984, Focusing techniques: a shortcut to improving probability judgements? *Organisational Behaviour and Human Performance*, 34, 174–194.

Flores, B.E., D.L. Olson and C. Wolfe, 1992, Judgemental adjustment of forecasts: a comparison of methods, *International Journal of Forecasting*, 7, 421–433.

Gardner, E.S. and E. McKenzie, 1985, Forecasting trends in time series, *Management Science*, 31(10), 1237–1246.

Goodwin, P. and G. Wright, 1993, Improving judgemental time series forecasting: a review of the guidance provided by research, *International Journal of Forecasting*, 9, 147–161.

Gorr, W.L., 1986a, Use of special event data in government information systems, *Public Administrative Review*, Special Issue, 532–539.

Gorr, W.L., 1986b, Special event data in shared databases, *MIS Quarterly*, 10(September), 239–255.

Granger, C.W.J. and P. Newbold, 1973, Some comments on the evaluation of economic forecasts, *Applied Economics*, 5, 35–47.

Guerard, J.B., 1987, Linear constraints, robust-weighting and efficient composite modelling, *Journal of Forecasting*, 6(3), 193–199.

Guerard, J.B. and C.R. Beidleman, 1987, Composite earnings forecasting efficiency, *Interfaces*, 17(5), 103–113.

Harvey, A.C. and J. Durbin, 1986, The effects of seat belt legislation on British road casualties: a case study in structural time series modelling, *Journal of the Royal Statistical Society, Series A*, 149(3), 187–227.

Henry, R.A. and J.A. Sniezek, 1993, Situational factors affecting judgments of future performance, *Organisational Behaviour and Human Decision Processes*, 54, 104–132.

Hopwood, W.S. and J.C. McKeown, 1990, Evidence on surrogates for earnings expectations within a capital market context, *Journal of Accounting, Auditing and Finance*, 5(Summer), 339–368.

Jones, G.V., 1984, Perception of inflation: polynomial not exponential, *Perception and Psychophysics*, 36, 485–487.

Kang, H., 1986, Unstable weights in the combination of forecasts, *Management Science*, 32(6), 683–695.

Kleinmuntz, B., 1990, Why we still use our heads instead of formulas: toward an integrative approach, *Psychological Bulletin*, 107, 296–310

Kurke, L.B. and H.E. Aldrich, 1983, Mintzberg was right!: a replication and extension of The Nature of Managerial Work, *Management Science*, 29(8), 975–984.

Laestadius, J.E., Jr., 1970, Tolerance for errors in intuitive mean estimations, *Organisational Behaviour and Human Performance*, 5, 121–124.

Lawrence, M.J., 1983, An exploration of some practical issues in the use of quantitative forecasting models, *Journal of Forecasting*, 1, 169–179.

Lawrence, M.J. and S. Makridakis, 1989, Factors affecting judgemental forecasts and confidence intervals, *Organisational Behaviour and Human Decision Processes*, 42, 172–189.

Lawrence, M.J. and M.J. O'Connor, 1992, Exploring judgemental forecasting, *International Journal of Forecasting*, 8, 15–26.

Lawrence, M.J. and M.J. O'Connor, 1993, Heads or models: the importance of task differences, *Working paper*, University of N.S.W.

Lawrence, M.J., R.H. Edmundson and M.J. O'Connor, 1985, An examination of the accuracy of judgemental extrapolation of time series, *International Journal of Forecasting*, 1, 25–35.

Lawrence, M.J., R.H. Edmundson and M.J. O'Connor, 1986, The accuracy of combining judgemental and statistical forecasts, *Management Science*, 32(12), 1521–1532.

Lawrence, M.J., R.H. Edmundson and M.J. O'Connor, 1994, A field study of forecasting accuracy, *Working paper*, University of N.S.W.

Lee, J.K., S.B. Oh and J.C. Shin, 1990, UNIK-FCST: knowledge-assisted adjustment of statistical forecasts, *Expert Systems with Applications*, 1, 39–49.

Lewandowski, R., 1982, Sales forecasting by FORSYS, *Journal of Forecasting*, 1, 205–214.

Lim, J.S., 1993, An empirical investigation of the effectiveness of time series judgemental adjustment using forecasting support systems, *Ph.D dissertation*, University of N.S.W.

Lim, J.S. and M.J. O'Connor, 1993, Graphical adjustment of initial forecasts: how good are people at the task? *Working paper*, University of N.S.W.

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

Lobo, G.J. and R.D. Nair, 1991, Analysts utilisation of historical earnings information, *Managerial and Decision Economics*, 12, 383–393.

Lyness, K.S. and E.T. Cornelius, 1982, A comparison of holistic and decomposed judgment strategies in a performance rating simulation, *Organisational Behaviour and Human Performance*, 29, 21–38.

MacGregor, D. and J.S. Armstrong, 1993, Judgemental decomposition: when does it work? *Working paper*.

MacGregor, D.G. and S. Lichtenstein, 1991, Problem structuring aids for quantitative estimation, *Journal of Behavioural Decision Making*, 4, 101–116.

MacGregor, D.G., S. Lichtenstein and P. Slovic, 1988, Structuring knowledge retrieval: an analysis of decomposed quantitative judgments, *Organisational Behaviour and Human Decision Processes*, 42, 303–323.

Mackinnon, A.J. and A.J. Wearing, 1991, Feedback and the forecasting of exponential change, *Acta Psychologica*, 76, 177–191.

Makridakis, S., 1988, Metaforecasting: ways of improving

forecasting accuracy and usefulness, *International Journal of Forecasting*, 4, 467–491.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen and R. Winkler, 1982, The accuracy of extrapolation (time series) methods: results of a forecasting competition, *Journal of Forecasting*, 1, 111–153.

Makridakis, S., S.C. Wheelwright and V.E. McGee, 1983, *Forecasting: Methods and Applications*, 2nd edn. (Wiley, New York).

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord and L.F. Simmons, 1993, The M2-Competition: a real-time judgmentally based forecasting study, *International Journal of Forecasting*, 9, 5–22.

Mathews, B.P. and A. Diamantopolous, 1986, Managerial intervention in forecasting: an empirical investigation of forecast manipulation, *International Journal of Research in Marketing*, 3, 3–10.

Mathews, B.P. and A. Diamantopolous, 1989, Judgemental revision of sales forecasts: a longitudinal extension, *Journal of Forecasting*, 8, 129–140.

Mathews, B.P. and A. Diamantopolous, 1990, Judgemental revision of sales forecasts: effectiveness of forecast selection, *Journal of Forecasting*, 9, 407–415.

Mathews, B.P. and A. Diamantopolous, 1992, Judgemental revision of sales forecasts: the relative performance of judgementally revised versus non-revised forecasts, *Journal of Forecasting*, 11, 569–576.

McNees, S.K., 1990, The role of judgment in macroeconomic forecasting accuracy, *International Journal of Forecasting*, 6, 287–299.

Meehl, P.E., 1957, When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, 4, 268–273.

Mentzer, J.T. and J.E. Cox, 1984, Familiarity, application, and performance of sales forecasting techniques, *Journal of Forecasting*, 3, 27–36.

Miller, G.A., 1956, The magic number seven plus or minus two: some limits on our capacity for processing information, *Psychological Review*, 63, 81–97.

Mintzberg, H., 1975, The manager's job: folklore and fact, *Harvard Business Review*, July/August, 49–61.

Moriarty, M.M., 1990, Boundary value models for the combination of forecasts, *Journal of Marketing Research*, 27(November), 402–417.

Moriarty, M.M. and A.J. Adams, 1984, Management judgment forecasts, composite forecasting methods, and conditional efficiency, *Journal of Marketing Research*, 21(August), 239–250.

Mosteller, F., A.F. Siegal, E. Trapido and C. Youtz, 1981, Eye fitting straight lines, *The American Statistician*, 35(3), 150–152.

Murphy, A.H. and B.G. Brown, 1984, A comparative evaluation of objective and subjective weather forecasts in the United States, *Journal of Forecasting*, 3, 369–393.

Newbold, P. and C.W.J. Granger, 1974, Experience with forecasting univariate time series and the combination of forecasts (with discussion), *Journal of the Royal Statistical Society, Series A*, 137, 131–149.

Newbold, P., J.K. Zumwalt and S. Kannan, 1987, Combining forecasts to improve earnings per share prediction, *International Journal of Forecasting*, 3, 229–238.

O'Connor, M., 1989, Models of human behaviour and confidence in judgement, *International Journal of Forecasting*, 5, 159–169.

O'Connor, M., W. Remus and K. Griggs, 1993, Judgemental forecasting in times of change, *International Journal of Forecasting*, 9, 163–172.

Pankratz, A., 1989, Time series forecasts and extra-model information, *Journal of Forecasting*, 8(2), 75–84.

Phelps, R.H. and J. Shanteau, 1978, Livestock judges: how much information can an expert use? *Organisational Behaviour and Human Performance*, 21, 209–219.

Raiffa, H., 1968, *Decision Analysis* (Addison-Wesley, Reading, MA).

Reinmuth, J.E. and M.D. Guerts, 1972, A Bayesian approach to forecasting effects of atypical situations, *Journal of Marketing Research*, 9(August), 292–297.

Remus, W.E., 1984, An empirical investigation of the impact of graphical and tabular data presentations on decision making, *Management Science*, 30(5), 533–542.

Remus, W.E., 1987, A study of graphical and tabular displays and their interaction with environmental complexity, *Management Science*, 33(9), 1200–1204.

Remus, W.E., 1990, Will information systems research generalize to managers? *Working paper*.

Sanders, N.R., 1992, Accuracy of judgemental forecasts: a comparison, *Omega International Journal of Management Science*, 20(3), 353–364.

Sanders, N.R. and K.B. Manrodt, 1994, Forecasting practices in US corporations: survey results, *Interfaces*, 24(2), 92–100.

Sanders, N.R. and L.P. Ritzman, 1990, Improving short-term forecasts, *Omega International Journal of Management Science*, 18(4), 365–373.

Sanders, N.R. and L.P. Ritzman, 1992, The need for contextual and technical knowledge in judgemental forecasting, *Journal of Behavioural Decision Making*, 5, 39–52.

Schnaars, S.P. and Mohr, I., 1988, The accuracy of Business Week's industry outlook survey, *Interfaces*, 18(5), 31–38.

Schroder, H.M., M.J. Driver and S. Streufert, 1967, *Human Information Processing* (Holt, New York).

Silverman, B.G., 1992, Judgment error and expert critics in forecasting tasks, *Decision Sciences*, 23(5), 1199–1219.

Simon, H.A. and R.K. Sumner, 1968, Pattern in music, in: Kleinmuntz, ed., *Formal Representation of Human Judgment* (Wiley, New York), 219–250.

Sparkes, J.R. and A.K. McHugh, 1984, Awareness and use of forecasting techniques in British industry, *Journal of Forecasting*, 3, 37–42.

Taranto, G.M., 1989, Sales forecasting practices: results from an Australian survey, Unpublished thesis, University of N.S.W.

Timmers, H. and W.A. Wagenaar, 1977, Inverse statistics

and misperception of exponential growth, *Perception and Psychophysics*, 21(6), 558–562.

Turner, D., 1990, The role of judgement in macroeconomic forecasting, *Journal of Forecasting*, 9, 315–346.

Tversky, A. and D. Kahneman, 1974, Judgement under uncertainty: heuristics and biases, *Science*, 185, 1124–1131.

Wagenaar, W.A. and S.D. Sagaria, 1975, Misperception of exponential growth, *Perception and Psychophysics*, 18(6), 416–422.

Walker, K.B. and L.A. McClelland, 1991, Management forecasts and statistical prediction model forecasts in corporate budgeting, *Journal of Accounting Research*, 29(2), 371–381.

West, M. and J. Harrison, 1989, Subjective intervention in formal models, *Journal of Forecasting*, 8, 33–53.

Willemain, T.R., 1989, Graphical adjustment of statistical forecasts, *International Journal of Forecasting*, 5, 179–185.

Willemain, T.R., 1991, The effect of graphical adjustment on forecast accuracy, *International Journal of Forecasting*, 7, 151–154.

Wind, Y., V. Mahajan and R.N. Cardozo, 1981, *New Product Forecasting* (Lexington Books, Lexington, MA).

Winton, E. and R.H. Edmundson, 1993, Trend identification and extrapolation in judgemental forecasting, *Working paper*, University of N.S.W.

Wolfe, C. and B. Flores, 1990, Judgemental adjustment of earnings forecasts, *Journal of Forecasting*, 9, 389–405.

Wood, R.E., 1986, Task complexity: definition of the construct, *Organisational Behaviour and Human Decision Processes*, 37, 60–82.

Biographies: Richard WEBBY is a lecturer in the School of Information Systems at the University of New South Wales. His research focuses on the development and empirical validation of software designed to support human judgement. His doctoral work examined the application of a forecasting support system to aid judgemental forecasters with both time series data and extra-model information.

Marcus O'CONNOR is a member of the academic staff of the School of Information Systems at the University of New South Wales. His research interests centre on the way people use information in making judgements. He typically uses the forecasting task as a vehicle to examine the ability of people to combine both causal and non-causal information in making forecasts.