Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

International Journal of Forecasting 25 (2009) 3–23

international journal of forecasting

www.elsevier.com/locate/ijforecast

Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning

Robert Fildes[a,*], Paul Goodwin[b], Michael Lawrence[c], Konstantinos Nikolopoulos[d]

[a] *Department of Management Science, Lancaster University, LA1 4YX, UK*
[b] *School of Management, University of Bath, BA2 7AY, UK*
[c] *Australian School of Business, University of New South Wales, Sydney, Australia*
[d] *Manchester Business School, Manchester M15 6PB, UK*

Abstract

Demand forecasting is a crucial aspect of the planning process in supply-chain companies. The most common approach to forecasting demand in these companies involves the use of a computerized forecasting system to produce initial forecasts and the subsequent judgmental adjustment of these forecasts by the company's demand planners, ostensibly to take into account exceptional circumstances expected over the planning horizon. Making these adjustments can involve considerable management effort and time, but do they improve accuracy, and are some types of adjustment more effective than others? To investigate this, we collected data on more than 60,000 forecasts and outcomes from four supply-chain companies. In three of the companies, on average, judgmental adjustments increased accuracy. However, a detailed analysis revealed that, while the relatively larger adjustments tended to lead to greater average improvements in accuracy, the smaller adjustments often damaged accuracy. In addition, positive adjustments, which involved adjusting the forecast upwards, were much less likely to improve accuracy than negative adjustments. They were also made in the wrong direction more frequently, suggesting a general bias towards optimism. Models were then developed to eradicate such biases. Based on both this statistical analysis and organisational observation, the paper goes on to analyse strategies designed to enhance the effectiveness of judgmental adjustments directly.

Keywords: Forecasting accuracy; Judgment; Heuristics and biases; Supply chain; Forecasting support systems; Practice; Combining; Forecast adjustment

1. Introduction

Supply chain planning is usually reliant on demand forecasts at the stock keeping unit (SKU) level. The accuracy achieved for these forecasts has consequences for companies at all levels of the supply

* Corresponding author.
E-mail addresses: R.Fildes@lancaster.ac.uk (R. Fildes), P.Goodwin@bath.ac.uk (P. Goodwin), michael.lawrence@unsw.edu.au (M. Lawrence), Kostas.Nikolopoulos@mbs.ac.uk (K. Nikolopoulos).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2008.11.010

chain, from the retailer to the raw materials supplier, and even for companies whose final product is 'make-to-order' (Yelland, 2006). Errors at each stage of the chain are potentially amplified, resulting in poor service or excess inventory levels. The forecasting task is difficult due to the inter-related nature of the data series, the presence of outliers, level and trend shifts (Fildes & Beard, 1992), and the impacts of the market and general economic environment. These data difficulties are compounded by the huge number of SKUs that often need to be forecast each period.

In order to plan and manage their supply chain, organisations typically set up a unit responsible for forecasting. Because of the size and complexity of the forecasting task, it is generally impossible for all SKUs to be given individual attention by demand planners. The most common approach to forecasting demand in support of supply chain planning involves the use of a statistical software system which incorporates a simple univariate forecasting method, such as exponential smoothing, to produce an initial forecast. For key products, these initial forecasts (hereafter called the 'system' forecasts) are reviewed, and may be adjusted by the company's demand planners to take into account exceptional circumstances expected over the planning horizon, or possibly to correct perceived inadequacies in the system forecast. This 'Sales and Operations Planning' process is usually carried out in a committee setting, where representatives from marketing, sales, production and logistics agree on the 'final forecast': a combination of a statistical forecast and managerial judgment.

Improved demand forecasting accuracy can lead to significant monetary savings, greater competitiveness, enhanced channel relationships, and customer satisfaction (Moon, Mentzer, & Smith, 2003). However, the topic has generated relatively little organisationally-based research, and little is known about the effects of the types of judgmental adjustment (e.g. large or small, positive or negative) on accuracy, or the extent to which the resulting forecasts are unbiased and efficient (i.e. make optimal use of the available information).

This paper uses extensive data gathered from four supply-chain companies to address these questions and to consider how the process of adjustment can be made more effective. Section 2 considers the literature on adjustment and proposes hypotheses about the adjustment process. Section 3 describes the forecasting processes in the four companies and gives details of the frequency and nature of the judgmental interventions. Section 4 examines the detailed hypotheses developed in the literature review, while Section 5 evaluates some potential solutions to the problems that have been identified. The final section offers our recommendations for improvements in both forecasting support systems and organisational forecasting processes.

2. Literature review and hypotheses

There is substantial evidence from the economic forecasting literature that statistical forecasts can be made more accurate when experts judgmentally adjust them to take into account the effects of special events and changes that were not incorporated into the statistical model (Donihue, 1993; McNees, 1990; Turner, 1990). However, few studies have investigated judgmental adjustment in the context of company forecasts of the demand for SKUs. The exceptions were six studies, all based in the same company, by Mathews and Diamantopoulos (1986, 1989, 1990, 1992, 1994) and Diamantopoulos and Mathews (1989). These showed that judgmental adjustments tend to improve accuracy, though sometimes only marginally, but that they may also introduce bias.

Experimental evidence on the same question generally suggests that forecasters often make unnecessary judgmental adjustments to statistical forecasts (Lawrence, Goodwin, O'Connor, & Onkal, 2006). In particular, they make adjustments even when they do not possess additional information about special events. There is evidence that this occurs because forecasters see patterns in the noise associated with random fluctuations in the time series (Harvey, 1995; O'Connor, Remus, & Griggs, 1993). In addition, it could be due to the widely observed illusion of control effect, where users who make adjustments exhibit greater confidence in their forecasts (Kottemann, Davis, & Remus, 1994; Langer, 1975). Lim and O'Connor (1995) even found that a tendency by forecasters to make damaging adjustments persisted, despite a computer display showing that they were reducing the accuracy.

However, experimental evidence also suggests that when an adjustment is made on the basis of events not reflected in the statistical forecast (e.g. a forthcoming sales promotion), it is likely to improve accuracy, as long as the information about the event is reliable (Goodwin & Fildes, 1999; Lim & O'Connor, 1996). Forecast adjustments made by experts in a company environment with access to reliable market intelligence, are likely to yield greater benefits than those obtained in experiments by student subjects. Most importantly, organisations employ substantial resources in the forecast adjustment activity, and economic rationality argues that they must view it as valuable. Thus, we hypothesize:

H_1: Judgmental forecast adjustments improve forecast accuracy.

The study by Mathews and Diamantopoulos (1990) found that managers were able to select the most inadequate system forecasts and then to adjust them in the correct direction. In addition, Diamantopoulos and Mathews (1989) found that larger judgmental adjustments were more effective in improving forecast accuracy than smaller ones. However, as this study was based on a single company, and the statistical method (Holt's method) had been fitted to only eight quarterly observations, it is possible that these findings will not hold for the different companies and forecasting systems analysed here. Nevertheless, there are a number of reasons why the result relating to the size of adjustments might apply more generally. Large adjustments are likely to be associated with reliable information about events which will have large anticipated effects not reflected in the system forecast. We would expect such adjustments to be associated with substantial improvements in accuracy. Smaller adjustments are likely to be less effective, for the reasons outlined below.

(a) If the information on which the adjustment is based is viewed as unreliable, the forecasters are likely to hedge their bets by reducing the size of the adjustment.
(b) Research on human decision making has shown a tendency across a number of decision environments for people to ignore or modify good advice, and to demonstrate excessive trust in their own judgement. This tendency has been observed in the illusion of control literature (Kottemann et al., 1994; Langer, 1975), the advice taking literature (Bonaccio & Dalal, 2006; Yaniv, 2004), where users typically discount good advice, and for computer mediated advice (Waern & Ramberg, 1996), where users are reluctant to trust this form of advice compared to human mediated advice. Furthermore, Lim and O'Connor (1995) showed that even when an impossibly good forecast was provided,[1] users still made small adjustments to it, which naturally reduced its accuracy. This literature consistently suggests a likelihood for forecasters to make small modifications to a provided forecast which have no sound basis, and therefore diminish its accuracy.
(c) Based on our observations of forecasting in companies and discussions with forecasters, we suggest that forecasters often tinker with the system forecasts merely to demonstrate that they have reviewed the forecasts, and are attending to the task.

We characterise an adjustment as either large or small on a relative, within company, basis, where the adjustment is expressed as a percentage of the system forecast. Thus we hypothesise:

H_{2a}: The forecasts selected for adjustment are those most in need of improvement.

H_{2b}: When adjustments are made, the sizes of the judgmental adjustments are positively associated with an improvement in accuracy.

The relationship between the volatility of a time series (as measured by the coefficient of variation of the raw data) and the relative accuracy of judgmental and statistical forecasts was examined by Sanders and Ritzman (1992). Their study concluded that when series had low coefficients of variation, statistical time series methods outperformed judgmental forecasters who had expertise relating to the variables to be forecast. However, the opposite was true when series had high coefficients of variation (over about 30%). Moreover, the experts increasingly outperformed statistical time series methods as the volatility of the series increased. This suggests that, for the series they examined, higher levels of volatility were more the result of the effects of special events, of which the judgmental forecasters had some prior

[1] The 'impossibly good forecast' was developed by averaging an exponential smoothing forecast with the next actual.

knowledge, than of noise. When the volatility in a series reflects noise or unanticipated discontinuities, there is evidence that judgmental forecasters will perform poorly relative to statistical methods as the volatility increases (O'Connor et al., 1993). This is because the forecasters' propensity to overreact to noise is exacerbated in conditions of high noise (Goodwin & Fildes, 1999; Harvey, 1995). Because our company time series were regularly affected by special events, like sales promotions, about which the forecasters usually had advance information, we hypothesise that:

H_3: Judgmental forecast adjustments improve the forecast accuracy more under high volatility than low volatility conditions.

While accuracy is the most important property for a forecast, two further properties are also important: bias and efficiency. Bias, itself, can be decomposed into two components (Theil, 1966). Mean bias is a systematic tendency for the forecast to be either less or greater than the actual. Regression bias is the extent to which the forecasts systematically fail to track the actual observations. For example, forecasts may tend to be too high when outcomes are low and too low when outcomes are high. Efficiency is the property that forecasts optimally incorporate relevant new information as it becomes available. While adjustments might improve accuracy, they may not be either unbiased or efficient. Lawrence, O'Connor, and Edmundson (2000) found that judgmental forecasts made by 13 large manufacturing organisations were generally neither unbiased nor efficient. They found that many forecasters faced a situation of asymmetric management incentives (depending on the sign of the forecast error), which resulted in biased forecasts. In particular, differences between over-stocking costs and under-stocking costs may lead to forecast bias, though the direction seems to depend on the organisational circumstances (Sanders & Manrodt, 1994, and Stewart in Fildes et al., 2003). In addition, it has been found, both in the UK (Fildes & Hastings, 1994) and in the US (Galbraith & Merrill, 1996), that around 60% of forecasts were negatively affected by organisational politics (also see Deschamps, 2004). The literature also provides plenty of evidence for over-optimism in management judgment, and has been shown to impact a wide variety of forecasts including security analysts' forecasts (Helbok & Walker, 2004), project time prediction (Buehler & Griffin, 2003) and capital budgeting (Flybjerg, Bruzelius, & Rothengatter, 2003). Moreover, in their study in a health products company, Mathews and Diamantopoulos (1989) found evidence suggesting an optimism bias in managers' revisions of forecasts, though in this case the adjustments may have been partly a reaction to systematic underestimation by the statistical forecast. We therefore hypothesise:

H_4: The judgmentally adjusted forecasts are biased.

There are a number of reasons why forecasts based, in part, on management judgment are likely to be inefficient; that is, where the forecasts could be improved by modifying them to take into account information available to the forecaster at the time. Human cognitive limitations mean that people will struggle to optimally incorporate in their forecasts the effects of information from multiple sources (Fildes, 1991). As a result, they may restrict their attention to only one or two sources. Moreover, when estimating these effects, forecasters may over-rely on the recall of single analogies from the past, and they may anchor too closely to these recalled effects (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007). In addition, the 'escalation of commitment' literature (Staw, 1976) demonstrates a strong reluctance in human judgment to modify a view already held about the future. Thus, evidence from actual sales of a lack of response to a promotion, for example, is likely not to be believed. No studies of both bias and efficiency have been carried out on forecast *adjustments* to SKU data. However, based on the limited evidence available to date, we hypothesise:

H_5: The judgmentally adjusted forecasts are inefficient.

While these hypotheses are of theoretical interest, more excitingly they also open up the prospect of improving the accuracy of companies' forecasting processes by eradicating any consistent biases, inefficiencies and size effects. We examine these possibilities later in the paper.

3. The data base and preliminary analysis

Data have been collected at SKU level from four companies, three in manufacturing with monthly forecasts (pharmaceuticals (A), food (B), and household products (C)), and one retailer (D) forecasting weekly.

For companies A–C, all SKUs in the company were examined. For the retailer, data on two product groups each supplied by an individual manufacturer were made available (D1 and D2). The data included one-step-ahead statistical systems forecasts, the final forecasts, and the corresponding actual outcomes. SKUs without the required continuous forecast history were excluded from the analysis, as were low volume SKUs, defined as those with actual outcomes or system forecasts of less than 10 units. In addition, SKUs where the final forecast is zero were not analysed, as these were thought to result from special circumstances such as the particular SKU being withdrawn from the market. These low volume items have long been seen as special cases, ideally requiring their own particular statistical forecasting models, and would not typically require the same attention as higher volume items. As a consequence, the 'intermittent demand' case has been examined elsewhere, focussing on inventory effects (Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, in press). However, we refer to the main findings of this study later, to allow comparisons between the two types of series.

In all four organisations the forecasting process was observed and discussions were held with the principal forecasters. Each organisation uses a broadly similar process to estimate their final forecasts. At the start of each forecasting period, the statistical 'system' forecasts are produced using the computer software. The software offers the user options to change the data base or the statistical model. Three companies use systems that are based on variants of exponential smoothing (two being commercial and one developed in-house), whilst one uses a commercial system related to Focus Forecasting (Gardner & Anderson, 1997). A forecasting meeting follows, generally involving forecasting, marketing, production and sales personnel, which then examines the system forecasts in the light of various pieces of marketing and other information, and agrees on the final forecast. Particular aspects of the recent forecasting performance, such as a large error, might be drawn to the group's attention. All company forecasters interviewed affirmed that the principal objective of the forecasting process was producing accurate forecasts, as opposed to targets or politically acceptable predictions, and said that their final forecasts were not subsequently changed by more senior management.

Table 1 summarises the full data base, showing the number of observations containing (i) the statistical forecast, (ii) the final forecast, and (iii) the actual outcome. The table also shows the percentage of forecasts adjusted and the number of SKUs contained in the data set. Companies A–C, all manufacturers which make monthly forecasts, adjust a substantially greater percentage of forecasts than Company D1 and D2, the retailer which makes weekly forecasts. We group the organisational data in this way because analysis shows that the groups have common characteristics; however, any substantive individual company differences will also be noted.

Various measures have been proposed by Mathews and Diamantopoulos (1987) to measure forecast adjustment. They argue for a 'symmetric measure', where the adjustment is measured relative to an average of the system and final forecasts; however, as Goodwin and Lawton (1999) have shown, their proposed measure does not have these properties, and we have therefore chosen their more intuitive measure, defined as $100 \times$ (Final forecast − System forecast)/System forecast. Fig. 1 shows the size of the relative adjustments in the two groups of companies. Both distributions are right skewed, but the retailer (D) tends to make larger adjustments. Table 2 shows the mean and median sizes of the relative adjustments categorised by the direction of adjustment (a positive adjustment is made when the final forecast exceeds the system forecast). It can be seen that positive adjustments for both groups of organisations are very much larger than the negative adjustments (which are bounded by zero).

4. Accuracy, unbiasedness and efficiency in the adjusted forecasts

4.1. Accuracy

Although the measurement of forecasting accuracy is controversial (Armstrong & Fildes, 1995; Clements & Hendry, 1995), the use of absolute percentage error measures is now general within company settings (Fildes & Goodwin, 2007). However, the disadvantages of such measures are well known. In particular, they suffer from being sensitive to extremes (Armstrong & Collopy, 1992), which would be problematic in our database, despite our large

Table 1
Database of SKU forecasts by company (3.7% of triples fell into the low volume category and were omitted from further analysis).

Companies	Data	Available data	Total complete triples[a]	% adjusted	No. of SKUs
A	Monthly	1/2003–12/2005	5,428	65	213
B		5/2004–12/2005	2,856	91	296
C		3/2004–12/2005	3,012	63	244
D1	Weekly	1/2004–52/2005	12,789	14	191
D2		1/2004–52/2005	44,899	10	592
Total			68984	21	1536

[a] A triple consists of the actual observation, the system forecast and the final forecast.

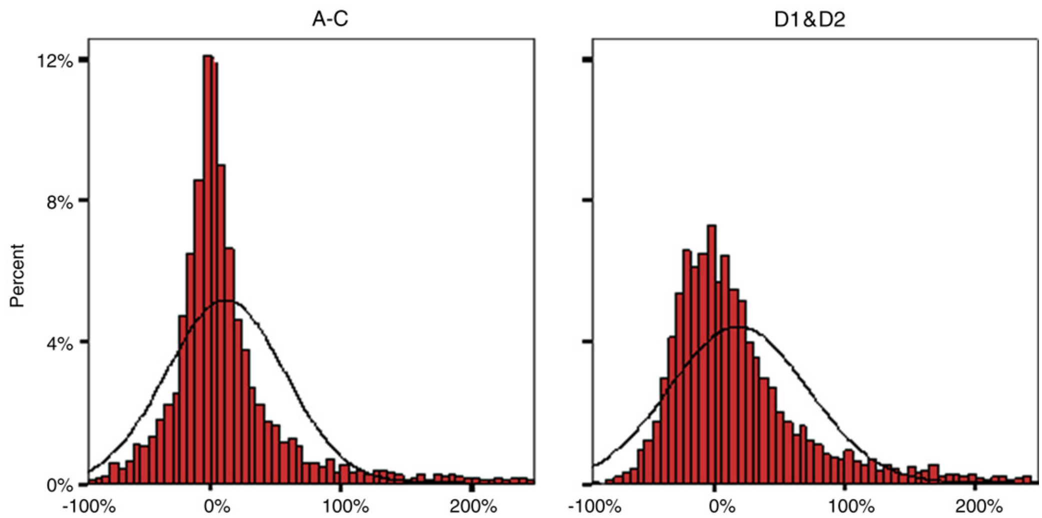

Fig. 1. Size of relative adjustments to the forecasts in two groups of companies.

Table 2
Mean and median relative adjustment by organisational group, and the direction of adjustment.

Org. group	Direction of adjustment	*N*	Relative adjustment		
			Mean (Trimmed) (%)	Median (%)	90th Percentile (%)
A–C	Positive adjustment	4013	57.2 (46.1)	20.1	132
	Negative adjustment	3392	19.0 (18.6)	13.1	46.9
D1 & D2	Positive adjustment	3049	60.3 (54.4)	32.5	144
	Negative adjustment	2409	22.6 (22.3)	20.2	44.2

number of observations. In evaluating the accuracy and bias of the statistical forecasts (SFC) and the final forecasts (FFC), we therefore report two more robust variants, the trimmed mean absolute percentage error (MAPE) and the median absolute percentage error (MdAPE). For the MAPE, we have removed outlying values using a 2% trim (a value chosen to include as many observations as possible, while still being compatible with removing outliers). Other accuracy measures such as the symmetric MAPE and the relative absolute error of FFC compared to SFC have been calculated in preliminary work, but tell the same tale and are therefore not reported.

A summary of accuracy for the groups of organisations is given in Table 3, where the errors of the system forecast, the final forecasts, and a naïve

Table 3
Forecast error by organisational group and the direction of adjustment.

Org. Group		No. of Observations, N (naïve)	Trimmed MAPE			MdAPE		
			Naïve	System forecast	Final forecast	Naïve	System forecast	Final forecast
A–C	No adjustment	3,174 (3,068)	37.5	60.5	60.5	22.2	13.6	13.6
	Positive adjust	4,013 (3,735)	39.8	29.1	39.6	25.8	20.2	17.6
	Negative adjust	3,392 (3,136)	41.4	46.9	26.6	25.2	20.7	15.7
D1& D2	No adjustment	50,427 (49,794)	17.5	19.3	19.3	12.7	13.5	13.5
	Positive adjust	3,049	27.6	32.1	64.9	20.9	21.2	43.2
	Negative adjust	2,409	24.5	40.9	28.5	14.7	25.0	20.9

(random walk) forecast are compared for each type of adjustment. The differences between the MAPEs and MdAPEs indicate that the absolute percentage errors are right skewed, with some extremely large errors present in the data, even after the 2% trimming. Because of the skew in the trimmed MAPE as a result of these extremes, we will focus on the interpretation of the MdAPE, noting any contrasting results.

For both groups of companies, a comparison of the MdAPEs of the unadjusted system forecasts with those of the system forecasts that were subsequently adjusted suggests that forecasters were able to identify the system forecasts that were most in need of adjustment, a result which is consistent with findings of Willemain (1991) and Mathews and Diamantopoulos (1990), and supports hypothesis H_{2a}. The poor performance of those forecasts left unadjusted for companies A–C, as measured by the trimmed mean, suggests that there were, however, a number of forecasts ($> 1\%$) where the forecasters failed to recognize the need to adjust.

Wilcoxon's signed paired rank tests were used to compare the APEs of the system and adjusted final forecasts for each group of companies, both overall and for the different directions of adjustment. Whilst there was a significant improvement overall for companies A–C (supporting H_1), there was no significant improvement where the adjustment was positive. Moreover, for the retailer (D1 & D2), adjustments did not significantly improve the accuracy at all. Indeed, in this case both the system and final forecasts were less accurate than the naïve forecasts. Lawrence et al. (2000) also found that the naïve forecasts were often the most accurate in their study of judgemental sales forecasting in Australian companies. Thus, while the forecasters in all companies could apparently identify situations where judgmental adjustments were most needed, only negative adjustments in the A–C companies significantly improved accuracy. The finding that negative adjustments are more effective is consistent with the results of the study of products subject to intermittent demand conducted by Syntetos et al. (in press).

4.1.1. Effects of size of adjustments and series volatility

We now investigate whether the accuracy resulting from judgmental adjustments is affected by either the size of the adjustments (H_{2b}) or the volatility of the series (H_3). We measured the size of the adjustment by its absolute size relative to the system forecast, defined as:

$$100 * \frac{|\text{Final forecast} - \text{System forecast}|}{\text{System forecast}}.$$

Sanders and Ritzman (1992) measured volatility using the coefficient of variation of the raw data. However, we are particularly interested in volatility which arises as a result of special circumstances, as this is where the judgemental adjustment is most needed. Further, we wish to eliminate volatility which is capable of being forecasted, as the adjustment process starts with the statistical forecast. Thus the volatility measure used is based on the coefficient of variation of the system forecast absolute error. Each series was categorised as belonging to either a high or low volatility group, depending on whether this measure was less or greater than the median value for that organisation. Finally, we define our measure of the effect of judgmental adjustments on improving accuracy as

$$FCIMP = 100 * \frac{(|\text{Actual} - \text{System forecast}| - |\text{Actual} - \text{Final forecast}|)}{\text{Actual}},$$

i.e. the difference between the absolute percentage forecast error from the system forecast and the absolute percentage final forecast error. Note that this variable is positive when the final forecast is more accurate than the system forecast (i.e. the adjustment has improved the system forecast) and negative when the adjustment has reduced the accuracy. Mathews and Diamantopoulos (1987) proposed certain alternative measures, where they normalise the numerator above by, alternatively, the system forecast or the adjustment. The measure adopted above has the advantage that it measures the improvement or degradation in MAPE introduced by the adjustment directly. For firms A–C, an ANOVA with the size of adjustment (split into quartiles), volatility and firm as explanatory variables showed firm as significant ($p = 0.023$), volatility as highly significant ($p = 0.002$), and size of adjustment as highly significant ($p < 0.001$). For the retailer (D1 & D2) the ANOVA results showed that firm, volatility and size were all highly significant. The results held both when extreme adjustments were eliminated and when volatility was measured by the standard deviation of the absolute system error relative to the mean actual.

As hypothesised in H_{2a}, the size of judgmental adjustments is positively associated with the size of accuracy improvements, and Fig. 2 shows the median improvement in absolute percentage error for different sizes of relative adjustment. For the positive adjustments, the smallest 25% of adjustments on average lead to lower forecast accuracy for groups A–C, but for the retailer (D1 & D2) positive adjustments reduce accuracy, on average, irrespective of their size. Negative adjustments pay off for all companies (apart from the largest adjustments by the retailer). Examining the means tells a similar story, although positive adjustments for the manufacturers have negative consequences for all sizes. Interestingly, these results are not entirely consistent with those of Syntetos et al.'s (in press) study of products subject to intermittent demand, which found that small adjustments to forecasts of zero demand are likely to be beneficial.

The results do not support H_3. While there is a significant association between volatility and forecast improvement, the improvements are greater for low volatility series, contrary to H_3. It seems that volatile series are more difficult to forecast, either because their 'jumpiness' is not associated with foreseeable events or because the forecaster has difficulty in accurately assessing the effect of events which are known to be occurring in the future.

4.2. Unbiasedness

All of the organisations studied aim to minimize the forecast error. An equal propensity to over- and under-forecast would suggest that the (trimmed) means and medians of the percentage final forecast error (PE) are zero (hypothesis H_4). These measures are displayed in Table 4. For both groups, whatever the direction of the adjustment, the distribution was not centred at zero. This also applies to the individual companies. There is a clear contrast between negative and positive adjustments. In all of the companies, negative adjustments are effective in that they tend to reduce the mean bias in system forecasts that are too high. It appears that, once they have decided to make a negative adjustment, forecasters are realistic about the likely levels of demand. Positive adjustments, on the other hand, tend to lead to final forecasts that overestimate (i.e., they are too optimistic). This tendency can also be seen in Table 5a, which shows that 66% of positive adjustments for companies A–C and 83% of those made by D1 and D2 led to forecasts that were too high. The retailer's positive adjustments are applied to system forecasts that are already, on average, too high (according to the mean percentage error), and therefore serve only to exacerbate the bias. Only 46% of the negative adjustments made by the two groups resulted in final forecasts that were too high (Table 5b).

Overall, for both groups of companies, the forecasts suffer from mean bias (using a binomial test) with the magnitude of bias often being large, supporting H_4. To test for both mean and regression bias (H_4), we estimated the following regression model:

$$Y_{ij,t} - F_{ij,t-1}(1) = \alpha_i + \beta_i F_{ij,t-1}(1) + \nu_{ijt}, \qquad (M1)$$

where $Y_{ij,t}$ represents the actual sales in the ith company for the jth SKU in period t, $F_{ij,t}(1)$ is

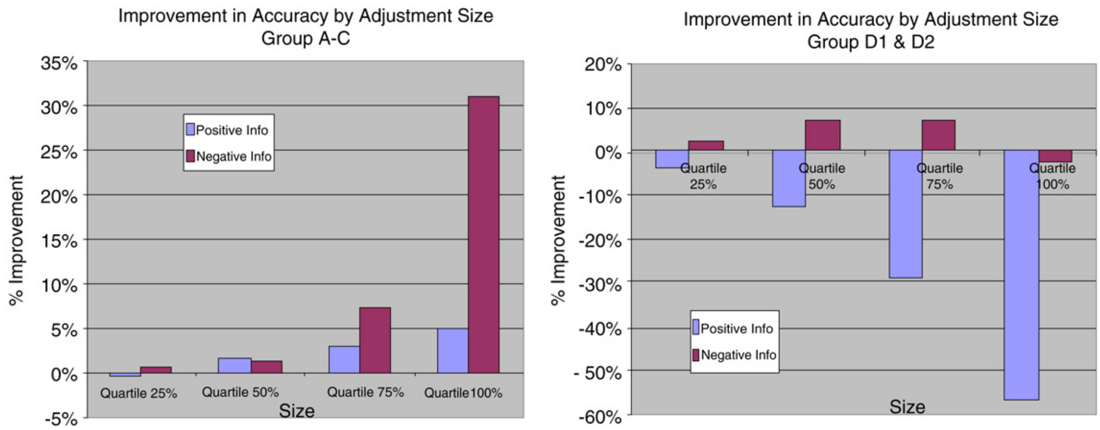

Fig. 2. Effect of size on improvements in forecast accuracy.

Table 4
Mean bias in adjusted forecasts by organisational group and direction of adjustment.

Org. group	Direction of adjustment		Mean percentage errors			Median percentage errors		
			Naïve	System forecast	Final forecast	Naïve	System forecast	Final forecast
A–C	No adjustment	3,174	−9.7	−44.3	−44.3	2.9	−1.4	−1.4
	Positive adjust	4,013	−9.0	5.7	−29.6	2.5	9.8	−9.4
	Negative adjust	3,392	−15.5	−38.1	−4.5	1.0	−14.0	2.2
D1 & D2	No adjustment	50,427	3.2	−4.6	−4.6	−0.1	−1.4	−1.4
	Positive adjust	3,049	10.1	−7.4	−57.5	12.1	0.0	−39.2
	Negative adjust	2,409	−1.8	−35.5	−3.5	4.8	−22.3	3.6

Table 5a
Evidence of optimism bias in positive adjustments.

Org. group	No. of observations	% of times positive adjustment is too large	% of times positive adjustment is in wrong direction	Total % of positive adjustments that are overoptimistic
A–C	4013	32.0	34.4	66.4
D1 & D2	3049	32.4	50.6	83.0

Table 5b
Evidence of optimism bias in negative adjustments.

Org. group	No. of observations	% of times negative adjustment is too large	% of times negative adjustment is in wrong direction	Total % of negative adjustments that are overoptimistic
A–C	3392	25.2	28.3	46.5
D1 & D2	2409	33.2	20.5	46.3

the one period ahead final forecast made at period t, and ν_{ijt} is random error. If the company forecasts are unbiased in aggregate, $\alpha_i = \beta_i = 0$ for each company i. The normal regression assumptions cannot be expected to hold, in that the data have very different levels and the errors can therefore be assumed to depend on the level of $Y_{ij,t}$, as well as the noise. The variables have therefore been normalised using the standard deviation of sales for each SKU. In addition, the data have been ordered by size of adjustment (in percentage terms) into three groups, and models have been estimated for each sub-group.

The sub-sample of forecasts that have been adjusted contains many extreme observations. For example, 1.9% of observations have forecast adjustments greater than 250%. We have removed all observations with such large values from the model building. In addition, when estimating model ($M1$) we have removed outliers (with absolute studentized residuals greater than 2.5) and high leverage points, to ensure that such points do not exert 'undue' influence on the regression coefficients. Sensitivity testing was carried out on these data filtering decisions, and the results we present appeared robust. After the adjustment procedure had been applied, the residuals proved well-behaved (with no evidence of non-normality and heteroscedasticity for most models), demonstrating the effectiveness of the normalisation process.

Initially the model was estimated jointly, with the parameters (and error distribution) assumed to be constant across companies. Using a general linear model, with the companies as factors, and $F_{ij,t-1}$ as the covariate, together with an interaction between them, leads to a rejection of the hypothesis that the biases are independent of company. The two separate sets of forecasts available for company D were also tested for equality of the regression parameters, while the other three companies were examined in a pairwise fashion, again leading to a formal rejection of statistical equivalence, despite face similarities. From our analysis of the summary statistics on bias, there was also evidence that the direction of the adjustment influences the magnitude of bias in the adjustments. We have therefore estimated models both for the direction of adjustment and, separately, for the individual companies as well as the groups.

Table 6 shows the results of estimating the equations for the two groups. For both groups (and all companies), whatever the direction of the adjustment, H_4 (the forecasts are biased) is supported. The small values of R^2 and β for companies A to C suggest that bias makes a relatively small contribution to the forecast errors, though its elimination is still likely to be worthwhile. The table shows that, in general, the constant term is positive, but that all the β coefficients of the final forecasts are negative. At the mean level of the final forecasts this leads to an upwards bias, in that the final forecast tends to be too high. The relative importance of the mean bias compared to the regression bias can be established using Theil's decomposition of the Mean Squared Error,

$$MSE = (\bar{Y} - \bar{F})^2 + (S_F - \rho_{Y,F} S_A)^2 + (1 - \rho_{Y,F} S_Y)^2$$

where F is the forecast of Y, $\bar{F}$ and $\bar{Y}$ are the respective means, with standard deviations S_F and S_Y, and $\rho_{Y,F}$ is the correlation between the two. The first term represents the mean bias, whilst the second measures the regression bias. Table 6 also shows the ratio of these two biases. There is no evidence of a consistent pattern where one type of bias is more prevalent than the other. Overall, the results confirm that bias contributes more to forecast inaccuracy when adjustments are positive.

Table 6
Regression and mean forecast bias by organisational group and direction of adjustment.

Org. group	Coefficient α	Coefficient β	R^2 (%)	Theil's decomposition Mean Bias/ Regression (%)
A–C				
- Positive adjust	0.103	−0.118	6.5	55
- Negative adjust	0.259	−0.059	2.4	319
D1&D2				
- Positive adjust	0.264	−0.366	40	17
- Negative adjust	0.497	−0.188	10	49

All coefficients significant at the <0.01% level.

Why were the forecasters applying positive adjustments so poorly? Discussions with the companies suggested that the reasons differed between companies A–C and the retailer (D1 & D2). In the latter case it emerged that, for many products, managers were confusing forecasts of demand with decisions on the levels of inventory required to meet customer service levels. For example, a forecast of 200 units might be adjusted upwards to 250 units so that the probability of a stock out was reduced to a level perceived to be acceptable. The managers were unaware of this confusion. Clearly there was a danger that their estimates, labelled as 'forecasts', were liable to misinterpretation, or even to further upward adjustments by others.

There was no evidence that this confusion applied to companies A–C. Here, there were two plausible explanations for the bias associated with positive adjustments: over-optimism and the mistiming of estimates of promotion effects. To identify the most likely explanation, we need to consider the two conditions under which positive adjustments can lead to forecasts that are too high. Either they can be made when a negative adjustment was actually required (i.e., the adjustment will be in the wrong direction), or they can be adjustments which are too large, despite being in the right direction.

A high percentage of positive adjustments were made in the wrong direction (Table 5a). These adjustments add substantially to the forecast errors, particularly when they are large. They might be caused by the effects of some promotion campaigns occurring later than the forecaster expected, so that an expected upward movement in demand does not materialise in the early stage of the campaign. However, if this is the case then one would expect subsequent positive adjustments to be in the correct direction. Our data provides little support for this. For example, for companies A–C, 43.3% of upwards adjustments that were made following a wrong side adjustment were also wrong sided. This suggests that unwarranted optimism is the cause of many positive adjustments, rather than the mistiming of adjustments.

4.3. *Efficiency*

With bias established, the next hypothesis to examine is that of efficiency. The most immediate data the forecaster can bring to bear in making the adjustment are the time series history, the latest system forecast and the most recent forecast errors. If the forecast is efficient, the forecast error should not be predictable by variables known to the forecaster. For all of the companies, the latest observation is known provisionally at the time of making the forecast. A suitable test of efficiency with this information set is the following model:

Define

$$\begin{aligned} e_{ij,t} &= Y_{ij,t} - F_{ij,t-1}(1) \\ &= \alpha_i SFC_{ij,t-1}(1) + \beta_{i,1} Y_{ij,t-1} + \beta_{i,2} Y_{ij,t-2} \\ &\quad + \gamma_1 e_{ij,t-1} + \gamma_2 e_{ij,t-2} + \nu_{ij,t}, \qquad (M2) \end{aligned}$$

where $SFC_{ij,t}$ is the system forecast made at t. The constant term has been suppressed, as the model's objective is to re-weight and combine the available information to explain the observed error. In the model, a significant coefficient for an explanatory variable indicates that the forecaster has not used the information represented by that variable efficiently, and that by re-weighting that variable, either judgmentally or statistically, the forecast could be improved. Before the final estimated equation can be established, the outliers must again be removed, and the errors rendered homoscedastic via normalisation (where, as before, the standard deviation of the actuals has been used). There is some limited evidence of seasonality in the errors, but its removal through the inclusion of dummy variables affects the results only slightly.

Due to space limitations, we present summary regressions for the two groups of firms A–C and D1 & D2. Both groups and all companies show signs of inefficiency, in that their forecast errors can be reduced by better weighting the available information on the past observations and forecast errors (see Table 7). Typically, the current forecast assigns too little weight to the latest observed error, possibly due to the fact that it is not always known exactly at the time the forecast is made. For firms A–C, there is some slight evidence that the most recent observation is mis-weighted. In addition, the system forecast is over-weighted. Interestingly, this result differs from that of Goodwin and Fildes (1999), who found in a laboratory study that forecasters ignored the statistical forecast when making judgmental interventions to take promotion effects into account. Overall, the results confirm those derived from examining bias: there is more inefficiency shown by the forecasters when they have made positive adjustments to the final forecasts.

We also attempted to pool estimates using a general linear model, but this leads to rejection of the hypothesis that the error model is constant across companies. No uniform pattern of response is observed across the companies. For example, company B fails to take into account both most recent sales and its previous forecast error. Company C shows the least sign of significant inefficiencies. The retailer D shows major inefficiencies, particularly with regard to the most recent errors.

Table 7
The efficiency of the final forecasts by company and adjustment direction.

Organisational group		No. observations	Model coefficients					R^2 (%)
			System forecast	Lag 1 Actual	Lag 2 Actual	Lag 1 Error	Lag 2 Error	
A–C	Positive adjust	2868	−0.120 (8.4)	0.028 (2.0)	n.s.	0.135 (8.5)	0.072 (4.7)	20
	Negative adjust	2427	−0.106 (10.2)	0.107 (7.6)	0.029 (2.2)	n.s.	n.s.	7
D1 & D2	Positive adjust	2566	n.s.	n.s.	−0.108 (4.5)	0.585 (31.7)	n.s.	76
	Negative adjust	2101	n.s.	n.s.	0.104 (5.9)	0.369 (19.4)	0.046 (3.6)	41

t statistics are in parentheses: n.s. implies that the coefficient is not significant at the 5% level.

From the analysis so far presented, we have found that final forecast errors (and the judgmental adjustments) are biased, inefficient, and affected by the direction and size of the adjustments. This leads to the core question of whether these features can be exploited to achieve more accurate forecasts.

5. Reaping the benefits of expert adjustment

In this section we evaluate four possible methods for improving the accuracy of the adjustments, based on the biases and inefficiencies we have identified. The first two methods use a statistical procedure to correct the adjustments. The others require training for forecasters, better market information, or restrictiveness within the forecasting software.

5.1. The Blattberg–Hoch approach

The '50% model, 50% manager' heuristic proposed by Blattberg and Hoch (1990) involves taking the mean of independent management judgmental and statistical forecasts. Blattberg and Hoch found that this simple strategy was successful because of the complementary strengths and weaknesses of statistical models and human judges. For example, models can consistently attach optimal weights to large volumes of data, while managers can recognise and interpret abnormal situations. In our case, application of the heuristic will involve taking a mean of the system and final forecasts. However, because the managers saw the system forecasts (indeed, they typically over-weight them) before making their adjustments, the system and final forecasts are not independent. This means that the heuristic will lead to a forecast which is equal to:

0.5 (System forecast) + 0.5 (System forecast + Adjustment) = System forecast + 0.5 Adjustment.

Thus, the method will simply act as a damper on the adjustments, restricting them to 50% of the change indicated by the managers. When there is a propensity to over-adjust forecasts (e.g. as a result of an optimism bias), this damping might improve accuracy. However, if the forecasters are anchoring on the statistical forecast and conforming to the anchor and adjustment heuristic (Tversky & Kahneman, 1974), they will be under-adjusting from the statistical forecast, and the damping is likely to be too severe.

5.2. Error bootstrap rules

Bootstrap rules, which model the relationship between the judgmental forecast and the available cue variables, have outperformed unadjusted raw judgments in many studies (Dawes, Faust, & Meehl, 1989). This is because these models average out the inconsistencies in the raw judgments. However, the evidence is primarily based on cross-sectional forecasting problems where the cues are not autocorrelated, as in time series data, and, even in these situations, bootstrapping has not always outperformed raw judgments. For example, Astebro and Elhedhli (2006) found that their linear additive bootstrap model performed worse than experts' forecasts of whether R&D projects would subsequently be commercialised because it was not able to represent their non-compensatory use of cues.

The time series evidence is much more limited (Lawrence & O'Connor, 1996; Lawrence et al., 2000). This may be partly because the serial correlation in the cues implies that there is a high degree of redundancy in the available information, a situation which is likely to enhance the quality of forecasts based on raw judgments relative to those of the bootstrap model. An alternative way in which the cue information can be used to improve on judgment in forecasting was demonstrated by Fildes (1991). He showed that modelling the relationship between forecast errors and cues enabled forecasts to be corrected for the mis-weighting of both a causal driver (GDP) and past actuals. This resulted in improved accuracy in out-of-sample data. Given that the final forecasts in our companies are inefficient and biased, this approach (which we will refer to as 'error bootstrapping') seems much more likely to result in improvements than conventional bootstrapping.

We have developed various models in order to examine whether the observed inefficiencies can be used to improve accuracy. Although lags of two periods were shown to be significant in the earlier efficiency analysis shown in Table 7, the longer lags typically have a low impact on the models' standard errors, once the most recent data is forced into the model, and the additional complexity of such rules seemed likely to limit their operational impact. The model estimated here combines the various raw information sources using only one period lags – the most recent data available to the forecasters.

$$Y_{ij,t} = \lambda_{1j} SFC_{ij,t-1}(1) + \lambda_{2j} Adj_{ij,t-1} + \beta_{1j} Y_{ij,t-1} + \gamma_{1j} e_{ij,t-1} + \nu_{ij,t}, \qquad (M3)$$

or alternatively

$$e_{ij,t} = (1+\lambda_{1j}) SFC_{ij,t-1}(1) + (1+\lambda_{2j}) Adj_{ij,t-1} + \beta_{1j} Y_{ij,t-1} + \gamma_{1j} e_{ij,t-1} + \nu_{ij,t},$$

where $SFC_{ij,t-1}(1)$ is the one-period ahead system forecast for the ith product, and the jth company, made in period $t-1$; Adj_{t-1} is the size of the judgmental adjustment made by the forecaster; and e_{t-1} is the last period's error. Once (M3) is estimated (for each company), it can be used to produce forecasts of demand.

To ensure a rigorous evaluation of the proposed models, the database was split into an estimation set of approximately 80% of the total data set for each company, and a test set of the remainder. For example, for Company A, the last 7 months constituted the test data. This design of a hold-out sample is more demanding for the model than the alternative of selecting 20% of the data at random as the hold-out sample. Model performance has been evaluated again using a trimmed mean (trimmed by removing 2% of the extreme final forecast errors). The test of equality of the model coefficients for positive and negative information was highly significant, and therefore the two classes of data have again been modelled separately.

Various different models were estimated, based on the size of the adjustments made, since preliminary analysis had shown that the size of the adjustment affected the model coefficients (e.g., set Adjustment = 0 for small values). However, fitting the default model to the entire data set (apart from outliers) proved the most successful in terms of improved accuracy. We have analysed two models: (1) the full model, incorporating the system forecasts, size of adjustment, past values of the errors and recent observations as cues; and (2) the optimal adjustment model, which uses only the system forecasts and size of adjustment. Error bootstrap models are shown in the Appendix for the two organisational groups. Full details of all of the individual company models are available from the corresponding author. They show that the companies differ, often quite substantially, in their ability to use information in their environment to achieve optimally accurate forecasts. However, a number of traits were common to all companies; in particular, for the A–C group of companies the past actuals and errors (whilst significant) were unimportant, in contrast to the retailer (D1 & D2). Second, negative forecast adjustments were almost optimal in companies A–C.

5.3. Avoiding small adjustments

The results displayed in Fig. 2 show that small adjustments were often ineffective in reducing forecast error. We argue that this was because they tended to be made when information was unreliable and/or the anticipated effects of special events were small, or because forecasters were either seeing false patterns in noise or were tweaking the forecasts to justify their role. This suggests that a strategy which stops forecasters from making these smaller adjustments would marginally enhance accuracy, but also, more importantly, it would free up the time of those involved in the forecasting process. There are a number of ways in which such a strategy could be implemented, including the training of forecasters and the use of software that prohibits adjustments below a pre-set percentage. We examined the potential accuracy gains that could be achieved through a strategy of avoiding adjustments smaller than 20% (some 42% of all adjustments) to see whether it would be worth implementing.

5.4. *Avoiding wrong-sided adjustments*

Our earlier analysis showed that wrong-sided adjustments are particularly damaging to forecast accuracy. We examined the potential gains that could be achieved if improvements in market intelligence could be used to eliminate *half* of these wrong sided adjustments. We did this by randomly allocating each wrong sided adjustment to either a *no change* group or a *change* group. Each member of the *no change* group had its adjustment set to zero, so that its forecast became the system forecast.

5.5. *A comparison of the methods*

We compared the accuracy of methods on both the in-sample and hold-out data using both the MAPE and MdAPE. We also ranked the results for each company (1 being most accurate), and summed the ranks. Of the two error bootstraps, the full bootstrap was the more accurate in-sample, so we present these results. Table 8 summarises the hold-out sample results for the A–C and D1 & D2 companies. The in-sample results are similar, and demonstrate the robustness of the methods.

It can be seen from the hold-out results that there are clear differences in the accuracy of the methods depending on whether the adjustment is positive or negative. When positive adjustments are made, the strategy of preventing forecasters from making smaller adjustments is ineffective. Because over-large positive adjustments are the main problem, preventing small positive adjustments offers little improvement. However, the other methods *are* effective when adjustments are positive. Rather than removing the small adjustments, the Blattberg–Hoch method, as applied here, acts by damping *all* adjustments by 50%. This method therefore improves accuracy by reducing the damaging impact of the large adjustments. Wrong-sided adjustments are also frequently associated with positive adjustments (especially for the D1&D2 group—see Table 5a). Not surprisingly, developing measures to prevent half of these wrong-sided adjustments also leads to substantial improvements over the final forecasts. However, the optimal adjustment model performs best, reflecting the serious errors associated with forecasters' estimates of the sizes of required adjustments when the direction of adjustment is positive. For the D1&D2 group, the improvements it yields over the final forecasts are substantial.

When information is negative, a contrasting set of results is obtained. In this case, comparison between the MAPEs and MdAPEs suggests that the judgmental adjustments are particularly effective because they are reducing many of the extreme errors in the system forecasts. However, there is a slight tendency to under-adjust: around 54% of negative adjustments are too small in both groups of companies (Table 5b). Table 8 shows that two of the improvement methods are ineffective at best when adjustments are negative. Indeed, if measures are taken to remove the smaller adjustments, accuracy will actually be reduced (especially for the D1 & D2 group). This is because the measure would eliminate those small adjustments, which are nevertheless in the correct direction. Similarly, by damping the size of the adjustments, the Blattberg–Hoch method only exacerbates the tendency to under-adjust. (The small improvement in the MdAPE arises from just one of the companies.) Preventing half of the wrong-sided adjustments is bound, by definition, to lead to improved accuracy. However, when information is negative there tend to be fewer wrong-sided adjustments (see Table 5b), so the benefits derived from measures designed to eliminate half of them are less than those obtained with positive information. The optimal adjustment model does not yield improvements for companies A–C, reflecting the fact that forecasters in these companies made negative adjustments which were highly accurate, but it does lead to improved accuracy for companies D1 & D2, where there was a confusion between forecasts and inventory decisions.

Table 8
Hold-out sample accuracy of forecast 'improvement' methods.

Org. group	Adjustment	Accuracy measures	System forecast	Final forecast	Blattberg-Hoch	Full model	Avoid smallest adjustments	Remove 50% of wrong side adjustments
A–C	Positive ($n = 639$)	MAPE	32.3%	42.5%	38.1%	32.1%	41.4%	35.0%
		MdAPE	21.0%	21.7%	18.1%	19.9%	19.9%	18.0%
		Sum of ranks	20	33	19	13	28	14
	Negative ($n = 720$)	MAPE	52.6%	29.8%	49.3%	29.6%	32.7%	28.6%
		MdAPE	19.2%	18.3%	16.9%	18.4%	16.0%	17.0%
		Sum of ranks	34	17	28	16	22	9
D1 & D2	Positive ($n = 627$)	MAPE	30.3%	47.7%	38.5%	35.5%	46.2%	40.0%
		MdAPE	26.1%	41.8%	31.1%	30.3%	40.2%	32.9%
		Sum of ranks	8	24	14	9	20	14
	Negative ($n = 612$)	MAPE	35.2%	23.6%	27.8%	18.9%	25.5%	21.7%
		MdAPE	28.0%	19.4%	23.2%	11.8%	20.9%	18.0%
		Sum of ranks	18	17	14	5	19	11

n = number of observations.

In summary, preventing smaller adjustments is not likely to lead to improvements in accuracy, whatever the information direction. The other methods are effective when information is positive. When it is negative they lead to smaller improvements in accuracy or none at all. As a strategy, the optimal adjustment procedure works well overall; it generally leads to improved accuracy, sometimes substantial, and with no damaging effects for either positive or negative adjustments for either group of companies.

6. Improving forecasting performance in practice

Our analysis has revealed major differences in the forecasting accuracies obtained by the companies. It also showed the potential for improvements that could be achieved by focusing on the more effective use of the available information and the removal of consistent biases. However, the methods of improvement that we examined would face possible obstacles if an attempt was made to implement them in many organisations. Because the Blattberg–Hoch and error bootstrap approaches are automatic correction procedures, their use may lead to the demotivation of forecasters, with less effort being applied to the original judgments (Belton & Goodwin, 1996). Alternatively, the nature of the biases may change over time, possibly as a result of training, or the company forecasters may seek to pre-empt the corrections by distorting their judgmental inputs into the process. Improvements in forecasting through training and better use of market intelligence, which are required by the other two improvement methods we examined, are also not straightforward to achieve, and there are many barriers to adopting new forecasting procedures (Schultz, 1984). In this section we use our experience of extensive meetings with the companies concerned and observation of their forecasting processes to provide explanations for the results and to identify how improvements might be achieved. We held between 5 and 11 meetings with staff at each company. These included semi-structured interviews with individual company personnel, formal presentations by managers of their forecasting processes, forecast review meetings (involving at least five staff), and discussions where we debriefed staff on the findings of our research. The staff we met were forecasters, logistics and supply chain managers, accountants, and sales and marketing personnel. At least two observers (or interviewers) were involved in the meetings, which were also usually tape-recorded.

6.1. The forecasters

Fildes and Hastings (1994) and Moon et al. (2003) identified motivation and training as potentially important in attaining accurate forecasts. Despite such a conclusion being apparently obvious, in none of the companies were the forecasters knowledgeable in the statistical aspects of forecasting such as error measurement or alternative forecasting methods. Nor were they aware of the many biases associated with judgmental interventions in forecasting. All the senior forecasters were, however, immersed in process and management issues relating to forecasting, with one acting as a regular presenter at professional forecasting events. It therefore appears that training and the use of appropriately targeted incentives might lead to improvements in accuracy.

6.2. The forecasting support system (FSS)

All four company systems were professionally developed, but they had inflexible interfaces and poor (or in one case non-existent) graphics. However, the format of the interface can be important in improving accuracy (Tashman & Hoover, 2001). The system forecasts were produced using models far removed from best practice (see, for example, Gardner & Anderson, 1997), nor had the chosen methods been tuned to produce the best possible results from the software. Standard exponential smoothing models are now known to require just such tuning in the choice of smoothing parameters, so this might explain the inadequacies (Gardner, 2006). This was underlined in the strong performance of the naïve forecast compared to the system forecast for companies B and D. In company A, perceived inadequacies in the automatic system forecasts led to a complex model-fitting process (see Goodwin, Lee, Fildes, Nikolopoulos, & Lawrence, 2007) and, typically, system forecasts that were less accurate than an automatic alternative. An analysis of screen displays shows that the forecasters did not have clear guidance on the previous actuals and previous errors. Summary error measures were not easily available, and those that were provided were subject to outlier and intermittent demand effects. Thus, a reliable assessment of the gains or losses in accuracy resulting from judgmental adjustments could not be made, and there was therefore little opportunity to learn from experience about the appropriateness of judgmental intervention in different circumstances. Improved statistical forecasting systems to provide better baseline forecasts, and accuracy monitoring systems using well-designed error measures, are therefore needed.

Although three of the systems had a ‘notes’ facility, whereby the forecaster could explain the reasons for the adjustments they made, they had none of the features that might make it easy to use and effective (Lee et al., 2007). Their use was spasmodic and incoherent, in that forecasters could not explain the past adjustments they had made to us by referring to their ‘notes’ system. Clearly, requiring forecasters to record reasons for their adjustments in a standard format (e.g. by selecting a reason from a list) might serve to reduce the number of relatively small, but damaging adjustments that may be based on misinterpreting noise as signal, or reflect gratuitous tweaking of the forecasts (Goodwin, 2000). A list of reasons would also allow forecasters to understand why and how market intelligence is so often misinterpreted. In addition, it would assist in the decomposition of market intelligence into key drivers, thereby lessening the likelihood of double counting. Finally, experimental evidence suggests that the incorporation of guidance systems such as those which allow the formal use of analogies (e.g. past promotions and their effects, see Lee et al., 2007), would improve the quality of judgments based on market intelligence.

6.3. Communicating and compiling the market intelligence

Forecasters in companies B, C and D identified promotions as the most important driver of their judgmental adjustments. Other important drivers included price changes (companies C & D), the weather (B & D), and inventories (A & D). Where market intelligence is strong and the direction of its effect is clear, there are major potential improvements in accuracy, as seen in the greater accuracy achieved though negative adjustments. However, the process by which such intelligence is gained, as Moon et al. (2003) point out in an examination of 16 case studies, is often very flawed, primarily through the lack of coordination and communication between the different organisational units involved in supply chain operations, sales and marketing. Here, while all companies apparently consult widely on important drivers, the evidence that is collected is not compiled through, for example, a database which could lead to learning by analogy from earlier exemplars. Each forthcoming event is, instead, treated as unique.

In addition, different sources of intelligence are not identified and quantified separately using a decomposition approach. This can lead to double counting or omission (MacGregor, 2001). None of the companies attempted to review the reasons for wrongly interpreting the direction of an adjustment, which, as we have seen, is a cause of major forecast error.

7. Limitations of the study and an agenda for future research

Despite the large number of forecasts and adjustments that we examined, our study has some limitations. First, the data was collected from just four UK-based companies and were one step ahead forecasts. Although they covered a range of industries, all of these organisations were similar in the expertise of their forecasting staff, the forecasting processes that they adopted, and the type of software that they used. In particular, given that the managers in our organisations had minimal training in forecasting, we cannot be sure that our results will apply to companies where forecasts are produced by more highly trained staff. Moreover, the manufacturers were not involved in collaborative forecasting with either their suppliers or their customers; the retailer's forecasts were however shared with some of their suppliers. A large body of recent research has attempted to assess the benefits of information sharing in supply chains (see for example, Helms, Ettkin, & Chapman, 2000, for an introduction and Smaros, 2007, for a description of some of the difficulties of collaboration); however, much of this research has involved simulations or mathematical modelling of simplified theoretical supply chains (e.g. Aviv, 2001). Clearly, future research needs to assess the benefits of information sharing in real supply chains. Despite all of these reservations, the results of our recent surveys of forecasters (Fildes & Goodwin, 2007), as well as the study by Moon et al. (2003), and our own observations in other companies, suggest that the forecasting practices we studied are typical of those currently found in a large number of supply-chain based companies.

Second, there is arguably no such thing as an objective statistical forecast. All forecasting involves some judgment, at least in the choice of the statistical method used and the length of the data history to which it is fitted. In some of our companies the forecasters had often already intervened in the statistical forecasting process to change the length of the data history or the smoothing parameters. In addition, the process of cleaning past data to remove the effect of unusual events was judgmental. Thus the judgmental adjustments were applied to statistical forecasts that already contained an element of the forecasters' judgment. In the future it would be interesting to compare the merits of adjustments to statistical forecasts that have, as far as possible, been generated automatically to those applied to forecasts that have already been subjected to significant intervention (Önkal, Gönül, & Lawrence, 2007).

Third, our finding that forecast adjustment reflects a general over-optimism needs further investigation to determine its underlying causes. Although the forecasters in the non-retail companies assured us that their objective was to produce as accurate forecasts as possible, our results do not allow us to assess the extent to which this bias resulted from overt or latent incentives to over-forecast, or whether it is a cognitive bias associated with the forecasting process.

Our results point to a number of additional key areas where future research efforts need to be directed. The statistical models we have examined have been based on cross-sectional analysis, pooling models across companies. With a larger database, time series cross-sectional models could be examined with a view to analysing the performance of products individually and producing even more effective combined forecasts. Such an analysis would deliver answers to questions such as, do the adjustments made to 'important' products prove more effective? Is there learning taking place from adjustment to adjustment?

There is clearly a need for much more research into the design of forecasting software. Software developments over recent decades have focussed on improving data handling and input–output processes. Management judgment, which, as we have seen, is a crucial component of demand forecasts in supply chain companies, remains relatively unsupported (Fildes, Goodwin, & Lawrence, 2006). In this regard, we think that there is scope for developing and testing software facilities that allow advice and information obtained from multiple sources to be used and combined in a structured way. Our early work on the use of a database of analogies to support estimates of the effects of special events, such as past promotions (Lee et al., 2007), could be extended to include a database of profiles of the week-to-week effects of sales promotions over the course of the campaign. In addition, the evidence from the retailer (company D) suggests the need for a decoupled system which clearly distinguishes between forecasting and the associated supply chain decisions. A system which allows the two to be compounded means that beneficial judgmental interventions are difficult to apply because there is no clear view of either forecast error or optimal inventories. Moreover, given the widespread use of forecasting review meetings, there is also scope for the development of group forecasting support systems which allow managers to feed independent estimates of required adjustments into the system. Such improvements in system design might also help to mitigate the pressures towards bias, both personal and organisational, that exists in many companies.

However, none of these developments will be effective if such facilities in systems are not demanded and then used by forecasters, perhaps because of political or cultural factors, or because of a lack of training. This indicates the need for organisation-based studies that use interpretive research methods to establish, at a deep level, the beliefs and values of managers engaged in forecasting. Such research would need to explore both the psychological processes that individual managers employ and the effects of interactions between managers within organisational contexts. The results of these studies should encourage the successful implementation of effective company forecasting processes.

8. Conclusions

Judgmental adjustment of statistical demand forecasts for SKUs is very common, with up to 80% of forecasts adjusted in some companies. The accuracy of such forecasts and their associated judgmental adjustments is crucial to supply chain operations and planning, and has a direct impact on profit and service levels. Moreover, without reliable forecasts, collaboration between retailers and manufacturers is likely to be limited. Hitherto the effectiveness of these judgmental adjustments has been moot, with very limited empirical evidence available. In this paper we have shown conclusively that the value of these adjustments depends on the company context, but where the forecasters' principal motivation is towards improved accuracy, they can add substantially to forecast accuracy (hypothesis H_1).

However, the company forecasts we observed proved to be biased and inefficient (hypotheses H_4 and H_5). Positive adjustments were far less effective than negative, and forecasts that overestimated demand were prevalent. The forecasters tended to over-weight the statistical system's forecast, which for some of the companies was itself deficient in comparison to a naïve forecast. We therefore developed various models for capitalising on the biases and inefficiencies, showing that the most appropriate model depends on the circumstances, and in particular the nature of the adjustment, positive or negative. For example, the effectiveness of the Blattberg-Hoch heuristic of 50% model + 50% man, as we applied it, is itself limited to positive adjustments. Based on over 12,000 judgmentally adjusted forecasts and 1536 SKUs, we can therefore conclude that, at least for the companies analysed here, they could improve their forecasts substantially by more effective codification and incorporation of available information, such as market intelligence — the basis of most adjustments. In particular, companies should put into place processes to avoid 'optimism' bias and the confusion of forecasts and decisions, either as part of the FSS or in the motivation, training and monitoring aspects of the organisational forecasting process. Avoiding small adjustments (hypothesis H_{2b}) would free time to focus on what is undoubtedly the most important issue, identifying the direction of the effect of market intelligence. In addition, the statistical forecasting systems employed by many of the companies could be improved to deliver a better baseline from which to make the judgmental adjustments.

Forecast adjustment is the only practical way for most organisations to improve their incorporation of key drivers into their disaggregated sales forecasts. While the evidence we present shows the benefits of adjustment, there remain plenty of opportunities for

Table A.1
Error bootstrap model coefficients and R^2 values by organisational group.

Org. group	Adjustment direction		System forecast	Past actual	Past error	Adjustment	R^2
A–C		M3	0.842 (62.5)	0.169 (13.5)	n.s.	0.424 (22.6)	0.97
	Positive $n = 1367$	Optimal Adjust	0.987 (211)	*	*	0.450 (25.2)	0.97
		Blattberg–Hoch	1	*	*	0.5	*
		M3	0.827 (59.3)	0.181 (13.2)	n.s.	0.804 (32.0)	0.97
	Negative $n = 1087$	Optimal Adjust	0.995 (197)	*	*	0.989 (43.1)	0.96
		Blattberg–Hoch	1	*	*	0.5	*
D1-D2		M3	0.392 (17.2)	0.590 (24.1)	n.s.	0.203 (8.90)	0.97
	Positive $n = 2063$	Optimal Adjust	0.864 (122)	*	*	0.272 (17.9)	0.94
		Blattberg–Hoch	1	*	*	0.5	*
		M3	0.465 (17.4)	0.573 (18.7)	0.089 (4.14)	0.371 (13.8)	0.97
	Negative $n = 1797$	Optimal Adjust	0.913 (104)	*	*	0.622 (21.4)	0.94
		Blattberg–Hoch	1	*	*	0.5	*

t-statistics are in parentheses. Key: M3: The full model using all information; Optimal Adjust: The optimal adjustment model using just the adjustment and system forecast; Blattberg-Hoch: The 50-50 model weighting the system forecast and the adjusted final forecast equally.

both companies and researchers to understand how such factors are best included to overcome the biases and inefficiencies we have identified. The result should be major improvements in accuracy.

Acknowledgements

This research was supported by Engineering and Physical Sciences Research Council (EPSRC) grants GR/60198/01 and GR/60181/01. Alastair Robertson contributed substantially to the statistical modelling.

Appendix. Error bootstrap models to make more efficient use of the judgmental adjustments

Table A.1 shows the estimated cue models based on the ($M3$) equation, which uses the system forecast, past cues and the judgmental adjustment to estimate the actual demand. If the final forecast were the best achievable, the coefficients of the system forecast (SFC) and the judgmental adjustment (Adj) would both be 1, with zero weight assigned to past actuals and past errors. The second model only used the judgmental adjustment and the system forecast. The data have been grouped into the manufacturing companies (forecasting monthly) and the retailer (forecasting weekly). For the monthly data the drivers of the past actual and past error have limited or no predictive power. However, the forecasters over-respond to market intelligence and other information, particularly in the case where information is positive, and therefore damping the adjustment improves accuracy. For negative information the final forecasts are close to optimal (giving a weight close to 1 to the adjustment). For the retailer, neither the system forecast nor the adjustment is close to optimal, so the scope for improvement is large. The implied coefficients from the Blattberg-Hoch model are shown for comparison.

References

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods – empirical comparisons. *International Journal of Forecasting*, *8*, 69–80.

Armstrong, J. S., & Fildes, R. (1995). On the selection of error measures for comparisons among forecasting methods. *Journal of Forecasting*, *14*, 67–71.

Astebro, T., & Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, *52*, 395–409.

Aviv, Y. (2001). The effect of collaborative forecasting on supply chain performance. *Management Science*, *47*, 1326–1343.

Belton, V., & Goodwin, P. (1996). Remarks on the application of the analytic hierarchy process to judgmental forecasting. *International Journal of Forecasting*, *12*, 155–161.

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*, 127–151.

Buehler, R., & Griffin, D. (2003). Planning, personality, and prediction: The role of future focus in optimistic time predictions. *Organizational Behavior and Human Decision Processes*, *92*, 80–90.

Clements, M. P., & Hendry, D. F. (1995). On the selection of error measures for comparisons among forecasting methods — reply. *Journal of Forecasting*, *14*, 73–75.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1673.

Deschamps, E. (2004). The impact of institutional change on forecast accuracy: A case study of budget forecasting in Washington state. *International Journal of Forecasting*, *20*, 647–657.

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

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

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

Fildes, R., & Beard, C. (1992). Forecasting systems for production and inventory control. *International Journal of Production and Operations Management*, *12*, 4–27.

Fildes, R., Bretschneider, S., Collopy, F, Lawrence, M., Stewart, D., Winklhofer, H., et al. (2003). Researching sales forecasting practice – commentaries and authors' response on "Conducting a sales forecasting audit" by M.A. Moon, J.T. Mentzer & C.D. Smith. *International Journal of Forecasting*, *19*, 27–42.

Fildes, R., & Goodwin, P. (2007). Against your better judgement? How organizations can improve their use of management judgement in forecasting. *Interfaces*, *37*, 570–576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, *42*, 351–361.

Fildes, R., & Hastings, R. (1994). The organization and improvement of market forecasting. *Journal of the Operational Research Society*, *45*, 1–16.

Flybjerg, B., Bruzelius, N., & Rothengatter, W. (2003). *Megaprojects and risk: An anatomy of ambition*. Cambridge, UK: Cambridge University Press.

Galbraith, C. S., & Merrill, G. B. (1996). The politics of forecasting: Managing the truth. *California Management Review*, *38*, 29–43.

Gardner, E. S. (2006). Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting*, *22*, 637–666.

Gardner, E. S., & Anderson, E. A. (1997). Focus forecasting reconsidered. *International Journal of Forecasting*, *13*, 501–508.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgment. *International Journal of Forecasting*, *16*, 85–99.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, *12*, 37–53.

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408.

Goodwin, P., Lee, W. Y., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). *Understanding the use of forecasting systems: An interpretive study in a supply-chain company*. Bath University Management School working paper.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes*, *63*, 247–263.

Helbok, G., & Walker, M. (2004). On the nature and rationality of analysts' forecasts under earnings conservatism. *The British Accounting Review*, *36*, 45–77.

Helms, M. M., Ettkin, L. P., & Chapman, S. (2000). Supply chain forecasting — collaborative forecasting supports supply chain management. *Business Process Management Journal*, *6*, 392–407.

Kottemann, J. E., Davis, F. D., & Remus, W. E. (1994). Computer-assisted decision making: Performance, beliefs, and the illusion of control. *Organizational Behavior and Human Decision Processes*, *57*, 26–37.

Langer, E. J. (1975). Illusion of control. *Journal of Personality and Social Psychology*, *32*, 311–328.

Lawrence, M., & O'Connor, M. J. (1996). Judgement or models: The importance of task differences. *Omega: International Journal of Management Science*, *24*, 245–254.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, *122*, 151–160.

Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*, 493–518.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*, 377–390.

Lim, J. S., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts — its effectiveness and biases. *Journal of Behavioral Decision Making*, *8*, 149–168.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting*, *12*, 139–153.

MacGregor, D. (2001). Decomposition for judgemental forecasting and estimation. In J. S. Armstrong (Ed.), *Principles of forecasting* (pp. 107–123). Norwell, MA: Kluwer.

Mathews, B. P., & Diamantopoulos, A. (1986). Managerial intervention in forecasting: An empirical investigation of forecast manipulation. *International Journal of Research in Marketing*, *3*, 3–10.

Mathews, B. P., & Diamantopoulos, A. (1987). Alternative indicators of forecast revision and improvement. *Marketing Intelligence and Planning*, *5*, 20–23.

Mathews, B. P., & Diamantopoulos, A. (1989). Judgemental revision of sales forecasts — a longitudinal extension. *Journal of Forecasting*, *8*, 129–140.

Mathews, B. P., & Diamantopoulos, A. (1990). Judgmental revision of sales forecasts — effectiveness of forecast selection. *Journal of Forecasting*, *9*, 407–415.

Mathews, B. P., & Diamantopoulos, A. (1992). Judgmental revision of sales forecasts — the relative performance of judgementally revised versus non revised forecasts. *Journal of Forecasting*, *11*, 569–576.

Mathews, B. P., & Diamantopoulos, A. (1994). Towards a taxonomy of forecast error measures — a factor-comparative investigation of forecast error dimensions. *Journal of Forecasting*, *13*, 409–416.

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

Moon, M. A., Mentzer, J. T., & Smith, C. D. (2003). Conducting a sales forecasting audit. *International Journal of Forecasting*, *19*, 5–25.

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

Önkal, D., Gönül, M.S., & Lawrence, M. (2007). Judgmental adjustments: Effects of prior adjustments and explanations. *Paper presented at the 27th International Symposium on Forecasting*.

Sanders, N. R., & Manrodt, K. B. (1994). Forecasting practices in United States corporations — survey results. *Interfaces*, *24*, 92–100.

Sanders, N. R., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, *5*, 39–52.

Schultz, R. L. (1984). The implementation of forecasting models. *Journal of Forecasting*, *3*, 43–55.

Smaros, J. (2007). Forecasting collaboration in the European grocery sector: Observations from a case study. *Journal of Operations Management*, *25*, 702–716.

Staw, B. M. (1976). Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational Behavior and Human Performance*, *16*, 27–44.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2008). The effects of integrating management judgement on intermittent demand forecasts. *International Journal of Production Economics*, in press (doi:10.1016/j.ijpe.2008.08.011).

Tashman, L. J., & Hoover, J. (2001). Diffusion of forecasting principles through software. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners*, (pp. 651–676). Norwell, MA: Kluwer.

Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North-Holland.

Turner, D. S. (1990). The role of judgment in macroeconomic forecasting. *Journal of Forecasting*, *9*, 315–345.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *85*, 1124–1131.

Waern, Y., & Ramberg, R. (1996). People's perception of human and computer advice. *Computers in Human Behavior*, *12*, 17–27.

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

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*, 1–13.

Yelland, P. M. (2006). Stable seasonal pattern models for forecast revision: A comparative study. *International Journal of Forecasting*, *22*, 799–818.

Robert Fildes is Professor of Management Science in the School of Management, Lancaster University, and Director of the Lancaster Centre for Forecasting. He has a mathematics degree from Oxford and a Ph.D. in statistics from the University of California. He was co-founder of the *Journal of Forecasting* in 1981 and of the *International Journal of Forecasting* in 1985. For ten years from 1988 he was Editor-in-Chief of the IJF. He was president of the International Institute of Forecasters between 2000 and 2004. His current research interests are concerned with the comparative evaluation of different forecasting methods, the implementation of improved forecasting procedures in organizations and the design of forecasting systems.

Paul Goodwin is Professor of Management Science at the University of Bath. He has an MSc in Management Science and Operational Research from the University of Warwick and a Ph.D. in Management Science from the University of Lancaster. His research interests are concerned with the role of management judgment in forecasting and decision making. He is a Director of the International Institute of Forecasters, an Associate Editor of the *International Journal of Forecasting* and a member of the editorial board of the *Journal of Behavioral Decision Making*. He also writes a column in *Foresight*, the international journal of applied forecasting that is aimed at practitioners.

Michael Lawrence is Emeritus Professor of Information Systems in the Commerce and Economics Faculty at the University of New South Wales, Sydney, Australia. He has a Ph.D. from the University of California, Berkeley (specialising in Operations Research). Before joining the university he worked in the USA for the Ciba-Geigy Corporation and Corning Inc. He has held visiting positions, including Lancaster and Bath Universities, England, during the research reported on here. He is past President of the International Institute of Forecasters (IIF), past editor of the *International Journal of Forecasting*, and served for many years on the Board of Directors of the IIF. He is an associate editor of *Managerial and Decision Economics*. His research interests centre generally on how to best support decision making where a significant component of the decision involves management judgment, with a specific interest in how to support the task of short term forecasting.

Konstantinos Nikolopoulos is Assistant Professor in Decision Sciences at Manchester Business School. He received his Ph.D. from the National Technical University of Greece in the research field of Forecasting Support Systems. His research interests are in time series forecasting, forecasting support systems, artificial intelligence and geometry. He has published several papers in refereed academic journals including the *International Journal of Production Economics*, *International Journal of Forecasting*, and *European Journal of Operational Research.*