

Use and misuse of information in supply chain forecasting of promotion effects

Robert Fildes, Paul Goodwin, Dilek Onkal

Abstract:

Demand forecasting is critical to sales and operations planning (S&OP) but the effects of sales promotions can be difficult to forecast. Typically, a baseline statistical forecast is judgmentally adjusted on receipt of information from different departments. However, much of this information has either no predictive value or its value is unknown. Research into base rate discounting suggests that such information may distract forecasters from the average uplift and reduce accuracy. This was investigated in situations where forecasters could adjust the statistical forecasts for promotions via a forecasting support system (FSS). In two ecologically valid experiments forecasters were provided with the mean level of promotion uplift, a baseline statistical forecast, and quantitative and qualitative information. The forecasters were distracted from the base rate and misinterpreted the information available to them. These findings have important implications for the design of organizational S&OP processes, and for the implementation of FSSs.

Key Words: Sales and Operations Planning; behavioral operations; information effects; forecaster behavior; judgmental forecasting

1. Introduction

Production and inventory planning, scheduling, logistics, marketing and finance in supply chain companies all rely on short-term disaggregate forecasts at the SKU level. Yet little research has been carried out into the way such forecasts are actually produced or the factors that influence their effectiveness (Thomé et al., 2012; Tuomikangas & Kaipia, 2014; Seifert et al., 2015). In contrast to the

academic research literature, the practitioner literature is awash with descriptions and recommendations as to how ‘Sales and Operations Planning (S&OP)’ processes can be used to effectively integrate cross-functional information to produce forecasts (e.g., Lapide, 2007; Stahl, 2010). The demand uplifts achieved through sales promotions campaigns can be particularly difficult to forecast because of the relative infrequency of such events. When promotion campaigns are due to take place, the forecasts within S&OP are usually produced as a combination of a simple baseline statistical forecast and a judgmental adjustment, which is an estimate of the promotion effect (Fildes & Goodwin, 2007). The adjustments are made to reflect information received from different departments such as sales and marketing. These adjustments may mirror individual and functional biases stemming from informational blind spots as well as other organizational misalignments in supply chain processes (Oliva & Watson, 2009, 2011).

In one of the few detailed case studies of forecasting practice, Goodwin et al. (2007) found that the benefits of judgmental adjustments based on additional information in a pharmaceutical company were slight and often negative. Other studies have found evidence that information use is inefficient and biased (Fildes et al., 2009; Franses & Legerstee, 2010, 2011, 2013) and, where promotions are concerned, adjustments can therefore have a deleterious effect on forecast accuracy (Trapero et al., 2013). While the consensus is that integrating diverse sources of information is valuable (Kremer et al., 2016) and that forecast information sharing affects and usually improves supply chain performance (e.g., Ozer and Raz, 2011; Ozer et al., 2011), no studies have examined the nature of the information that was available and how it was used.

One possible cause of damaging judgmental adjustments is that too much attention is paid to information relating to single, isolated past events, such as one past promotion. Because the effects of such events are subject to noise, they offer little or no diagnosticity¹ when forecasting the effect of similar future events. A second possible cause is that information with unknown diagnosticity is overweighted

¹ “The diagnosticity of a piece of information is a measure of its helpfulness and usefulness for making a judgment (or forecast) in empirical studies.” Qiu, Pang, & Lim (2012).

when the judgmental adjustment is estimated. For example, a sales uplift of 80% may be estimated when a top celebrity is recruited to front an advertising campaign for a product, but there may be an absence of information to support such a judgment and its typical effects may be unknown. The attention paid to these two types of information may lead to the neglect of available base rate information, which may, indicate, for example, that promotions in the relevant product group or category lead to an average uplift of 50% (e.g. Kahneman and Tversky, 1973).

Using controlled experimentation with realistic simulations of the supply chain forecaster's task environment when sales promotions are imminent, this paper aims to help address the gap in our knowledge of how information is used in forecasting. (Note that the term 'supply chain forecaster' refers to a demand forecaster based in a company that forms part of a supply chain including retailers and manufacturers.) It does this by identifying how supply-chain-based forecasters respond when they have base rate information indicating the average sales uplift achieved during sales promotions, together with other information that has little or unknown diagnosticity - a situation that is common in S&OP settings. Promotions were shown by Fildes and Goodwin (2007) to be the most important reason behind judgmental adjustments of demand forecasts, and they are used in this study as an integral part of providing an ecologically valid supply chain forecasting task.

Our research makes three important contributions to the literature. First, it addresses a gap in our knowledge of how diverse information is used by judgmental forecasters when predicting the effects of sales promotions in the typical organizational context of a forecasting support system. Other researchers have investigated information use -and, specifically, the role of irrelevant information - in decision making in other, quite different contexts. For example, Gaeth and Shanteau (1984) examined the appraisal of soil types and learning effects when irrelevant materials were contained in soil. The participants in a study by Hutchinson and Alba (1991) attempted a visual product classification task when only one piece of information amongst others had diagnosticity. Returning to the classic Kahneman, Tversky experiment concerned with the classification of an individual as either an engineer or lawyer, Schwarz et al. (1991) focused on the effects of task framing in the neglect of base rates. Shelton (1999) examined auditor

judgments and the effects of experience on the ability to discount irrelevant information. These and other studies cited elsewhere in the text all emphasize the effects of irrelevant information, but the tasks they have focused on are all far removed from the day-to-day operational tasks faced by demand forecasters. Human judgment is highly dependent on the context and nature of the task so the findings of these studies may not apply when it comes to the important task of demand forecasting in supply chains. Second, this study investigates how the well-known base rate neglect phenomenon applies to judgmental time series forecasting. Expanding on this, with both qualitative and quantitative information available, we investigate how these different types of information are (mis) weighted. Third, the study is designed to investigate whether there is a tendency to underestimate the effects of special events when adjusting statistical forecasts within a forecasting support system. Such under estimation might result from a tendency to anchor on the statistical forecasts. Taken together, the findings have important implications for ‘value-added’ analysis whereby the adjustments made through the organizational S&OP process are evaluated as to their effectiveness in improving forecasting accuracy.

The paper is divided into five further sections. Following a review of the relevant literature in Section 2, we set out our hypotheses. Section 3 describes our methodology, the participants and the experimental setting. The fourth and fifth sections contain the detailed results of two experiments with associated discussion. Finally, Section 6 summarizes our conclusions as well as providing suggestions for further work and implications for practice.

2. Literature Review and Hypothesis Development

2.1 Base rate distractors

Previous work has acknowledged the importance of an expanded information set in order to enhance supply chain forecasting performance. Such information may come from internal sources such as marketing and operations (Fildes & Hastings, 1994) as well as from information shared by other supply chain partners (e.g., Eksoz et al., 2014; Önkal & Aktas, 2011). Such an extended information set has been shown to be valuable in enhancing accuracy with consequential stock service level improvements (Cui et al., 2015; Trapero et al., 2013). However, there appears to be surprisingly little empirical work on the actual *use* of information in supply chain forecasting, particularly in the important case of sales promotions. When determining the extent to which a baseline statistical forecast should be adjusted to take into account the effect of a forthcoming sales promotion, forecasters will usually have access to a diverse range of information, both quantitative and qualitative. In addition to a base rate figure that shows the average sales uplift achieved by past promotion campaigns, this will typically include historic demand data, past and current baseline statistical forecasts, quantitative information on the most recent sales promotion as well as qualitative information on factors judged to be relevant to the success of the forthcoming promotion.

A potential problem is that some of this information, will have little or no predictive value, that is no diagnosticity. For example, while a base rate founded on a large sample of previous promotions is likely to provide a reliable estimate of the typical sales uplift, the uplift achieved in the most recent promotion is a sample of just one observation. Using the base rate and most recent uplift to forecast the uplift of a future promotion is essentially a Bayesian task of combining a prior mean uplift (the base rate) with the evidence from the sample of one to produce a posterior mean. This requires the use of the following formula (assuming that the actual uplifts are normally distributed).

$$m'' = \frac{\sigma^2 m' + n\sigma'^2 m}{n\sigma'^2 + \sigma^2}$$

where m'' is the posterior mean;

m' is the prior mean;

m is the observed sample mean (in this case it equals the observed past promotion as the sample is of size 1);

σ^2 is the variance of the prior distribution;

σ'^2 is the sample variance (which is undefined given that we have a sample of 1);

n is the sample size.

If σ^2 is not known to forecasters, and since σ'^2 is undefined, the mean promotion uplift cannot be optimally updated based on this single observation (unless the forecaster assigns a subjective variance estimate to the single observation). Thus, the optimal forecasting strategy is to estimate an uplifted value close if not identical to the prior. Despite this, there is evidence that it is mostly this latest uplift that gets used in promotion forecasting in practice (Cooper et al., 1999). Recent observations tend to be more salient (Hutchinson and Alba, 1991) and the tendency to focus on them –particularly the most recent observation - rather than on the underlying performance of the system that produced these observations, has been found in several other judgmental forecasting studies (e.g., Andreassen & Kraus, 1990; Lawrence & O'Connor 1992, 1995; Bolger & Harvey, 1993).

However, the tendency to use the previous promotion effect in the adjustment to the statistical forecast, thereby neglecting the base rate, is likely to depend on the salience of this effect. Goodwin and Fildes (1999) found that previous promotion effects had no influence on judgmental time series forecasts for promotion periods when the time series was highly noisy. Under these conditions, the previous effects were submerged in the large random movements of the series and hence were not salient. In contrast, given the prevalence of a recency bias in judgmental forecasting, it seems likely that the proximity of the latest promotion to the current period will increase its salience and hence amplify its influence on the forecast.

A second problem is that some information, such as an announcement that a particular celebrity has been recruited to lead a promotion campaign, *may* be associated with an uplift in sales, but in the

absence of relevant data, the likely size of this effect is unknown and may even be zero. In these and similar circumstances, the safest strategy will be to adjust the statistical baseline forecast by estimating an uplift equivalent to the base rate. After all, some of the promotions that were used to estimate the base rate may also have employed celebrities (or had other characteristics that are similar to the forthcoming promotion), but information that would establish this is unlikely to be immediately accessible. Deviating from the base rate in such cases implicitly relies on the unsupported assumption that the celebrity effect (or the effect of any similar characteristic of the promotion that is being forecast) is not already embedded in the base rate. However, even though its diagnosticity is unknown, information in a narrative form is likely to act as a powerful distractor from the base rate (Önkal, Sayim, & Gönül, 2013). For example, in a classic study of judgmental decision making, Tversky and Kahneman (1974) showed that information on statistical base rates is often neglected or discounted and unreliable narrative information is preferred. Kahneman and Lovallo's (1993) notion of the 'inside view' would also suggest that the availability of a set of reasons as to why a promotion will, or will not, be a success, will cause attention to be devoted to the specific characteristics of the particular promotion that is being forecast. As a result, the focus on average sales uplifts (the 'outside view') will be lost. The above discussion suggests the following hypothesis:

H1: Adjustments to statistical baseline forecasts to take into account forthcoming promotion effects will deviate from base rates when information with no, or unknown, diagnosticity is provided.

How might a focus on the most recent promotion be translated into an adjusted forecast? One possibility is that people will employ the anchor-and-adjust heuristic (Bolger & Harvey, 1993; Epley & Gilovich, 2006; Lawrence & O'Connor, 1995; Tversky & Kahneman, 1974) and that the statistical baseline forecast will act as an anchor. This would lead to the forecast being formed as a weighted average of the baseline statistical forecast and the previous promotion. The effect of this would be a forecast that fell between the two values leading, on average, to an underestimate of the promotion effect.

Ironically, the use of a weighted average would imply that higher statistical forecasts lead to smaller *percentage* adjustments because in such cases the statistical forecast will be closer to the uplifted sales achieved in the previous promotion. An alternative suggested by de Baets and Harvey (in press) is that a weighted average of the estimated series mean and the statistical forecast acts as an initial anchor. The presence of sales uplifts reflecting multiple past promotions would clearly be expected to have a significant upwards effect on the estimated series mean, compared to cases where only a single past promotion is observed. Nevertheless, their study also found a tendency to under adjust from this anchor and hence, typically, underestimate promotion effects when multiple past promotions were observed.

In addition, when a forecaster has access to multiple items of qualitative information some of which are positive (suggesting that the promotion will be a relative success) and some of which are negative (suggesting the contrary), there is some evidence to suggest that the negative information may be more potent (Rozin & Royzman, 2001). This is consistent with Prospect Theory, which assumes that people have an aversion to losses (Kahneman & Tversky, 1979), and hence may be more vigilant in their response to negative information indicating a potential loss than they would be to positive information. The reasons for the greater influence of negative information are complex but researchers such as Peeters and Czapinski (1990) have suggested that in many environments negative events are rarer but can have more important implications for survival, so it pays to be especially watchful for dangerous negative events. In our context, this negativity bias implies that in situations where positive and negative reasons are equally likely to be present and where their diagnosticity is unknown, there may be a tendency to underestimate future promotion effects. These factors lead to the following hypothesis:

H2: Adjustments made to statistical baseline forecasts to take into account the effects of forthcoming promotions will tend to underestimate these effects when information with no, or unknown, diagnosticity is provided.

In summary, in the S&OP forecasting process where information relating to a forthcoming promotion is provided in quantitative, graphical and qualitative forms, the literature suggests that information with zero or unknown diagnosticity is likely to distract a demand forecaster from the normative base rate for promotional events.

2.2 Moderating factors

In practical contexts, a number of other factors may moderate the extent to which both the latest promotion and qualitative information arising from S&OP discussions lead to a tendency to under-forecast future promotion effects (De Baets, 2017). Before commencing the forecasting task, forecasters may have a prior view of the likely impact of promotions based on their (potentially imperfect) recall of earlier promotions (Reimers & Harvey, 2011), their own experiences, or industry beliefs. This may also serve to reduce the weight that is attached to the base rate. Secondly, forecasters in organizations may be subject to motivating factors that cause them to consciously or unconsciously bias their forecasts. The motivation of the forecaster may also affect the way in which sets of information in verbal statements will be assessed and aggregated in forecasting (Eroglu & Croxton, 2010). In some situations forecasters may prefer the variable-to-be-forecast to take on high or low values (e.g., a desire for high sales). Such desirability of outcomes may lead to an overblown optimism referred to as ‘desirability bias’ (e.g., Windschitl et al., 2010). As the Oliva and Watson (2009) case study shows, such a bias is a common feature of the S&OP forecasting process.

Despite these potential biases, forecasters in organizations are also likely to be motivated to produce accurate forecasts. Indeed, supply chain forecasters identified accuracy as their most important objective in the survey by Fildes and Goodwin (2007). Moreover, prestige and reputational concerns and the knowledge that one’s forecast will be evaluated may lead to a ‘reality constraint’ so that factors favoring biased forecasts, whether overly optimistic or pessimistic, may well be tempered (Lerner & Tetlock, 1999).

In addition to motivational influences, forecasters all come to the task with relevant past experience, which may affect the weightings they give to the different pieces of information they are presented with, whether in a real S&OP process or a simulated process. This had been observed with auditors (Shelton, 1999) and, in demand forecasting, Franses (2014) found that more experienced forecasters in a pharmaceutical company produced more accurate adjustments.

Finally, the most recent sales figure (as opposed to sales in the most recent promotion period) and the most recent forecast error may have an influence on the size of adjustment made for the forthcoming promotion. For example, an additional upwards adjustment might be made to reflect a relatively high last observation as it might be seen as mirroring a recent change in the baseline level of sales (e.g., a recent increase in the popularity of a product). If this observation is well above the forecast for that period (thus leading to a large positive forecast error), its salience, and hence its influence on the adjustment, is likely to be enhanced.

In summary, little is known about the way forecasters use the information they face whilst producing their judgmental adjustments of statistical baseline forecasts when products are due to be promoted. Yet it is an important issue in that inaccurate demand forecasts can be costly in terms of surplus inventory or the loss of customer goodwill and sales. It is also important theoretically in that little research has examined the general problem of which this is an example: the interpretation of diverse information in a time series context. In the remainder of this paper, we investigate whether and how information of different types distracts forecasters from using base rate information efficiently, with the effect of decreased accuracy.

3. Methodology and Design of Experiment 1

In order to test the hypothesis developed above, while controlling for prior expectations of promotion effects, self-reported knowledge of forecasting and different types of motivation, we have adopted a behavioural experimental approach. Controlled laboratory experiments are increasingly used to investigate demand forecasting behavior and related biases (Kremer et al., 2011; Siemsen, 2011; Harvey

& Reimers, 2013; Moritz et al., 2014), as they allow for systematized examinations of crucial factors affecting forecasting performance. They are also now common in the operations literature (Gans & Croson, 2008; Croson et al., 2013; Zhao et al., 2013).

We report in detail on one experiment (labelled as Experiment 1) which built on the experience gained from a number of preliminary experiments. The responses in these preliminary experiments provided information on such features as screen design and the number of time series that could be used, as we discuss below. The participants in Experiment 1 were management students, studying for either bachelors, masters or doctoral degrees at the Universities of Bath (UK), Bilkent (Turkey), and Lancaster (UK). They had all studied some forecasting. While they did not have the same experience as commercial forecasters, they did have at least as much statistical training as many practicing forecasters. Evidence provided by earlier studies, and recently by Kremer et al. (2016), strongly suggests there are few differences between student participants and practising forecasters in a context such as that simulated here. We pick up on this issue in the second experiment that involved executives, as discussed below.

The participants were asked to assume the role of a forecaster for a large company which supplies a wide range of products to supermarkets. They were told that their task was to predict the sales of a number of these products that would be subject to a sales promotion. Each participant was given a briefing describing the task, together with base rate information on the average percentage uplift in sales achieved by promotions at this supermarket.

Once the experiment started, other information was provided through an FSS including a graphical display (see Figure 1 for a typical screenshot) designed to have features and a format that is similar to those found in some widely used commercial forecasting systems (e.g. ForecastProTM). The realism of both the system and the participants' task was intended to enhance the ecological validity of our experimental task and its findings (Rogers & Soopramanien, 2009) and was based on our earlier field-based research in companies (Goodwin et al., 2007; Fildes et al., 2009). This was considered especially important since many experimental tasks in the extant literature do not allow the assessment of

the combinations of factors that apply in practical contexts, and the current task was structured to replicate as closely as possible the forecaster's everyday experiences on the job.

Key elements in the design of the task include the length of sales history presented graphically (24 periods), the number of past sales promotions (here chosen as one, a simplification for promotional-intensive products), the types of products (we aimed for everyday products familiar to participants) and the promotional information. The demand generating process and the forecast function (described below) were also chosen to replicate the types of series and forecasting methods facing practicing demand forecasters: exponential smoothing has therefore been used (see for example, Fildes et al. (2009 and Alvarado-Valencia et al. (2017)) where all the companies examined used this as the basis of their forecasting and adjustment process). The choice of average uplift is also potentially important. An analysis of promotional effects reported by companies provided a wide range of possible average uplifts with Nakamura et al. (2014) estimating a range of 20% to 40% while an analysis of US store data gave much higher estimates with a range of 127% to 519%². These estimates gave us a wide choice and an early experiment using 80% was contrasted with participants' prior views (which provided a median estimate of 50%). Since this last figure was within the range of plausible uplift values, we chose to use it in the experiments reported here. This average uplift was highlighted both in the cover story and in information presented on the computer screen during the trial run.

The participants first saw product details for a particular SKU (the SKUs were presented in random order), a corresponding time series sales history of 24 periods and the corresponding statistical forecast for all periods including the 25th.

The data were generated according to the rules:

² Thanks are due to Shaohui Ma who provided these figures as part of research reported in Ma et al. (2014).

$$Sales_t = BaseLineForecast_t + PromInd_t * Promotional Effect_t + \varepsilon_t$$

$$BaseLineForecast_t = 0.2 * Sales_{t-1} + 0.8 * BaseLineForecast_{t-1} + (Perturbation \text{ for } t = 25)$$

$$Promotional Effect \sim Uniform(80, 120)$$

$$PromPeriod_t \sim Int(Uniform(1, 24))$$

$$PromInd_t = 1 \text{ when } t = PromPeriod_t, = 0 \text{ otherwise}$$

$$\varepsilon_t \sim Normal(0, stddev).$$

The initial *BaselineForecast* was set to 200 with a promotional effect between 80 and 120 units so that it was, on average, 50% of the typical sales in non-promotion periods. The past promotional period occurred randomly over the 24 period history. The standard deviation (*stddev*) had values of 40 and 80. On the rare occasions where the simulated observation turned out negative, a value of 0 was substituted.

The FSS provided a simple exponential smoothing forecast, as shown, apart from a random perturbation in period 25. This was done by assigning each series a value of 0, or $\pm 50 * U(0.4, 0.6)$, i.e., a random perturbation of between 20 and 30 in absolute value. This limited the collinearity between the forecast, previous sales observation and previous error, allowing its influence on the adjustment to be estimated more precisely. It was made clear to the participants that the baseline forecast did not include any promotional effects. For promoted periods, t , the previous baseline forecast was not updated. The timing of the single promotion in the historical data was generated uniformly for an integer period between 1 and 24. The timing and effect of this promotion varied across SKUs with a mean sales uplift (relative to the baseline forecast) of 50%.

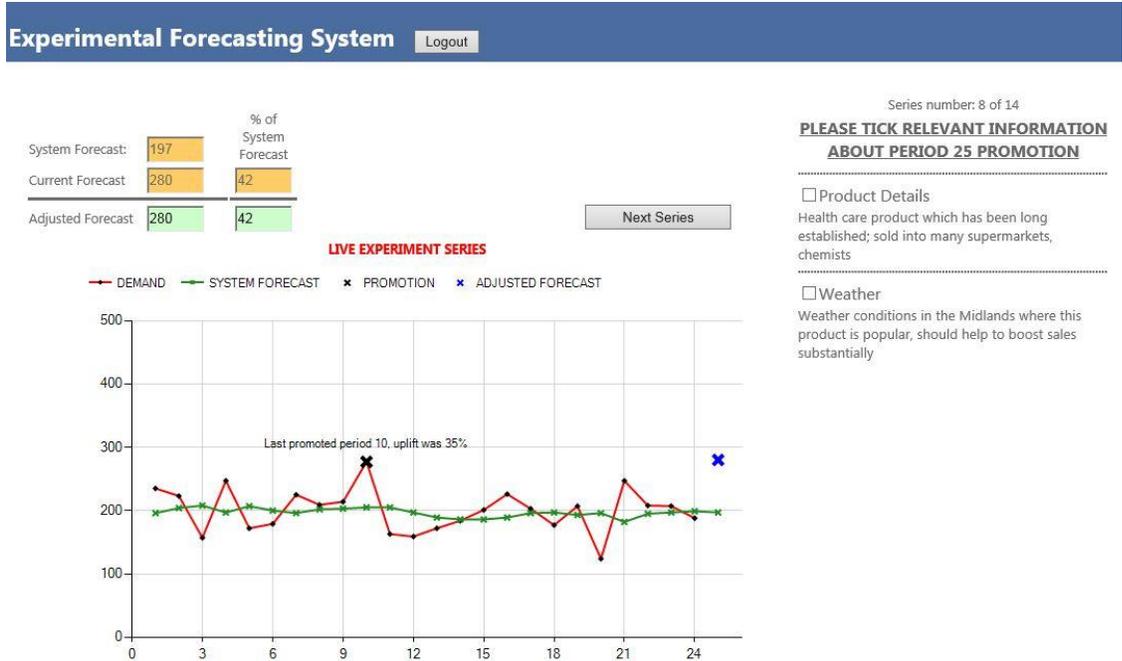


Figure 1. Screenshot of experimental forecasting support system

In addition to the historical sales series and statistical forecast, the screen displayed, for each SKU, between zero and four written statements which gave reasons suggesting why the level of sales uplift in the forthcoming promotional effect would be above or below the average (‘positive’ and ‘negative’ reasons). These reasons related to the amount spent on the promotion (e.g., “Over £1m is being spent on the promotion, double the usual size”), market research (e.g., “Focus groups have been quite negative about the promotional packs, but we can’t change these at this late stage.”), weather factors (e.g., “This product is mainly sold in the North where the weather conditions should be good for high sales according to the latest forecast.”) and campaign effectiveness (e.g., “We were hoping for a celebrity endorsement of our product as part of the campaign, but negotiations have not been successful and, unfortunately, we will have to run the campaign without this endorsement”), four reason types in all. The reasons were tested on nine experts for their plausibility and relevance to promotional events. A full list of reasons is available from the authors. Half of these were positive and half were negative, 12 in each category for each of the four reason types. The number of reasons displayed at any one time, the appearance of positive or negative reasons and the order of their display were all randomized. Having

been presented with all of this information, the participants were invited to use their judgment to adjust the baseline forecast for each SKU to take into account the forthcoming promotion for that product. These statements, typical of issues arising in S&OP discussions (See Goodwin et al., 2007), had no impact on the simulated promotional effects. In organizational S&OP such information is typically not supported by any evidence on promotional impact (although an exception has been observed in brewing, see Nikolopoulos & Fildes, 2013).

In order to control for the possible moderating effect of motivation, participants were randomly assigned to three treatments that were designed to provide different types of motivation. The first group were told that they would be rewarded when a promotion uplift exceeded 50%; although this was beyond their control, it was thought that the possibility of this reward might lead to desirability bias. The second group were told that they would be rewarded for the accuracy of their forecasts. A third (control) group were given a reward merely for participating in the experiment. The best two forecasters in each treatment received an Amazon voucher in the first and second groups; a prize draw was used to select the two winners in the control group. This led to a 3 (motivation type) between subjects x 12 (SKUs) within subjects design.

Before embarking on the experiment, participants were asked to indicate their prior view of what a typical percentage sales uplift would be for a fast-moving consumer good that was being promoted. They then made forecasts for the two SKUs that were used as a trial run to familiarize themselves with the FSS, before making forecasts for 12 additional SKUs that formed the basis of the experiment. For each SKU, they had the option of indicating which, if any, of the displayed reasons had led them to make their adjustment. During the trial run they were provided with an assessment of why the earlier promotion had, or had not been, a success – though no empirical evidence was produced to support the assessment. They also received overall feedback on their accuracy after forecasts had been made for both trial-run SKUs. No feedback was provided in the main part of the experiment. At the end of the experiment participants completed a questionnaire designed to assess their knowledge of forecasting, their

engagement in the task, their expectations regarding the accuracy of their judgmental adjustments and their interpretation of the reasons that were provided.

Because of the complexities of designing experiments that provide a realistic simulation of the supply-chain forecaster’s task, a number of preliminary experiments were run, involving over 200 participants. These enabled us to fine-tune the design and screen display, to eliminate potential confounded factors and to identify the key issues that merited further investigation. These experiments included different numbers of series, fixed numbers of reasons, forcing participants to select a primary and secondary reason to support their adjustment, having an average promotional uplift of 80% and including a trend in the data. The results of these experiments were consistent with those that we discuss next, suggesting that our results are robust. For brevity, these earlier results will not be reported here but they are available from the authors.

4. Analysis and Results of Experiment 1

126 participants completed the experiment. We then excluded respondents who did not make any but the very smallest average adjustments (i.e. their mean adjustment was less than 0) as this suggests either a limited understanding of promotional effects in retailing or no engagement with the experiments. The results are therefore based on a sample of 112 participants. As indicated above, participants responded to a post-experimental questionnaire. The main results of interest are summarized in Table 1.

Table 1 Questionnaire responses

Scale: (1) None / low expectations, to (5) High / high expectations - depending on question

Question	Mean	Std.Dev.
Rating of overall knowledge of demand forecasting	2.77	0.86
Expectations of statistical forecast performance	3.03	0.77
The provided reasons had a direct influence on my forecasts	3.46	1.07
Confidence in my final adjusted forecast	2.66	0.94

The results show that participants were generally motivated by the experiment and responded to the reasons provided. Typically, they did not ‘write-off’ the potential performance of the statistical baseline forecasts, despite the fact that they were bound to have large errors in a promotion period. This may reflect some acknowledgment of the statistical forecasts’ usefulness in establishing a reliable baseline for judgmental adjustment. The participants also indicated a lack of confidence in the accuracy of their adjusted forecasts which is reasonable given the level of uncertainty associated with the promotion effects.

We conducted a preliminary data analysis prior to developing a full linear model to explain the size of the participants’ adjustments. The participants’ median prior estimate of the percentage uplift achieved in supermarket promotions was 50%. However, their median estimated uplift during the experiment was only 30%, significantly lower ($p < .001$) than the base rate of 50%, a result consistent with a neglect of the base rate. This provides support for both H1 and H2. The distribution of these percentage adjustments was broadly normal with a few positive outliers. Only 25% of the adjustments were greater than the base rate of 50%. However, some were as high as 200%, which is quite possible for the sorts of products we have included in our experimental design.

4.1 Statistical modelling

Statistical modelling was used to identify the factors that were determining the sizes of adjustments made by participants and, in particular, whether the previous promotion effect and reasons were distracting them from the base rate. This also enabled the extent of any distraction to be estimated after taking into account the potential moderating factors discussed in Section 2.2. The nature of the experiment, where each respondent is sequentially given a number of series in random order, together with random information cues, requires a more sophisticated analysis than a standard ANOVA or regression.

Individual participants can be expected to have a random response to the series and to the cues. The advantages of using linear mixed effects models for this situation (Verbeke & Molenberghs, 2000) have been summarized as “they allow the researcher to simultaneously consider all factors that potentially contribute to the understanding of the structure of the data...including standard fixed effects and covariates” compared to standard approaches (Baayen et al., 2008). The statistical model is as follows:

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{b}_i \sim N(0, \mathbf{D})$$

$$\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}_i)$$

where \mathbf{Y}_i is the n_i dimensional response vector for respondent i , representing the promotional estimates for the i th series. X_i and Z_i are the $n_i \times p$ and $n_i \times q$ dimensional matrices of the factors influencing the response while $\boldsymbol{\beta}$ is the p - dimensional vector of fixed treatment effects and \mathbf{b}_i is the q dimensional vector of random effects. The covariance matrices are potentially important to the model building. \mathbf{D} and $\boldsymbol{\Sigma}$ are assumed independent. A repeated measures design is needed as the observations of the promotional uplift estimates from a given subject cannot be assumed independent of each other, for example in the sequence in which they were made. The standard assumption made for the variance-covariance matrix of the random effects, \mathbf{D} , is that the respective variances of the \mathbf{b}_i differ but are independent of each other –this is called the variance component assumption. In addition, the sensitivity of the estimated effects to changes in this assumption has been tested through an assumed autoregressive structure to capture any carry-over effect between the repeated observations, i.e., an AR(1) structure was assumed for \mathbf{D} . SAS 9.3 has been used to estimate the equations using restricted maximum likelihood.

The key features of the linear mixed effects model are set out below:

- The dependent variable is the adjustment percentage transformed into $\log(1 + \textit{Proportional_uplift})$ to ensure better error distributional characteristics (Davydenko & Fildes, 2013).
- The effects of variables relating to the past forecast history were assumed to be random effects as they depend on the individual participant. These variables were: the log of the respondent’s prior estimates of promotional effects, the log of the last forecast percentage error (measured as

$\log(\text{Forecast}/\text{Actual})$), the log of the uplift achieved in the last promotion (i.e. actual promoted sales over the baseline forecast), the log of latest forecast for the promoted period and the timing of the previous promotion.

- The effect of the series noise variance was treated as a fixed effects class variable.
- Participants' responses to the information cues were treated as random effects specific to the individuals.
- The numbers of positive and negative reasons were treated as fixed effects class variables, i.e. treatments.

Formally, let $Y_{ij} = \log(1 + \text{Proportional_uplift})$, the uplift estimated by the i th participant for the j th series.;

Then the model estimated is;

$$Y_{ij} = \beta_0 + \beta_{1i} * \ln(\text{Last_promotional_uplift})_j + \beta_{2i} * \ln(\text{Last_Actual})_j + \beta_{3i} * \ln(\text{Current_Stat_forecast})_j + \beta_{4i} * \ln(\text{Prior})_i + \lambda \text{Noise}_j + \beta_{5i} * (\text{Timing_past_promotion}_j) + \gamma_i \text{Reasncat}_{ij} + \text{error}_{ij}$$

where β_1, \dots, β_4 , are random effects, Noise is treated as fixed, and Reasncat as random.

Reasncat is defined as the number of positive reasons minus the number of negative reasons. In addition, the results presented have points of high leverage removed. Leverage was measured using Cook's D (eliminating points with $D > .002$ – approximating one of the recommended cut-offs of $4/n$). Various modelling choices also needed to be resolved, in particular how to characterize the number of negative and positive reasons. On a methodological note, since there is no 'optimal' route to establishing a model we sought to test out the robustness of our chosen approach by examining alternatives. Several were considered, including using both variables (with an interaction) and one variable together with the difference between positive and negative reasons. Using the variable Reasncat (as defined above as the number of positive reasons minus the number of negative reasons) proved the most parsimonious specification with minimum BIC. In addition, various interactions were included but did not add any explanatory power. Any non-linear effect of timing was also checked but a linear model proved adequate. A sensitivity check on the assumption of the correlation structure of the repeated measures did not show any substantive differences.

4.2 Results of modelling

The results from the model are shown in Table 2 which, as stated earlier, excludes observations with high leverage. However, as a check on the robustness of our findings, the results from estimating the model with the full set of observations remained broadly the same (1560 observations were reduced to 1309 after excluding high leverage points and non-compliant responders). The parameter coefficients are interpreted as percentage effects so for example, four negative reasons and no positive reasons ($Reasncat = -4$) lowers the average adjustment by 10.1% ($=100[1-\exp(-0.1067)]$). It can be seen that both the previous promotion uplift and the reasons were significantly associated with the adjustments made by the participants, consistent with H1. Higher uplifts in the previous promotion were associated with higher adjustments. This effect was slightly greater the more recent the promotion. As expected, lower levels of noise were also associated with higher estimated uplifts suggesting that high noise was making the effects of the previous promotion less salient. The significant negative coefficient for the statistical baseline forecast is consistent with participants placing their estimate of the uplifted sales between the baseline forecast and the previous promotion. This would account for the tendency to underestimate the expected uplift of 50%.

Table 2 Model of the adjustment: Dependent variable is $\log_e(1+Proportional_uplift)$

Effect	Estimate	p-value†
Intercept	0.505	<.0001
ln(last promotion uplift)	0.275	<.0001
ln (last actual)	0.037	0.001
ln (last stats forecast error)	0.040	0.105
ln(current stats forecast)	-0.128	<.0001
ln(Prior)	0.035	0.014
Low Noise	0.021	0.018
Timing of past promotion	0.001	0.036
Reasncat = -4	-0.107	<.0001
Reasncat = -3	-0.136	<.0001
Reasncat = -2	-0.114	<.0001

Reasncat = -1	-0.077	<.0001
Reasncat = 0	-0.078	<.0001
Reasncat = 1	-0.052	0.001
Reasncat = 2	-0.023	0.112
Reasncat = 3	-0.027	0.100

[Available n= 1560; sample size after deleting high leverage points=1309]

[*Reasncat* = No. of positive reasons supplied – No. of negative reasons]

† All tests are one-sided apart from that for low noise (and intercept).

Figure 2 shows the relationship between $\log_e(1+Proportional_uplift)$ and the difference between the number of positive and the number of negative reasons. The effects are compared with situations where there are an equal number of positive and negative reasons. It can be seen that, in general, the greater the number of positive reasons relative to the number of negative, the larger the upwards adjustment. This suggests that participants were balancing the reasons against each other, indicating that they were using a compensatory strategy. Broadly speaking, the greater the balance of reasons in one direction, then the greater was the distraction from the base rate, despite the unknown diagnosticity of these reasons, a result that is consistent with H1.

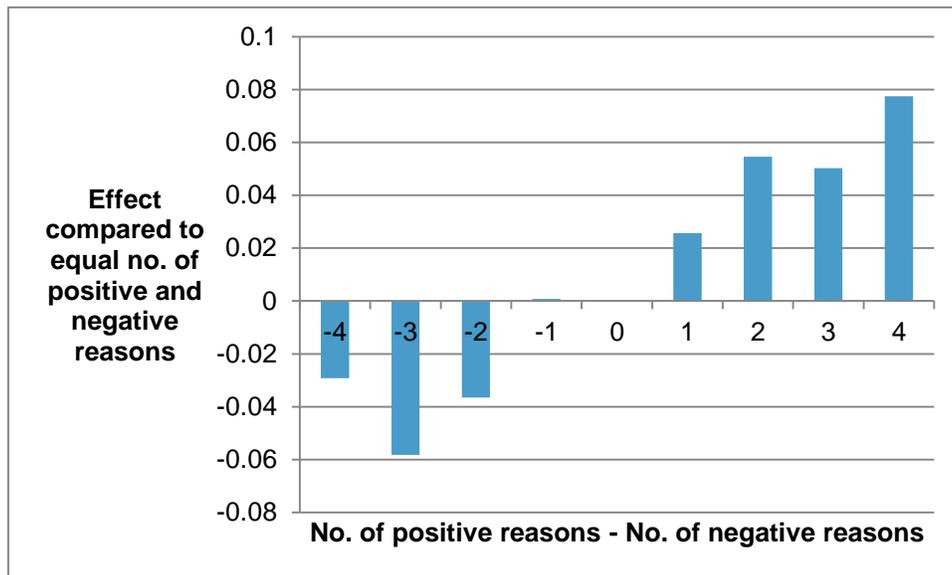
Did negative reasons have a greater influence than positive reasons? An analysis of ‘contrasts’ showed that having one more positive reason is more impactful than one more negative reason, but there is little difference between having two more positive compared two more negative reasons. Overall, the results suggest that positive reasons have slightly more effect than negative ones. This supports the argument that the propensity to underestimate the promotion effects resulted from the tendency to place the adjusted forecast between the baseline forecast and the previous promotion, rather than the alternative of attaching greater weight to negative reasons.

We investigated whether a number of additional variables had a moderating effect on the results. Table 2 shows that participants who came to the experiment with higher prior expectations of promotion uplifts tended to make larger upwards adjustments to the baseline forecasts. However, this carry-over effect to their individual SKU adjustments was small (as shown by the coefficient of the $\ln(Prior)$). There was also an apparent country effect between the participants based in the UK and Turkey ($p < .001$) with

the latter providing lower forecasts of uplifts. Once individual priors were included the effect was insignificant. This probably reflects the different retail environments that the participants were familiar with. In addition, Table 2 shows that adjustments tended to be higher when the most recent actual sales figure was higher (this turned out always as a non-promotion period). As discussed earlier, a high recent sales figure might be interpreted as a signal that the underlying level of sales has increased so that a greater adjustment to the statistical forecast is needed.

Figure 2. Effects on respondent’s estimate of uplift of differences between the number of positive and negative reasons

(Effects are measured relative to situations where there are an equal number of positive and negative reasons. The dependent variable is $\log_e(1+Proportional_uplift)$.)



There were no other substantive or significant effects on the size of the adjustment relating to the different motivation treatments or the characteristics of the participants, such as their knowledge of statistical forecasting, apart from the finding that participants’ motivation in the task proved significant in

increasing their average uplift. Telling participants that they would be rewarded if the uplift exceeded 50% (in an attempt to induce desirability bias) had no significant effect, nor did rewarding accuracy.

4.3 Discussion of Experiment 1

Overall, the results of Experiment 1 suggest that, when making their forecasts, participants were distracted from the 50% base rate by the previous promotion uplift and by the reasons despite this information having either no or unknown diagnosticity. In particular, they appeared to set their adjustment between the baseline statistical forecast and the previous promotion resulting in a tendency to under forecast the forthcoming promotion effect. These findings are consistent with earlier results on base rate neglect which suggest that base rates are eclipsed by more salient information with potentially lower predictive value. In fact, the introduction of any reasons surrounding the promotion had a negative effect on the calibration of the adjustment, a result established by comparing the situations where no reasons are offered with that where positive and negative reasons counterbalanced each other. It appears that the availability of any reason will distract forecasters from the base rate. Our findings also replicate studies that show how managers introduce variation into a system even when they know that demand is constant, as did our participants when they departed from the 50% uplift (Sterman & Dogan, 2015).

None of our motivation treatments had a significant effect on the size of adjustments: the treatments were intended to control for possible motivational effects. The absence of a desirability bias for those who were rewarded for higher than average uplifts was perhaps surprising. This once again demonstrates that it is difficult to replicate motivational and associated political effects that occur in the field in the laboratory. A small reward of a voucher for a higher sales forecast or an accurate one is not the same as the incentive to please the boss with a high forecast or the incentive to bring prestige and resources to one's department by producing reliable forecasts.

5. Experiment 2

In Experiment 1 participants had access to both potentially distracting information cues, namely the previous promotion uplift and the reasons. This did not allow the effects of information with zero diagnosticity and information with unknown diagnosticity to be examined separately. Experiment 2 had a simpler design. Participants were randomly assigned to one of two treatments. In the ‘previous promotion’ treatment, the series contained sales obtained in a previous promotion, but no reasons relating to the forthcoming promotion were displayed. In the ‘reasons’ treatment, reasons for the success or otherwise of the forthcoming promotion were displayed but no previous promotion effects appeared in the time series. In this case, for each SKU, either two negative, two positive or zero reasons were displayed with the number of reasons being selected at random. The design was based on the results of the previous experiment where the influential variable proved to be the difference between the number of positive reasons offered and the number of negative reasons. The effects had proved approximately linear between -2 and 2 reasons. In the ‘reasons’ treatment the promotion always appeared in period 18 eliminating any possible timing effects. Unlike Experiment 1 there was no random perturbation of the statistical baseline forecast and no motivation treatments were included. Each participant made forecasts for period 25 for 14 series, including two trial series. In all other respects the experiment was the same as Experiment 1.

The participants were 30 executives undertaking an Executive MBA module on forecasting so the experiment also enabled us to test whether the effects observed in Experiment 1 are also valid for experienced executives.

5.1 Results of Experiment 2

While the two sets of cues, past promotions and reasons, can be embedded in a single analytical model, analysis shows there are interaction effects that annul any gains in efficiency in the estimation of coefficients. Hence, separate mixed linear effects models were estimated for the adjustments made by participants in the two treatments. The overall mean (median) adjustment made was 49.4% (39.8%). The

models had the same underlying structure as the one used to analyse Experiment 1, except that the number of positive minus the number negative reasons was represented by a single variable, rather than a series of dummy variables, reflecting the approximately linear relationship referred to above. As before, the dependent variable was $\log_e(1 + \textit{Proportional_uplift})$ and high leverage points were removed. This time, $\ln(\textit{last actual})$ and $\ln(\textit{last stats forecast error})$ were not included in the list of independent variables because the lack of random perturbation in the statistical forecast meant that they would be collinear with that forecast.

The results from the group of executives taking part in this experiment generally support those reported earlier. As shown in Table 3, for the ‘previous promotion’ group, the previous promotion had a highly significant effect on the estimated uplift for the forthcoming promotion. Also, as in Experiment 1, a lower level of noise led to higher estimated uplifts consistent with the notion that the effect of the previous promotion was less salient under conditions of high noise. However, unlike Experiment 1, a higher statistical forecast was associated with a larger upwards adjustment, probably due to multicollinearity with the last actual (which was broken in Experiment 1). The effect of these factors was that the overall adjustment had a (trimmed) mean of 43%, and a median of 40%, (insignificantly different from 50%, but substantially and significantly less than the observed means of the past promotions, 57.5%).

Table 3 shows that participants in the ‘reasons’ group were significantly influenced by the provided reasons when they estimated the promotion uplift. The greater the balance in favour of positive reasons, the greater was the upwards adjustment they tended to make. For this group, on the occasion when no reasons were provided, their mean estimate was 52% (close to the base rate), although the median estimate was only 43%. As with the ‘previous promotion’ group, higher statistical forecasts were associated with larger upwards adjustments but this more experienced group were more firmly fixed on their prior promotional estimates. However, as expected, the level of noise did not have a significant effect, as there was no previous promotion uplift to be submerged in high noise.

Table 3 Model of promotional adjustment for the two treatments: ‘Previous promotion’ and ‘Reasons’

Effect	Treatment 1: Previous promotion		Treatment 2: Reasons	
	Estimate (n=159)	p- value	Estimate (n=161)	p- value
Intercept	-4.830	<0.001	-2.946	0.001
ln(last promotion uplift)	0.678	<0.001	n/a	n/a
ln(current stats forecast)	0.889	0.001	0.598	0.003
ln(Prior)	0.418	0.153	1.160	0.133
Low Noise	0.190	0.001	-0.022	0.608
No. of pos. minus no. of neg. reasons	n/a	n/a	0.032	0.0416
Mean adjustment	51.9		46.6	
Median adjustment	42.9		33.8	

All tests are one-sided except for those for *Low Noise* and the intercept.

In conclusion, Experiment 2 clearly demonstrated that, when the information cues were presented separately, both past promotions with zero diagnosticity and qualitative information with unknown diagnosticity distracted participants in their estimates of promotional adjustments from the normative adjustment of 50%. This provides further support for H1. Also, the effects observed earlier with a diverse group of business students have been replicated with experienced executives. However, the support for H2 (concerning the effects of information on under-adjustment) was weaker with the mean and median adjustments being closer to 50% than in Experiment 1, though they were still generally below this figure. This may be because each group in the experiment had only one main source of potential distraction (either the previous promotion or the reasons). The lower average adjustments for the ‘reasons’ group may reflect a tendency to anchor on the baseline forecast and the fact that this group did not observe a previous promotion that might have ‘pulled’ their estimates away from the anchor.

6. *General Discussion and Conclusions*

The efficient use of information by demand forecasters can be crucial given the negative effects of forecast errors on production, distribution and inventory planning. For example, Kremer et al. (2016) estimate that a percentage improvement in accuracy translates into a similar percentage reduction in safety stock. Given their supply chain repercussions, promotions pose particularly sharp challenges to S&OP decision-makers. The results of this experiment-based study suggest that the provision of information relating to promotions can be detrimental to forecast accuracy when, despite its salience, it has either no or unknown diagnosticity. This finding has important implications both for the design of forecasting support systems that are commonly used in supply-chain-based organizations and for the extent to which supply chains can operate efficiently. The systems typically place an emphasis on the provision of information to the forecaster in an amenable and accessible format, irrespective of its predictive value. Our results suggest that these ‘passive’ systems may be inimical to accuracy. In particular, participants appeared to adopt a version of the ‘last-lift’ heuristic, the most common promotional forecasting method used in practice. Their mistake was to ignore the average uplift, instead focusing on the last observed value.

Both facets of the participants’ sub-optimal forecasting suggest that FSSs need to be redesigned. Systems that actively evaluate and filter information before presenting it may lead to improved accuracy. Parikh et al. (2001) found that the provision in an FSS of informative guidance, which they defined as the provision of unbiased, relevant information without a specific suggestion, was superior in promoting learning compared to systems that suggested how information should be used. However, the emphasis needs to be on the provision of diagnostic and salient information. For example, in promotion forecasting, a system that identifies analogous past promotions and provides estimates of their average effect has been found to improve forecast accuracy in a laboratory study (Lee et al., 2007). However, as Lim and O’Connor (1995) and now Dietvorst et al. (2015) have shown, changing the habit of misweighting information remains difficult. Such intentional and unintentional misuse of information,

and the prevalence of habitual (mis)weighting schemes, further support the call for effective redesign of FSSs to aid predictive performance. But the redesign of the organization's FSS alone is not sufficient as the S&OP process also impacts information sharing and the salience of individual pieces of information, thought to be relevant, but of unknown diagnosticity. As a consequence, both the FSS and the S&OP process need to be considered together, incorporating some of the ideas laid out by Oliva and Watson (2009). The research question this raises is whether an S&OP process that incorporates an analytical 'notes' system into the supporting FSS and attempts to summarize past promotions (perhaps via short stories including explanations of available information and reasons for selecting particular forecasting model(s)) could be effective. In addition, the influence of prior beliefs about promotion effectiveness merits further attention. Research in psychology on conservatism indicates that in some circumstances such beliefs can be difficult to change, despite the provision of base rate information (Hilbert, 2012). However, Gaeth and Shanteau (1984) suggest training may help and there is some support for the idea that more experienced forecasters, like the auditors in Shelton (1999), may be less prone to strong responses to irrelevant qualitative information (as may have been the case in our experiment 2).

Is the underestimation of promotion effects we found in our laboratory experiments typical of what happens in the field? The evidence is sparse. Our study suggests the bias results from the poor design of FSSs but no field study has given details of the characteristics of the specific FSS that was used in promotion forecasting and the role it played, if one was used at all. Some field studies have reported that judgmental adjustments tend to suffer from optimism bias (Fildes et. al., 2009; Franses & Legerstee, 2011), while Muller (2011) reached the opposite conclusion. However, these studies focused on adjustment behaviour in all periods, rather than specifically on promotion periods. Further work on forecasting of promotional events is clearly needed to disentangle the confounding factors including the effects of a promotion forecast.

An obvious question is: how were the participants expected to know that information they were presented with had either no or unknown diagnosticity? A deeper reflection reveals that the diagnosticity of the information was self-evident. A sample of just one previous promotion when promotion effects

were subject to unknown levels of variation and the average promotional uplift is known clearly lacks diagnosticity. The presentation of reasons, such as “In this campaign, we will outspend our competitors by 100%”, without any supporting information on the typical effect of this factor on the number of units sold means that self-evidently the reasons had unknown diagnosticity. While the failure of participants to discount the extraneous information may not have been a surprise, our model has enabled us to measure the degree to which different types of such information impact judgmental forecasts in a realistic situation. The evidence we present is that the range of information taken into account is wide and damaging to accuracy.

However, like most experimental studies, this work has limitations. One issue is whether the participants felt obliged to deviate from the 50% base rate, otherwise why were they being invited to take part in the experiment. Simply entering a 50% uplift for every SKU may have seemed too easy or may have been perceived as signaling disengagement with the forecasting task. Yet, in this respect the experiment was an accurate reflection of the field. For example, Fildes et al. (2009) found that forecasters in companies tended to make lots of unjustified adjustments to statistical forecasts. There is evidence that these adjustments are often made so that forecasting staff can signal that they care about the forecasts produced, displaying a sense of ownership of the forecasts, or to simply justify their organizational roles (Önkal & Gönül, 2005). Having information that provides an apparent rationale for such adjustments is likely to increase their prevalence.

The participation of students in Experiment 1 may be regarded as another limitation despite their motivation and knowledge of forecasting. However, this is unlikely to affect the substantive conclusions, as Experiment 2 indicated. (also see Kremer et al. (2016)). In addition, while the on-screen simulation mirrored the operational realities of forecasting closely, the demand model and the promotional effects were based on a simple statistical model. The results may also depend on the features of the baseline statistical model, where the smoothing parameter is known to affect responses (Kremer et al., 2011). Building on the results presented here, when much of the information available to the forecaster has no diagnosticity, future research could usefully examine the behavior of forecasters when the statistical

model captures some promotional drivers. A second issue is whether the different types of information examined here (or observed in the S&OP process) are interpreted differently and given different weights when adjustments are made. A limited investigation supports this view.

In summary, given the limitations of current forecasting systems (Fildes et al., 2006), there appears to be substantial scope for design innovations. These may include structured support on filtering and integrating qualitative and quantitative information, targeted to individual forecasters, as well as support on design of collaborative forecasting systems that reach across different supply chain partners operating under diverse information platforms. These suggestions have implications for the organizational design of the S&OP process and further work on such innovations promises to enhance communication between forecasters and decision makers with extensive impact on overall supply chain performance.

References

Alvarado Valencia, J.A., Barrero, L.H., Önköl D., & Dennerlein, J. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting*, 33, 298–313.

Andreassen, P. B. & Kraus, S.J. (1990). Judgmental extrapolation and the salience of change. *Journal of Forecasting*, 9, 347-372.

Baayen, R. H., Davidson, D. J. & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.

Bolger, F., & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology*, 46, 779-811.

Cooper, L. G., Baron, P., Levy, W., Swisher, M. & Gogos, P. (1999). PromoCast TM, A new forecasting method for promotion planning. *Marketing Science*, 18, 301-316.

Croson, R., Schultz, K., Siemsen, E. & Yeo, M.L. (2013). Behavioral operations, The state of the field. *Journal of Operations Management*, 31, 1-5.

Cui, R., Allon, G., Bassamboo, A., & Mieghem, J.A.V. (2015). Information Sharing in Supply Chains, An Empirical and Theoretical Valuation. *Management Science*, 61, 2803-2824.

Dawes, R.M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571-582.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting*, 29, 510-522.

De Baets, S., (2017). Allowing for promotion effects in forecasting, Effects of judgment and formal forecasts. PhD thesis, Ghent University.

De Baets, S. & Harvey, N. (in press). Forecasting from time series subject to sporadic perturbations: effectiveness of different types of forecasting support. *International Journal of Forecasting*.

Dietvorst, B. J., Simmons, J.P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology-General*, 144, 114-126.

Eksoz, C., Mansouri, S.A. & Bourlakis, M. (2014). Collaborative forecasting in the food supply chain, A conceptual framework. *International Journal of Production Economics*, 158, 120-135.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17, 311-318.

Eroglu, C., & Croxton, K.L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26, 116-133.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment 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., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments, an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25, 3-23.

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

Franses, P.H. (2014). *Expert Adjustment of Model Forecasts*, Cambridge: Cambridge University Press.

Franses, P.H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29, 331-340.

Franses, P.H., & Legerstee, R. (2011). Combining SKU-level sales forecast from models and experts. *Expert Systems with Applications*, 38, 2365-2370.

Franses, P.H., & Legerstee, R. (2013). Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? *International Journal of Forecasting*, 29, 80-87.

Gans, N., & Croson, R. (2008). Introduction to the special issue on behavioral operations. *Manufacturing & Service Operations Management*, 10(4), 563-565.

Gaeth, G.J. & Shanteau, J.(1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance*, 33 (2), 263-282.

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., Lee, W-Y., Fildes, R., Nikolopoulos, K. & Lawrence, M. (2007). *Understanding the Use of Forecasting Systems, An Interpretive study in a Supply-Chain Company*. University of Bath, School of Management Working Paper Series, 2007.14.

Harvey, N., & Reimers, S. (2013). Trend damping: Under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology, Learning, Memory, and Cognition*, 39, 589-607.

Hilbert, M. (2012). Toward a synthesis of cognitive biases, how noisy information processing can bias human decision making. *Psychological Bulletin*, 138, 211-237.

Hogarth, R. (1987). *Judgment and Choice, The Psychology of Decision 2nd edition*. Chichester, UK: Wiley.

- Hutchinson, J.W. & Alba, J.W.(1991). Ignoring irrelevant information: situational determinants of consumer learning, *Journal of Consumer Research*,18 (3), 325-345.
- Kahneman, D. & Lovallo, D. (1993). Timid Choices and Bold Forecasts - a Cognitive Perspective on Risk-Taking. *Management Science*, 39, 17-31.
- Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, 47, 263
- Karelaia, N. & Hogarth, R. M. (2008). Determinants of linear judgment, A meta-analysis of lens model studies. *Psychological Bulletin*, 134, 404-426.
- Kremer, M., Moritz, B. & Siemsen, E. (2011). Demand forecasting behavior, system neglect and change detection. *Management Science*, 57, 1827-1843.
- Kremer, M., Siemsen, E. & Thomas, D.J. (2016). The sum and its parts, judgmental hierarchical forecasting. *Management Science*, 62, 2745–2764
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480-498.
- Lapide, L. (2007). Sales and operations planning S&OP mindsets, *Journal of Business Forecasting*. Spring, 21-31.
- Lawrence, M., & O'Connor, M. (1992). Exploring judgemental forecasting. *International Journal of Forecasting*, 8, 15-26.
- 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.
- Lawrence, M., & O'Connor, M. (1995). The anchor and adjustment heuristic in time- series forecasting. *Journal of Forecasting*, 14, 443-451.
- 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.

- Leitner, J. & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – A review of the literature. *European Journal of Operational Research*, 213, 459-469.
- Lerner, J. S. & Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255-275.
- Lim, J.S. & O'Connor, M. (1995). Judgemental adjustment of initial forecasts, its effectiveness and biases. *Journal of Behavioral Decision Making*, 8, 149-168.
- Moritz, B., Siemsen, E. & Kremer, M. (2014). Judgmental forecasting, Cognitive reflection and decision speed. *Production and Operations Management*, 23, 1146–1160.
- Nakamura, R., Pechey, R., Suhrcke, M., Jebb, S. A., & Marteau, T. M. (2014). Sales impact of displaying alcoholic and non-alcoholic beverages in end-of-aisle locations, An observational study. *Social Science & Medicine*, 108, 68-73.
- Nikolopoulos, K., & Fildes, R. (2013). Adjusting supply chain forecasts for short-term temperature estimates, a case study in a Brewing company. *IMA Journal of Management Mathematics*, 24, 79-88.
- Oliva, R. & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and Operations Management*, 18, 138-151.
- Oliva, R. & Watson, N. (2011). Cross-functional alignment in supply chain planning, A case study of sales and operations planning. *Journal of Operations Management*, 29, 434-448.
- Önkal, D. & Aktas, E. (2011). Supply chain flexibility, Managerial implications. In D. Önkal and E. Aktas (Eds.), *Supply Chain Systems-Pathways for Research and Practice*. Croatia: Intech Publ., 75-84.
- Önkal, D. & Gönül, M.S. (2005). Judgmental adjustment, a challenge for providers and users of forecasts. *Foresight: The International Journal of Applied Forecasting*, 1, 13-17.
- Önkal, D., Sayim, K. Z., & Gönül, M.S. (2013). Scenarios as channels of forecast advice. *Technological Forecasting & Social Change*, 80, 772-788.
- Özer, Ö. & Raz, G. (2011). Supply chain sourcing under asymmetric information. *Production and Operations Management*, 20, 92 – 115.

- Özer, Ö., Zheng, Y. & Chen, K.-Y. (2011). Trust in forecast information sharing. *Management Science*, 57, 1111 – 1137.
- Parikh, M., Fazlollahi, B., & Verma, S. (2001). The effectiveness of decisional guidance, an empirical evaluation. *Decision Sciences*, 32, 303-332.
- Payne, J.W., Bettman, J.R. & Johnson, E.J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press.
- Peeters, G., & Czapinski, J. (1990). Positive–negative asymmetry in evaluations: The distinction between affective and informational effects. In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology*, Vol. 1, New York: Wiley, pp. 33-60.
- Qiu, L., Pang, J., & Lim, K. H. (2012). Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence. *Decision Support Systems*, 54, 631-643.
- Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting*, 27, 1196-1214.
- Rogers, G. & Soopramanien, D. (2009). The truth is out there! How external validity can lead to better marketing decisions. *International Journal of Market Research*, 51, 163-180.
- Rozin, P. & Royzman, E.B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296-320.
- Schwarz, N. Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation, the contextual relevance of irrelevant information. *Social Cognition*, 9(1), 67-84.
- Shelton, S.W. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *The Accounting Review*, 74(2), 217-224.
- Seifert, M., Siemsen, E., Hadida, A.L. & Eisingerich, A.B. (2015). Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36, 33-45.

- Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data, The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, 249, 245-257.
- Siemens, E. (2011). The usefulness of behavioral laboratory experiments in supply chain management research. *Journal of Supply Chain Management*, 473, 17-18.
- Stahl, R.A. (2010). Executive S&OP, Managing to achieve consensus. *Foresight: The International Journal of Applied Forecasting*, Fall, 34-38.
- Sterman, J.D. & Dogan, G. (2015). "I'm not hoarding, I'm just stocking up before the hoarders get here." Behavioral causes of phantom ordering in supply chains. *Journal of Operations Management*, 39-40, 6-22.
- Thomé, A.M.T., Scavarda, L.F., Fernandez, N.S. & Scavarda, A.J. (2012). Sales and operations planning: A research synthesis. *International Journal of Production Economics*, 138, 1-13.
- Trapero, J. R., Pedregal, D.J., Fildes, R. & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting*, 29, 234-243.
- Tuomikangas, N. & Kaipia, R. (2014). A coordination framework for sales and operations planning S&OP, Synthesis from the literature. *International Journal of Production Economics*, 154, 243-262.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty, Heuristics and biases. *Science*, 85, 1124-1131.
- Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.
- Webby, R., M. O'Connor, B. Edmundson. 2005. Forecasting support systems for the incorporation of event information, An empirical investigation. *International Journal of Forecasting*, 21, 411-423.
- Windschitl, P.D., Smith, A.R., Rose, J.P. & Krizan, Z. (2010). The desirability bias in predictions: Going optimistic without leaving realism, *Organizational Behavior and Human Decision Processes*, 111, 33-47.
- Zhao, X., Zhao, X. & Wu, Y. (2013). Opportunities for research in behavioral operations management. *International Journal of Production Economics*, 142, 1-2.