Allowing for promotion effects in forecasting: Effects of judgment and formal forecasts

Shari De Baets

Advisors: Prof. Dr. Dirk Buyens

Prof. Dr. Karlien Vanderheyden

Submitted at Ghent University

Faculty of Economics and Business Administration

In partial fulfilment of the requirements for the

Degree of Doctor in Applied Economics

2016

ADVISORS

Prof. Dr. BUYENS Dirk, Department of Human Resource Management and Organizational Behavior, Ghent University

Prof. Dr. VANDERHEYDEN Karlien, Department of People and Organisation, Vlerick Business School

ADVISORY COMMITTEE

Prof. Dr. BAECKE Philippe, Department of Marketing, Vlerick Business School

EXAMINATION COMMITTEE

Prof. Dr. VAN KENHOVE Patrick, Department of Marketing, Ghent University (chairman)

Prof. Dr. BUYENS Dirk, Department of Human Resource Management and Organizational Behavior, Ghent University

Prof. Dr. VANDERHEYDEN Karlien, Department of People and Organisation, Vlerick Business School

Prof. Dr. BAECKE Philippe, Department of Marketing, Vlerick Business School

Prof. Dr. VAN DEN BROECK Herman, Department of Human Resource Management and Organizational Behavior, Ghent University

Prof. Dr. VEREECKE Ann, Department of Management Information Science and Operation Management, Ghent University

Prof. Dr. VANHOUCKE Mario, Department of Management Information Science and Operation Management, Ghent University

Prof. Dr. HARVEY Nigel, Department of Experimental Psychology, University College London

Prof. Dr. ÖNKAL Dilek, Department of Business Administration — Bilkent University

The work on this dissertation was made possible by the Intercollegiate Centre for Management Science (ICM).

Acknowledgements

Some say undertaking a PhD can get quite lonely. I have never experienced it as such and I have many people to thank for this. My advisors, Karlien & Dirk, have been there with constant support and guidance. Dirk, thank you for guiding me throughout this process and helping me focus. Your tremendous experience on all matters PhD and Career has been priceless. Karlien – where do I start? I am going to borrow a quote from Marc, who once said: “It doesn’t matter what the topic is. After you have talked with Karlien, you feel better”. This PhD and the start of the Vlerick Forecasting Centre have been a journey for us, and I sincerely hope we are not yet done with our travels. Your support means the world to me.

I am thankful for the scholarship from CIM. It didn’t just fund me these past three years; it also encouraged me to explore the broader academic community by going abroad. I consider myself very fortunate to be welcomed at University College London by Nigel. Thank you for being such a great mentor, by sharing your knowledge, your enthusiasm, and your experience. Every time we talk, I feel a little bit smarter. And thank you for introducing me to the community of judgmental forecasters. They have been welcoming and warm. Those words inevitably lead me to Dilek: you are a kind-hearted woman with vast knowledge. I hope to follow your example in my further career.

Knowledge can only be gained by working hard and being open to new ideas. Being in a field that lies somewhat at a cross-roads of different theoretical backgrounds and applications, I have encountered many opportunities to learn from others. Philippe, thank you for being a part of my guidance committee and helping me grow during these past years. Your analytical mind has been invaluable. Mario, I hope that we can philosophize some more together in the future on the importance of judgment in a future of machine learning and intelligence. Ann, you have the ability to question things I take for granted. Your different viewpoint shakes up my thought process – and encourage me to

think things through more thoroughly. Herman, you challenge me to think more high-level and to place myself within a broader field and framework.

The past six years at Vlerick wouldn't have been the same without my colleagues from P&O and the Leuven office – thank you all for the lovely chats, lunches, and seminars we shared together. My colleagues from UCL, thank you for making me feel right at home. One person spans both countries: Marc, you were my mentor in those first years. You introduced me to the field of decision making and therefore, to the future topic of my career. Whether it was that first day at Vlerick, or the first day in London, you have been there, supporting me. I cannot thank you enough for that.

Other colleagues, past and present, have become part of my life outside of Vlerick as well. Eva, Smaranda – you rock. I can always count on you for a good talk: whether it is to complain or to share a laugh, you are always there for me. And I do hope we can see many more ballets together. Freya, who speaks my language (books, books, books), Heidi, who embodies the warmth of South-Africa in her personality – I am grateful for your friendship. And Jana, who turned me into a hippie (gardening, sewing, ..), you were always there for a chat. Do we really have to live so far apart?

To my outside-work friends, Wim, Thomas. Thank you for the time we spend together, indulging in the nerdy aspects of life. You have no idea how much that time means to me. Valerie, my family-in-the-fifth-degree. I'm so happy we found each other again. Thank you for your support in those final stages of completing my PhD.

To my parents: you have been there over the years, always supportive, always understanding. Several people have told me: "you have such cool parents", to which I reply: "I know". You have taught me the value of lifelong learning and discovering new things. You lead by example. When I grow up, I want to be just like you. Thank you for being awesome.

Above all, I want to thank Koen. Thank you for being supportive, believing in me throughout it all, and supporting my crazy idea of getting a PhD. You were there to cheer me on if I felt down, and applied the brakes when I was working too much. Above all, you provided me with laughter. Thank you for being you.

Shari De Baets,

Gent, January 13th, 2017

Table of Contents

Chapter 1

Introduction

Introduction

Abstract

In this first Chapter, an in-depth background is provided for the reader on forecasting and its pivotal role in today's business world. The different methods (statistical, judgmental, combination), and their respective benefits will be highlighted. All three papers included here are aimed at unifying the strengths of judgmental and statistical forecasting. This doctoral thesis looks at a special situation, in which the potential of both can be realized: time series data, disturbed by exceptional events (promotions). At the beginning of this doctoral thesis, we set out to investigate combined judgmental and statistical forecasting in the real world. The dataset proved to be insightful and provided potential for thorough analysis. The combination used by the company involved was that of judgmental adjustment: a statistical model forecasted and a forecaster subsequently chose whether or not to adapt this forecast. In some cases, this adjustment was severely damaging to forecasting accuracy. Therefore, in a first study, we set out to apply a forecast support system that could mitigate this harmful effect. In other cases however, judgmental adjustment was beneficial to forecasting accuracy. To dig deeper into the beneficial effects of judgment in forecasting, we chose an experimental design in our second study. Unexpected, yet robust results required more investigation: judgment in itself outperformed the combination with statistics. Our third and final study digs deeper into this finding with a large-scale between-subjects experimental study. The quality of the statistical forecast was investigated, with its impact on different types of error. In the concluding chapter, these results are discussed, as well as the way forward, as there is still much potential for further research.

1. Literature background

Forecasting is "[the] *explicit processes* for determining what is likely to happen in the future" (Armstrong, 1999, p. 192). Fischhoff (Fischhoff, 1994, p. 387) then defines a forecast as "a set of probabilities attached to a set of future events." Forecasting is the driver of a wide variety of processes in the organization, including the introduction of new products, budget planning, and supply chain optimization (Fildes & Goodwin, 2007). For instance, the marketing department will forecast the expected number of sales, which will influence production planning, inventory, promotional planning, pricing, and financial prognosis (Smith, McIntyre, & Achabal, 1994). Improved sales forecasting may yield significant monetary savings, enhanced competitiveness and customer satisfaction (Fildes, Goodwin, & Lawrence, 2006; Moon, Mentzer, & Smith, 2003). Effective forecasting is thus a necessary skill in today's business world (Fildes & Goodwin, 2007; Giullian, Odom, & Totaro, 2000). Business leaders realize that they must understand what the future holds if they are to effectively grasp opportunities and have a solid starting point for their strategy (Makadok & Walker, 2000; Makridakis & Gaba, 1998; Shim, 2000; Titus, Covin, & Slevin, 2011). If a company fails to identify important future events or critical changes correctly, there is little chance that its strategy and planning will be successful (Makridakis & Gaba, 1998). Given the complexity of the business environment, the amount of data organizations have at their disposal is immense. To remain competitive, it is then vital to have the skills and tools to sift efficiently through the large amounts of data gathered (Economist Intelligence Unit, 2011).

Generally, forecasts are by experts (judgment), models, or a combination of both. The extant body of literature has focused on investigating the effects of these three types of forecasting on forecasting accuracy. Models can generate predictions based on the logical and systematic processing of information and can handle large amounts of data (Goodwin & Wright, 2010). The key advantage of models is that their predictions have a high degree of consistency and generate fewer errors than human judgment (Blattberg & Hoch, 1990). Models can thus improve forecasting performance by increasing the consistency of predictions (Hoch & Schkade, 1996). However, when an error does occur, it is more likely to be large and impactful: small input errors can produce large output errors (Stewart, 2001). Moreover, models are myopic; they fail to consider unexpected changes, and cannot deal with unstable environments or missing data (Armstrong & Collopy, 1998; Goodwin, 2002; Hughes, 2001; Taleb, 2007). In a world of change and uncertainty, this is especially troublesome (Economist Intelligence Unit, 2011; O'Connor, Remus, & Griggs, 1993). The model that best fits the historical data is also not necessarily the best fit for the future, implying that predictions based solely on historical data will not be sufficient, and do not guarantee future successes (Makridakis & Taleb, 2009).

In contrast, experts are capable of identifying new variables that influence predictions and are capable of subjective assessment (Blattberg & Hoch, 1990). Yet, a wide range of errors, such as biases and inconsistencies (Kahneman, 2011; Tversky & Kahneman, 1974), influence human judgment. Well known among researchers, these flaws represent a blind spot for the decision maker. Although we accept our limitations where memory is concerned (e.g., by making use of memory aids), no action is undertaken to counter our flawed judgment. Either people are unaware of their flaws and biases in judgment, or they are unwilling to accept them (Makridakis & Gaba, 1998).

When experts and models interact, experimental evidence suggests that people often make *unnecessary* adjustments to statistical forecasts based on their own judgment, even when they have no additional information (Goodwin, 2000; Lawrence, Goodwin, O'Connor, & Önkal, 2006). This is hypothesized to be because of the human tendency to discern patterns in noise or random numbers (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009), or the illusion of control effect, where forecasters are too optimistic and place excess weight on positive signals (Durand, 2003; Kotteman, Davis, & Remus, 1994). Moreover, those making forecasts are in general *overconfident* in the accuracy of their forecasts (Arkes, 2001; Lawrence, et al., 2006) and suffer from self-serving attribution bias, whereby they overestimate the importance of their own judgment when making adjustments to statistical forecasts (Hilary & Hsu, 2011; Libby & Rennekamp, 2012). Surprisingly, in a study by Lim and O'Connor

(1995), the tendency to adjust forecasts persisted despite giving participants feedback on their declining accuracy.

Reimers and Harvey (2011) summarize four factors leading to errors in forecasting. First, people tend to dampen trends in noisy series, thereby underestimating the steepness of trends in data series. This is arguably the result of an anchor-and-adjust heuristic, whereby people anchor on the last data point and then adjust (insufficiently) upwards (Bolger & Harvey, 1998; Harvey, 2007). Second, people tend to add noise of their own when making forecasts to make them more representative of the observed variation in the data (Harvey, 1995). Noise is thus really a problem for forecasters, given that it can mask any true patterns and lead to the misinterpretation of false signals (Goodwin & Fildes, 1999). Third, trend damping occurs more in downwardly than upwardly trending series (Harvey & Bolger, 1996; O'Connor, Remus, & Griggs, 1997). Fourth, people presume independent data to be positively serially correlated (Bolger & Harvey, 1993). Next to these unconscious errors, deliberate violations are also possible. We find examples in managers of poorly performing firms known to increase earnings forecasts to improve market perceptions (Rogers & Stocken, 2005); with analysts, who influence their forecasts to maintain a good relationship with management (Libby, Hunton, Tan, & Seybert, 2008; Washburn & Bromiley, 2013); and with sales representatives, who intentionally overestimate future sales to ensure product availability (Todd, Crook, & Lachowetz, 2013).

Despite these errors, studies have shown that human judgment is capable of improving on statistical models. Judgment enables human experts to process contextual information and environmental changes, which a model cannot (Cheikhrouhou, Marmier, Ayadi, & Wieser, 2011; Lawrence, et al., 2006; Whitecotton, Sanders, & Norris, 1998). Contextual information is what gives forecasting with judgment an edge over statistical forecasting, given that it can explain deviations in the data and identify future events that may cause anomalies (Armstrong & Collopy, 1998; Marmier & Cheikhrouhou, 2010). Webby and O'Connor (1996) confirm in their review of judgmental versus statistical forecasts that the accuracy of the judgmental process improves with contextual information or "domain knowledge." The concept of domain knowledge, defined by Lawrence et al. (2006, p. 499) is "…any information relevant to the forecasting task other than the time series—i.e. non-time series information."

In turn, this represents an un-modeled component in the statistical method. This can be, for instance, technical knowledge (knowledge about the data analysis and forecasting procedures), causal knowledge (an understanding of the cause-and-effect relationships involved), or product knowledge (Sanders & Ritzman, 1992; Webby & O'Connor, 1996). For example, domain experts may have knowledge of recent events that have not yet been included in the time series; unusual events that have occurred in the past, but not expected to occur again in the future; or unusual events that have not yet occurred, but expected to occur in the future (Armstrong & Collopy, 1998; Sanders & Ritzman, 2004). One particular important kind of information are so-called "broken leg" or "soft" cues, which concern impactful changes or events that cannot be captured in a formal model; e.g., a 90% probability of going to the movies would not hold if one has just broken a leg (Kleinmuntz, 1990). The ability to recognize these cues for what they are is crucial in forecasting and is what separates experts from novices (Goodwin & Wright, 2010). Moreover, statistical models may classify the effects of such broken leg cues as noise (Goodwin, 2000). Table 1 provides a comparison of the two main groups of methods for forecasting based on the literature review above.

	Statistical methods	Judgmental methods
Advantages	Consistent Processing power (speed, amount of data)	The unexpected Non-quantifiable information
Disadvantages	Exogenous events	Inconsistency & bias Limited processing power

Table 1. Advantages and disadvantages of statistical and judgmental methods.

While the (dis)advantages of the two groups of methods are painted by a broad brush, it does highlight an important quality: the potential for synergy. The advantages and disadvantages of both types of methods seem to balance each other out and hold the promise for improved forecasting accuracy. This potential seems to be situated especially in the case of exogenous events, in predicting the unexpected and in predicting rare occurrences with little quantifiable historical information. This doctoral research has focused on forecasting a type of special events: promotions. These can be seen as a disturbance in a time series. In contrast to the broken leg cues mentioned above, these special events have a degree of predictability. However, it is not an easy task. Usually, the amount of promotions in the past on which we can base ourselves, is limited. Additionally, the promotion disturbs the time series, complicating the baseline definition on and after the promotional event. This specific type of time series holds the potential for improved accuracy when judgmental and statistical methods are combined. The judgmental component can prove especially beneficial in such situations, by detecting the disturbance in the time series, providing the right interpretation and processing the information not ‘known’ to the statistical model. Importantly, practitioners report promotions to be the most common reason for adapting a statistical forecast (Fildes & Goodwin, 2007).

While research heavily values statistical methods, surveys indicate that between 40% and 50% of forecasting in practice involves judgment, and possibly even more (Webby & O’Connor, 1996). For example, Fildes and Goodwin (2007) showed that the most prevalent method to be the judgmental adjustment of a statistical forecast (33.1%), followed by the statistical method alone (25.0%) and judgment alone (24.5%). The method least used was averaging statistical and judgmental forecasts (17.7%). Judgmental adjustment, most often used, takes the statistical model as starting point, and allows for an adjustment based on judgment. Given its cost efficiency and its ease of use, it is not surprising that this method is common practice (Turner, 1990). The downside is biases and unnecessary adjustments, which are persistent and damaging to accuracy (e.g., Eroglu & Croxton, 2010; Fildes, et al., 2009; Webby & O’Connor, 1996).

Figure 1 represents the conceptual model of the doctoral thesis. This model does not indicate direct measurements and relationships; rather, it indicates the concepts and their values in each study. In every study, the task consists of forecasting from time series disturbed by promotions. A combination of judgment and statistics is used, leading to a certain level of forecasting accuracy. In the different studies, we can find different levels of expertise of the forecaster. Additionally, the context is also defined per study, as this can range from simple to complex.

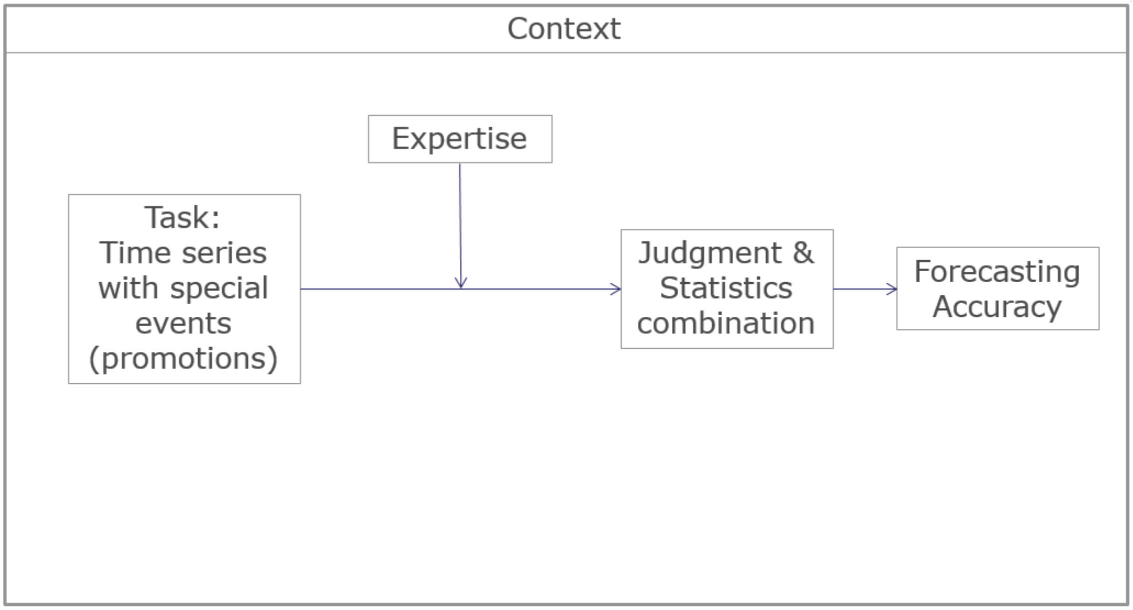

Context
Expertise
Task: Time series with special events (promotions)
Judgment & Statistics combination
Forecasting Accuracy

Figure 1. Conceptual model

2. Discussion of the three papers and research objectives

The three papers in this doctoral dissertation will tackle the combination of judgment and statistical models in three ways. The first paper, entitled "*Investigating the added value of integrating human judgement into statistical demand forecasting systems*", recognizes the problem of judgmental adjustment and attempts to resolve it via testing a method that integrates judgment into the model itself. This method can be seen as a form of forecasting support. Decision support systems are an important way of facilitating forecasting and heightening accuracy. Armstrong (2001c, p. 784) defines them as "…a set of procedures (typically computer based) that supports forecasting. It allows the analyst to easily access, organize, and analyze a variety of information. It might also enable the analyst to incorporate judgment and monitor forecast accuracy." Forecast support systems typically include a database with time series data, a number of statistical methods, and the possibility of integrating managerial judgments (Fildes, et al., 2006). According to Fildes et al. (2006), future decision support systems should be able to distinguish when judgment is appropriate. Moreover, when human intervention does occur, the system should provide support in debiasing the flaws that occur in judgment discussed earlier. According to Larrick (2004), debiasing can be by changing the cognitive strategies of the decision maker or by expanding the possible cognitive strategies with external techniques. Examples of the first approach include counterfactual thinking or training in decision rules, probabilities, or biases. Forecasting support systems (or the more widely applicable "decision support systems") are an example of the latter.

Interestingly, in an experimental study concerning the use of decision support systems, Goodwin, Fildes, Lawrence and Nikolopoulos (2007) found that participants ignored the "advice" of the system on which model to use (in the form of an "optimize" button). Only in 14.1% of the forecasts examined was the optimize button used, and only in 9.2% of the total examined forecasts was the advised method eventually chosen. Similarly, Lim and O'Connor (1995) found a tendency among forecasters to persist with damaging adjustments in subsequent forecasts, despite feedback that they were reducing accuracy. Decision makers seem to discount advice from statistical forecasts (Önkal, Goodwin, Thomson, Gönul, & Pollock, 2009), resist being debiased (Larrick, 2004), and statistical tools in the organization are distrusted and quickly fall into disuse (Zbaracki, 1998). Forecasters overvalue their own opinion over that of someone else and over that of a formal forecast (Harvey & Harries, 2004; Önkal, et al., 2009).

Research objective 1: to design a forecast support system that can improve on judgmentally adjusted forecasts, in such a way that it is beneficial for forecasting accuracy and, simultaneously, allows for judgmental input.

In our first paper, we therefore set out to investigate a method that would allow forecasters to input their own judgment, without harming accuracy. Judgmental forecasting is an inherently cross-disciplinary field, including but not limited to Psychology, Operational Research, Management Science and Statistics, with methods ranging from fundamental experimental research to applied empirical research. This paper can be situated in the scientific field 'Operations Research', as it focusses on the practical application of the developed model in an empirical context of stocks and sales.

The conceptual model of this doctoral thesis is shown in Figure 2. The task involved in the first paper consists of time series with special events. Specifically, this dataset consists of two types of periods or products: normal products or exceptional products (products with an extra). The data were sales data collected in an organizational context. The expertise of the forecasters was high, as they were professionals. The paper focusses on designing a Forecast Support Systems that optimizes the combination of judgment and statistical methods.

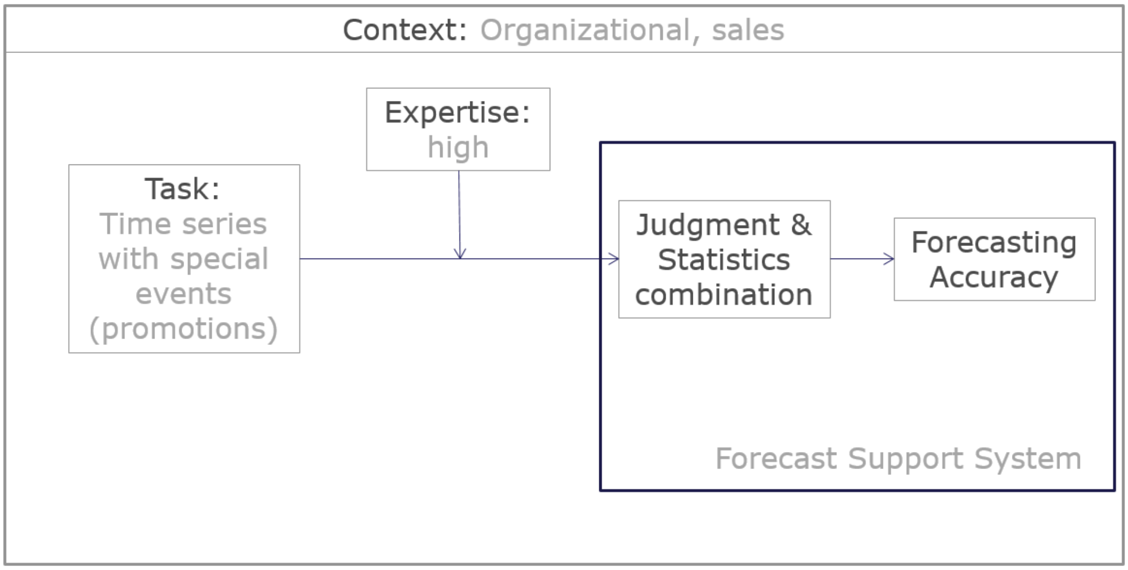

Context: Organizational, sales
Expertise: high
Task: Time series with special events (promotions)
Judgment & Statistics combination
Forecasting Accuracy
Forecast Support System

Figure 2. Conceptual model – paper 1

By integrating the judgmental component as a variable in the prediction model, we aimed to improve accuracy while retaining the input of the forecaster, thereby increasing their acceptance of the formal method. The challenge for an effective decision support system, is to design it as such that it informs forecasters about which elements they should leave to the statistical model and where adjustments would be beneficial (Goodwin & Fildes, 1999). Based on the work of Franses & Legerstee (2013), we provide an automation of this decision process.

After conducting the study in paper 1, it was deemed important to investigate these findings on the combination of judgment and statistics even further. While performing studies in the organizational context are of high practical relevance, we turn to experimental research for finding basic relationships between judgmental forecasting, statistics, and accuracy. Therefore, the second and third paper both employ an experimental design with non-professional forecasters. Time series are again perturbed by promotions, and are labeled as sales figures. In contrast to the more applied paper in Chapter 2, the papers in Chapter 3 and 4 focus on extracting the fundamental relationships found in judgmental forecasting. The framework used in these papers leans more closely to that of experimental psychology than operational research.

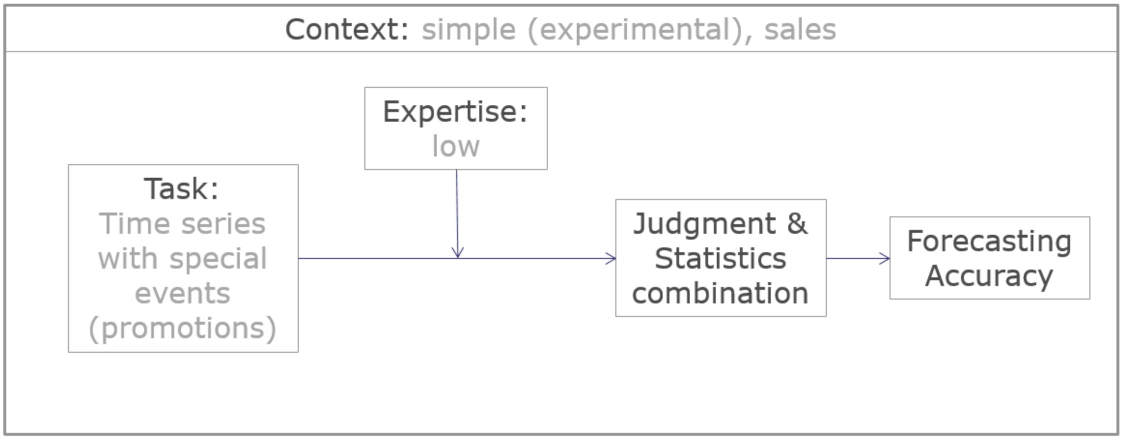

Context: simple (experimental), sales
Expertise: low
Task: Time series with special events (promotions)
Judgment & Statistics combination
Forecasting Accuracy

Figure 3. Conceptual model – paper 2 and 3.

In the second paper, entitled "*Enhanced anchoring effects produced by the presence of statistical forecasts: Effects on judgmental forecasting*", we perform three experiments aimed at entangling the effects of judgmental adjustment when making predictions from series perturbed by promotions. The second research objective for this doctoral paper therefore, is as follows:

Research objective 2: to investigate judgmental and statistical forecasting when faced with time series perturbed by promotions.

We found evidence of the statistical forecast being detrimental to forecast accuracy (Experiment 1). We manipulated the presentation of the statistical forecast history (Experiment 2) and lowered the proportion of promotions (Experiment 3). Our findings in the paper provide a cautionary tale in the application of statistical forecasting in time series with disturbances.

Subsequently, in the third paper, entitled "*Forecasting from time series subject to sporadic pertubations: Effectiveness of different types of forecasting support*", we worked further on the effects of combining judgment and formal methods in forecasting normal and promotional periods. We compared unaided judgmental forecasting with judgment aided by different types of statistical forecasts. Based on the results of the previous paper, we tested whether the detrimental effect of the statistical forecast was due to its crude form (not distinguishing between normal and promotional periods – which is realistic (Fildes, et al., 2009; Trapero, Pedregal, Fildes, & Kourentzes, 2013) but not generally used in research designs (e.g., Goodwin & Fildes, 1999; Goodwin, Fildes, Lawrence, & Stephens, 2011), or perhaps due to the experimental design. The latter was unrealistic as a professional forecaster will normally not have to forecast the same time series with and sometimes without statistical forecasts. Thus, in our third paper, the research objective is as follows:

Research objective 3: investigate the added value of providing a statistical forecast compared to unaided judgment.

An additional research objective specifies RO 3 further:

Research objective 3 bis: does the type of provided statistical forecast matter?

We compare unaided judgment, with judgment aided by a statistical forecast based on cleansed series as used in research, and with judgment aided by a statistical forecast based on non-cleansed series as used in practice. We investigate if providing a statistical forecast is beneficial and whether the type of statistical forecast provided matters for forecasting accuracy. We also looked into the proportion of promotions: does it matter whether the time series is often perturbed or only occasionally? Additionally, we looked into the effects of providing a near-perfect statistical forecast which included a forecast for the promotion. Overall conclusions are provided in the final chapter.

3. References

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (1999). Forecasting for environmental decision making. In V. H. Dale & M. E. English (Eds.), *Tools to Aid Environmental Decision Making* (pp. 192 - 225). New York: Springer-Verlag.

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Boston: Kluwer Academic Publishers.

Armstrong, J. S., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. In G. Wright & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 269 - 293). New York: John Wiley & Sons.

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

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

Bolger, F., & Harvey, N. (1998). *Heuristics and biases in judgmental forecasting*. Chichester: John Wiley & Sons.

Cheikhrouhou, N., Marmier, F., Ayadi, O., & Wieser, P. (2011). A collaborative demand forecasting process with event-based fuzzy judgements. *Computers & industrial engineering, 61*(2), 409 - 421.

Durand, R. (2003). Predicting a firm's forecasting ability: the roles of organizational illusion of control and organizational attention. *Strategic Management Journal, 24*(9), 821 - 838.

Economist Intelligence Unit. (2011). Game changer: How companies are responding to a fast-changing business environment. *The Economist*.

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*(6), 570-576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 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*(1), 3 - 23.

Fischhoff, B. (1994). What forecasts (seem to) mean. *International Journal of Forecasting, 10*(3), 387 - 403.

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*(1), 80 - 87.

Giullian, M. A., Odom, M. D., & Totaro, M. W. (2000). Developing essential skills for success in the business world: a look at forecasting. *The Journal of Applied Business Research, 16*(3), 51 - 61.

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

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega 30, 30*(2), 127 - 135.

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*(1), 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting, 23*(3), 391 - 404.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*(3), 242 - 253.

Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. *Technological Forecasting & Social Change, 77*, 355 - 368.

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

Harvey, N. (2007). Use of heuristics: insights from forecasting research. *Thinking and Reasoning, 13*, 5 - 24.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: effects of data presentation format on judgemental forecasting. *International Journal of Forecasting, 12*, 119 - 137.

Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20*(3), 391 - 409.

Hilary, G., & Hsu, C. (2011). Endogenous overconfidence in managerial forecasts. *Journal of Accounting and Economics, 51*(3), 300 - 313.

Hoch, S. J., & Schkade, D. A. (1996). A Psychological Approach to Decision Support Systems. *Management Science, 42*(1), 51 - 64.

Hughes, M. C. (2001). Forecasting practice: organisational issues. *Journal of the Operational Research Society, 52*, 143 - 149.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kleinmuntz, D. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological Bulletin, 107*(3), 296 - 310.

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

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 316 - 337). Oxford, UK: Blackwell Publishing.

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

Libby, R., Hunton, J. E. H., Tan, H. T., & Seybert, N. (2008). Relationship incentives and the optimistic/pessimistic pattern in analysts' forecasts. *Journal of Accounting Research, 46*(1), 173 - 198.

Libby, R., & Rennekamp, K. (2012). Self-Serving Attribution Bias, Overconfidence, and the Issuance of Management Forecasts. *Journal of Accounting Research, 50*(1), 197 - 231.

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

Makadok, R., & Walker, G. (2000). Identifying a distinctive competence: forecasting ability in the money fund industry. *Strategic Management Journal, 21*(8), 853 - 864.

Makridakis, S., & Gaba, A. (1998). Judgment: its role and value for strategy. In G. Wright & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 1 - 38). Chichester: John Wiley & Sons.

Makridakis, S., & Taleb, N. N. (2009). Living in a world of low levels of predictability. *International Journal of Forecasting, 25*(4), 840 - 844.

Marmier, F., & Cheikhrouhou, N. (2010). Structuring and integrating human knowledge in demand forecasting: a judgemental adjustment approach. *Production planning & control, 21*(4), 399 - 412.

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). Judgemental forecasting in times of change. *International Journal of Forecasting, 9*, 163 - 172.

O’Connor, M., Remus, W., & Griggs, K. (1997). Going up going down: how good are people at forecasting trends and changes in trends? *Journal of Forecasting, 16*, 165 - 176.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting, 27*, 1196 - 1214.

Rogers, J. L., & Stocken, P. C. (2005). Credibility of management forecasts. *The Accounting Review, 80*(4), 1233 - 1260.

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.

Sanders, N. R., & Ritzman, L. P. (2004). Integrating judgmental and quantitative forecasts: methodologies for pooling marketing and operations information. *International Journal of Operations and Production Management, 24*(5-6), 514 - 529.

Shim, J. K. (2000). *Strategic business forecasting: the complete guide to forecasting real world company performance*. New York, Washington DC: St. Lucie Press.

Smith, S. A., McIntyre, S. H., & Achabal, D. D. (1994). A two-stage sales forecasting procedure using discounted least squares. *Journal of Marketing Research, 31*(1), 44 - 56.

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners*. Boston: Kluwer Academic Publishers.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Titus, V. K., Covin, J. G., & Slevin, D. P. (2011). Aligning strategic processes in pursuit of firm growth. *Journal of business research, 64*(5), 446 - 453.

Todd, S. Y., Crook, T. A., & Lachowetz, T. (2013). Agency theory explanations of self-serving sales forecast inaccuracies. *Business and Management Research, 2*(2), 13 - 21.

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*(2), 234 - 243.

Turner, D. (1990). The role of judgement in mactroeconomic forecasting. *Journal of Forecasting, 9*, 315 - 346.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124 - 1131.

Washburn, M., & Bromiley, P. (2013). *Managers and analysts: An examination of mutual influence*. Academy of Management Journal.

Webby, R., & O'Connor, M. (1996). Judgmental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting, 12*, 91 - 118.

Whitecotton, S. M., Sanders, D., & Norris, K. B. (1998). Improving predictive accuracy with a combination of human intuition and mechanical decision aids. *Organizational Behavior & Human Decision Processes, 76*(3), 325 - 348.

Zbaracki, M. J. (1998). The rhetoric and reality of total quality management. *Administrative Science Quarterly, 43*, 602 - 636.

Chapter 2

Investigating the added value of integrating human judgement into statistical demand forecasting systems

Investigating the added value of integrating human judgement into statistical demand forecasting systems

Abstract

While academia point towards the benefits of a statistical approach, business practice continues to rely on judgmental approaches for demand forecasting. In the dynamic economic environment of today, a combination of both approaches may prove especially relevant. The question remains as to how this combination should occur. This study compares two different ways of combining statistical and judgmental forecasting, employing real-life data from an international publishing company that produces weekly forecasts on regular and exceptional products. Two methodologies that are able to include human judgment in a forecasting model are compared. In a restrictive judgement model, expert predictions are incorporated as restrictions on the forecasting model that determines the optimal number of magazines. In an integrative judgment model, this information is taken into account as a predictive variable in the demand forecasting process. The proposed models are compared on error metrics and analysed in-depth with regard to the properties of the adjustments (direction, size) and of the data itself (volatility). The integrative approach proves to have a positive effect on accuracy in all scenarios. However, in those cases where the restrictive approach proved to be beneficial (medium sized and big adjustments, downward adjustments and adjustments in case of high volatility), the integrative approach limited these beneficial effects. Additionally, this study includes the link with demand planning by using the forecasts as input for an optimization model, to determine the ideal number of SKUs per Points of Sale. Hence, it makes a distinction between SKU forecasts and SKU per PoS forecasts. Importantly, the latter enables expressing performance as a measure of profitability, which proves to be higher for the integrative approach than for the restrictive approach.

Keywords: Demand forecasting, judgmental forecasting, human judgment

1. Introduction

Accurate demand forecasting is a first vital step for supply chain management (Fildes, et al., 2006). Such forecasts have consequences for decisions within the organisation (e.g., manufacturing, marketing, logistics) and within the larger supply chain (e.g., suppliers, retailers). Forecasting however, is not an easy task and errors can have potential negative effects (e.g., Worthen, 2003). While the optimization of the forecasting process can yield significant advantages such as increased profitability and increased customer service levels (Moon, et al., 2003), empirical research remains fairly limited. Specifically, few studies have focussed on real company data (Sanders, 2009) and until recently (Fildes, et al., 2009; Franses, 2013; Franses & Legerstee, 2011; Syntetos, Kholidasari, & Naim, 2016; Trapero, et al., 2013), the topic of judgmental adjustments in the operational domain has been overlooked. While experimental research has provided a solid basis for investigating judgmental forecasting, organisation-based research is needed to further understand how forecasts are made (Franses, 2013; Sanders, 2009). Recent examples are the papers from Franses and Legerstee (2009, 2011, 2013), who work with data from a pharmaceutical company; Trapero, et al. (2013) with manufacturing company data, and Fildes, Goodwin, Lawrence and Nikolopoulus (2009), who compare four UK-based companies on the efficacy of their judgmental adjustments. We build further on their work by investigating the effect of judgmental forecasting within the context of a publishing company. We extend their studies in four ways: first, we investigate whether the approach of formally including expert knowledge as an additional explanatory variable as proposed by Franses and Legerstee (2011), although unsuccessful in the pharmacy industry, is more successful in an industry which handles both normal and exceptional products. Incorporating the expectations of experts has the possibility of improving the predictive performance of these models significantly, especially in the case of promotions (Fildes, Goodwin, & Onkal, 2016; Goodwin, 2002; Goodwin & Fildes, 1999; Trapero, et al., 2013). This is because the knowledge of expert forecasters represents a previously unmodelled component (Lawrence, et al., 2006). Although forecasting models are generally superior to human judgement based on the same information, experts can still add value because they are better able to recognize when predictions should be adapted based on additional information or the existence of exceptional events (Goodwin & Fildes, 1999; Jones & Brown, 2002). A study by Sinha and Zhao (2008) in the field of datamining demonstrated the added value of expert knowledge on a wide set of classifiers. The current study explicitly compares two different approaches: a restrictive approach and an integrative approach. In the former, predictions are restrictions of the demand forecasting model: i.e., the traditional case of judgmental adjustment of a statistical forecast. In the integrative approach, this information is taken into account as a predictive variable in the demand forecasting model itself. The advantage of this approach is two-fold: the forecasters retain their input and feeling of ownership, while the damaging effects of unnecessary adjustments are mitigated. In addition, this method should motivate forecasters to make the correct adjustment. The more accurate the adjustments of the forecaster were in the past, the more likely this variable will be picked up as a significant predictor for future demand.

Second, we integrate the papers of Fildes et al. (Fildes, et al., 2006) and Franses and Legerstee (2011), by providing an in-depth analysis of size and direction of adjustments, and volatility of the data. Regarding the latter, human judgment has been said to be especially relevant in the context of high volatility products due to special events such as promotions (Sanders & Ritzman, 1992). Our dataset gives a unique insight into volatility defined in two ways: as the variation in the data series (Sanders & Ritzman, 1992), and volatility defined as exceptional products.

Third, this study distinguishes between two levels of granularity in the forecasting process: SKU and SKU per PoS. Kremer, Siemsen, and Thomas (2015) have pointed out that judgmental forecasting research has not yet studied the effects on hierarchical forecasting. The accuracy of judgmental forecasts on the top-level and the bottom-level are generally not the same (Kremer, Siemsen, & Thomas, 2015). The integration of human judgement in forecasting models on SKU per PoS level may create an additional advantage, in that the algorithm can determine in which PoS the factors taken into account by the expert are most influential.

Fourth, to the best of our knowledge, this is the first study that includes data on both sales numbers and profit margins. Steenburgh et al. (2003) already showed that even small improvements in predictive performance can have a serious impact on a firms profitability. In order to investigate both aspects, this study makes a distinction between the forecasting system and the optimization system, as recommended in previous research (Fildes, et al., 2009). The forecasting system aims to predict the demand of a product, whereas the goal of the optimization model is to maximize profit, taking price, production cost, delivery cost, recollection cost and expected revenue into account. Whereas the forecasting system enables expressions of accuracy with measures such as MAPE or MdAPE, the optimization model indicates the results via a profitability metric.

2. Background Literature

The potential for accurate demand forecasting has only increased because of the exponential increase in computational power, the rise of the internet and the decrease in data warehousing costs. However, due to accelerated product lifecycles and unpredictable customer demand, forecasting remains challenging (Merzifonluoglu, 2015). Data analytics and human judgment need to be combined to achieve better results, rather than being seen as a dichotomy (Ransbotham, Kiron, & Prentice, 2016). However, it has proven to be a significant challenge to successfully combine the analytical power of computers with human judgment of organisational forecasters. Forecasters usually have a choice between leaving the prediction of the statistical model as it is, or adjusting it in a number of ways. Judgmental adjustment takes the statistical model as starting point, and allows for an adjustment based on human judgment. Given its cost efficiency and it being easy-to-use, it is not surprising that this method is common practice (Turner, 1990). The downside is the possibility of biases and unnecessary adjustments (Eroglu & Croxton, 2010; Webby & O’Connor, 1996). The latter is hypothesized to be caused by the illusion of control effect, where people have increased confidence in forecasts they have adjusted (Kotteman, et al., 1994), and the tendency of humans to see patterns in noise (Fildes, et al., 2009). On the other hand, forecasting models have difficulties incorporating the effect of exceptional events such as promotions or new product launches (Goodwin, 2002; Goodwin & Fildes, 1999; Scarpel, 2015). As a result, expert judgment is complementary with the process of mechanically analysing large amounts of data (Alvarado-Valencia, Barrero, Önkal, & Dennerlein, 2016; Blattberg & Hoch, 1990; Goodwin, 2002; Goodwin, et al., 2011). The question remains how to optimally integrate expert judgment with forecasting models. Recently, Alvarado-Valencia, Barrero, Önkal and Dennerlein (2016) compared three integration methods: judgmental adjustment, 50/50 combination and divide-and-conquer (restricting information access). Judgmental adjustment performed best, possibly because the other techniques restricted access to information. They conclude that bias reduction through information restriction does not work. Other research with decision support systems has illustrated that pointing the forecasters towards the damaging effect (in terms of forecasting accuracy) of their adjustments is not the solution. The tendency to adjust persists despite warning (Lim & O'Connor, 1995). This tendency to discount advice is especially troublesome, given that it implies that conscious efforts toward de-biasing judgmental adjustments may be in vain. Goodwin, Fildes, Lawrence and Stephens (2011) focussed on two possibilities to counter harmful judgmental adjustments: restrictiveness (limiting options) and guidance (providing information and explanation). Similar to previous studies in advice taking (e.g., Lim & O'Connor, 1995), the provision of guidance had no significant effect. Restricting users options to make adjustments to the forecast even led to a significant reduction in forecasting accuracy. An additional problem poses itself in the case of restricting judgmental adjustment: the acceptability of the system in the eyes of the forecaster (Goodwin, et al., 2011). Reducing harmful adjustments can come at the price of reduced acceptance (Goodwin, et al., 2011), feelings of loss of control and ownership (Goodwin, 2002) and decreased trust of the final forecast (Önkal, et al., 2009).

While some researchers work on de-biasing the judgmental component, others focus on improving predictive performance by applying more advanced statistical techniques (e.g., Kourentzes & Petropoulos, 2015). However, surveys have indicated that judgment persists in business forecasting (Fildes & Goodwin, 2007) and simultaneously, that forecasting accuracy in business practice is not improving (Armstrong, Green, & Graefe, 2013). We therefore propose to re-visit the possibility of incorporating human judgment factors in the model itself in order to improve the accuracy of predictions (Franses & Legerstee, 2011). This way of working has the benefit of potentially mitigating the harmful effects of judgment in two ways: first, by discounting those judgments that are biased or with great error, based on the historic performance. Second, it takes the widespread practice of incorporating judgment in forecasting into account. Forecasters will still be able to submit their judgment and show that they are attending to the task (Fildes, et al., 2009). Importantly, forecasters will not be put off by the system (Silver, 1991) and retain their sense of ownership. The model will incorporate human judgment as a predictor if these adjustments have proved to add value in the past. In sum, while previous attempts at optimizing the judgment-statistics combination via a forecast support system have demonstrated potential damaging effects on forecaster performance and forecasting accuracy, the integrative model collects input from both sources, and should therefore mitigate potential harmful effects of human judgment, while recognizing the role of acceptance of the forecaster/user. In the dataset of Franses and Legerstee (2011), the judgmental component proved only beneficial if the statistical model was not performing well. Overall, the original statistical model does not seem to improve by including judgment. However, ignoring the value of expert judgment can be especially dangerous in time series with disturbances, such as promotions and other exceptional events. We therefore propose an integrative approach that can be of significant value for industries dealing with exceptional events, such as promotions.

3. Integrative judgment forecasting model

In order to gain a more in-depth understanding of the effects of integrative judgment versus restrictive judgment, we investigate adjustment sizes, direction of the adjustment and the volatility of the data series. Previous research has shown that downward adjustments are more likely to be beneficial than upward adjustments (Fildes, et al., 2009; Franses & Legerstee, 2009; Syntetos, Nikolopoulos, Boylan, Fildes, & Goodwin, 2009). A possible explanation here is that downward adjustments are only made in the presence of evidence that a downturn may arise, while upward adjustments are mostly a reflection of over-optimism and wishful thinking of the forecaster (Fildes, et al., 2009). An integrative approach should counter the negative effects of positive adjustments as follows: In the integrative model, the historical human judgment forecast is added as an additional predictive variable. The model determines whether or not to take this variable into account based on past performance. Only if adjustments have been historically sufficiently accurate and have added value, the model will take them into account. Goodwin et al. (2011) state that a forecast support system which integrates judgment and statistics, should support two stages of this task: first, the decision whether or not the adjustment adds value and second, the determination of how large the adjustment should be. The integrative approach takes this into account, by looking at the significance and weight of the human judgment predictor.

In addition to the direction of the adjustment, previous research suggests a relationship between the size of the adjustment and forecasting accuracy, such that mostly big adjustments are beneficial and small adjustments should be avoided as to not harm forecasting accuracy (Fildes, et al., 2009; Syntetos, et al., 2009). This can be explained by the tinkering with data effect, where forecasters make small adjustments to show that they are working on the task and feel responsible and in control of the forecasting process (Fildes, et al., 2009). Large adjustments on the other hand, are an indicator of knowledge available to the forecaster that is not yet incorporated in the system. However, Fildes et al. (2009) find a tipping point in the case of negative adjustments, such that very large adjustments are equally damaging for forecasting accuracy. Indeed, the effect of adjustment size on forecasting accuracy is curvilinear (inverted U-shape), such that both small and very large adjustments are damaging for forecasting accuracy. Similarly, Trapero et al. (2013) found adjusted forecasts of promotional periods to be potentially beneficial, but not when they were overly large. In other words, the integrative approach should mitigate the damaging effects of both too small and too large adjustments.

A third factor that may play a role in forecast accuracy is volatility. Forecasting models typically have problems dealing with exceptional events (Goodwin & Fildes, 1999). These exceptional events can occur in several ways, for example the occurrence of the Olympic games when predicting aviation traffic in a country or predicting the sales of a special issue of a magazine. Since little past information about these events is available, it is difficult for computerized models to take this effect into account. Especially in these situations, the incorporation of human judgement can add significant value and could enhance the accuracy of the models (Goodwin & Fildes, 1999). In other words, expert judgment can prove beneficial in volatile data series. However, a distinction must be made between volatile data series because of special events or because of noise. In the former case, human judgment is expected to outperform models (Sanders & Ritzman, 1992) while in the latter case, models outperform judgment (O'Connor, et al., 1993). In this study, volatility is determined by the presence of promotions. In concordance with previous literature, we therefore expect that human judgment will have added value over the statistical model if promotions are present (high volatility). In contrast, in low volatility series (periods without promotions), judgment is expected to damage forecasting accuracy (Sanders & Ritzman, 1992). The damaging effect of judgment in low volatility series should be mitigated by the integrative approach, compared to the restrictive approach.

In addition to the above mentioned qualities of the data, this study decomposes forecasting models on different levels of granularity. Kremer et al. (2015) indicated the lack of cross-over between judgmental forecasting and hierarchical forecasting research. Yet, in many business situations, the forecasted SKUs are distributed over multiple points of sales (PoS). Hence, an organization has the choice between generating forecasts on SKU level and rolling this down to SKU per PoS level by using business rules, or making the forecasts on PoS level directly per SKU. For statistical forecasts, both options are easily implemented. However, in the latter situation, integrating human judgement is challenging since the number of SKU and PoS combinations is typically very high. This makes it impossible to apply expert adjustments on the lowest level of granularity. As an alternative, in a traditional restrictive approach, the forecasts on SKU per PoS level could be rolled up again on SKU level, which allows adjustments based on human judgement expertise. Next, these corrections would be applied over all PoS equally. However, this equal correction over all PoS is not always optimal. For example, if the expert systematically takes weather into account to make adjustments to the statistical forecast, an equal division would not incorporate geographical location effects per PoS: the influence of weather might be greater in PoS in touristic and commercial areas than in other rural areas. Thus, a trade-off exists between level of detail (granularity) and potential error (Zotteri & Kalchschmidt, 2007). In this situation, the use of an integrative judgement model adds value, in that not all PoS will be treated equally. Since expert adjustments are taken into account as an explanatory variable in the forecasting models on SKU per PoS level, it is able to distinguish the PoS that are typically more influenced by the expert related factors than others.

While accuracy is an important indicator of forecasting performance, this indicator remains in the theoretical realm. Several researchers highlight the need to link forecast accuracy to measures linked to other business performance (Kerkkänen, Korpela, & Huiskonen, 2009; Mahmoud, DeRoeck, Brown, & Rice, 1992; Mentzer, Bienstock, & Kahn, 1999; Moon, et al., 2003). Previous studies have indicated links between improved demand forecasting accuracy and inventory management (e.g., Clarke, 2006; Oliva & Watson, 2009; Syntetos, Nikolopoulos, & Boylan, 2010) and delivery times (Shan et al., 2009). This study focusses specifically on the practical consequences of demand forecasting accuracy on financial results. By doing this, we are able to express the effect of increased forecasting accuracy on the profitability of the organisation. The forecasts are used as an input for an optimization algorithm that determines the ideal number of SKUs per point of sales, taking revenue and costs into account. Hence, the focus of this optimization algorithm is to maximize profitability and not accuracy. Although the optimization algorithm on itself is out of the scope of this research, this enables us to express the results of this study in a more business relevant profit metric. To the best of our knowledge, this is the first study that takes this effect into account.

In sum, this study looks at the potential beneficial effects of integrating expert judgment in the statistical forecasting model. In contrast to Franses en Legerstee (2011), our time series are disturbed by exceptional events (promotions), which should increase the potential for an added value of judgmental intervention. We analyse our dataset in-depth according to the paper of Fildes et al. (2009) to provide a basis for comparison of the datasets characteristics. Additionally, we follow Kremer et al. ‘s (2015) recommendation to distinguish between the different hierarchical levels of forecasting. Lastly, we believe this to be the first study that links changes in forecast accuracy to a measure of direct profitability.

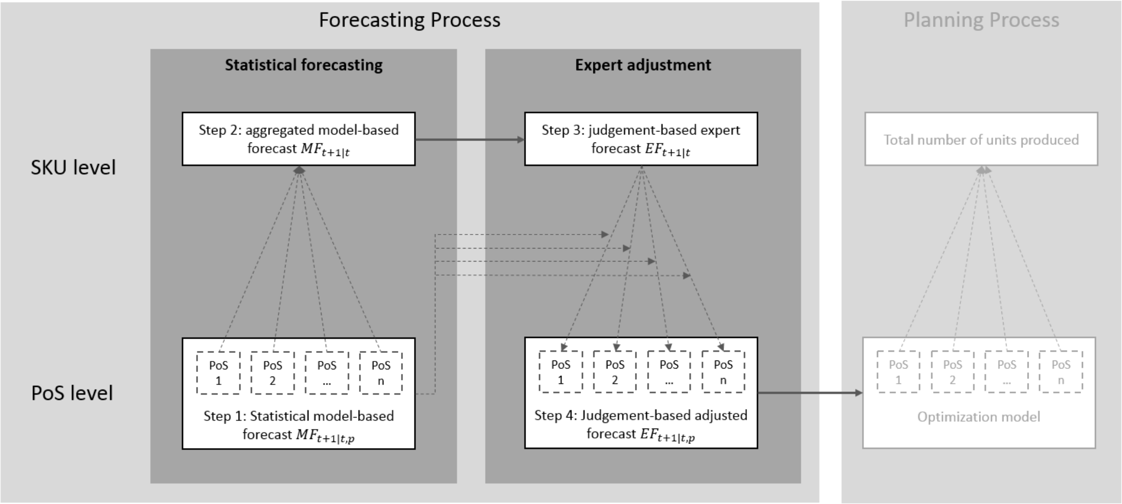
4. Materials and Methods

4.1. Data

Data from a European publishing company were collected. The company distributes weekly and monthly magazines to 5902 points of sales. The data included statistical system demand forecasts, judgemental forecasting adjustments, optimization recommendations for maximizing profit and corresponding actual outcomes. The products are divided into two categories: regular products and exceptional products. An exceptional product is defined as a regular product with an extra: a magazine that includes another magazine, a collectible, dvd, cd, etcetera. In total 3 575 263 data points were collected over a time period of 16 months. These data points contain forecasting data on SKU and PoS level. This data can be rolled up to 1312 aggregated forecasts on SKU levels. Data cleaning led to the deletion of cases with missing information, magazines with less than 100 sales (in comparison, mean sales per SKU = 30 604) (Fildes, et al., 2009), and two cases in which the system forecast deviated extremely from the final sales (APE larger than 2000%), resulting in a final of 1223 aggregated forecasts. Out of those 1223, 850 were classified as regular products (69.50%) and 373 as exceptional products (30.50%) by the company. The forecasting process was discussed with the company and follows a fixed procedure discussed below.

4.2. Methodology

Figure 1 shows the original forecasting process of the company. This process takes place on two levels of granularity: SKU per Point of Sales (PoS) and SKU level (the aggregated demand forecast per magazine).



Forecasting Process
Planning Process
Statistical forecasting
Expert adjustment
SKU level
PoS level
Step 2: aggregated model-based forecast $MF_{t+1|t}$
Step 3: judgement-based expert forecast $EF_{t+1|t}$
Total number of units produced
PoS 1
PoS 2
PoS …
PoS n
Step 1: Statistical model-based forecast $MF_{t+1|t,p}$
Step 4: Judgement-based adjusted forecast $EF_{t+1|t,p}$
Optimization model

Figure 1. Restrictive Judgment Process

In a first step, the demand per PoS is estimated using a statistical forecasting support system. This forecast is denoted as $MF_{t+1|t,p}$, which represents the statistical forecast for Point of Sales p for time t+1 at origin t. This forecast is generated by several time series based forecasting techniques such as moving average, exponential smoothing or ARIMA models. Each of these techniques is based on lagged sales data. We will illustrate this using an autoregressive model of order two, similar to the one used in Franses and Legerstee (2011) :

$$D_{t,p} = \alpha_{1,p} D_{t-1,p} + \alpha_{2,p} D_{t-2,p} + u_{t,p} \qquad (1)$$

Where $D_{t,d}$ represents the Demand at time t at Point of Sales p and u_t is an unobserved error term. Based on this forecasting support system the statistical model-based forecast for each PoS can be obtained by:

$$MF_{t+1|t,p} = \alpha_{1,p} D_{t,p} + \alpha_{2,p} D_{t-1,p} \qquad (2)$$

Since the number of PoS is typically very large (i.e., 5902) it is impossible to evaluate each forecast using expert judgment. Therefore, in a second step, this data is rolled up to a higher level using the following equation:

$$MF_{t+1|t} = \sum_{d=1}^{n} MF_{t+1|t,p} \qquad (3)$$

Where n represents the total number of PoS. In a third step, this aggregated forecast is sent to an expert, who can adjust the forecast. This judgement-based expert forecast can be denoted as $EF_{t+1|t}$, which represents the expert forecast on an aggregated SKU level for time t+1 at origin t. This expert forecast can be formulated by the following equation (Franses & Legerstee, 2011) :

$$EF_{t+1|t} = \lambda MF_{t+1|t} + \beta_t X_{t+1|t} \qquad (4)$$

This assumes that the judgement-based expert forecast can be decomposed into a weighted average of the statistically-based model forecast and own expert knowledge. This weight is represented by λ, which can range between 0, when the expert ignores the model forecast, and 1, when the expert completely accepts the model forecast. It is important to mention that in practice, a company often does not know how much the expert relies on the model forecast. In addition, equally frequently unknown are the factors that are taken into consideration by the expert to adjust the forecast, represented by $X_{t-1|t}$in equation (4) (Fildes & Goodwin, 2007). In a fourth step, these adjusted forecasts are drilled down again over all PoS equally as followed:

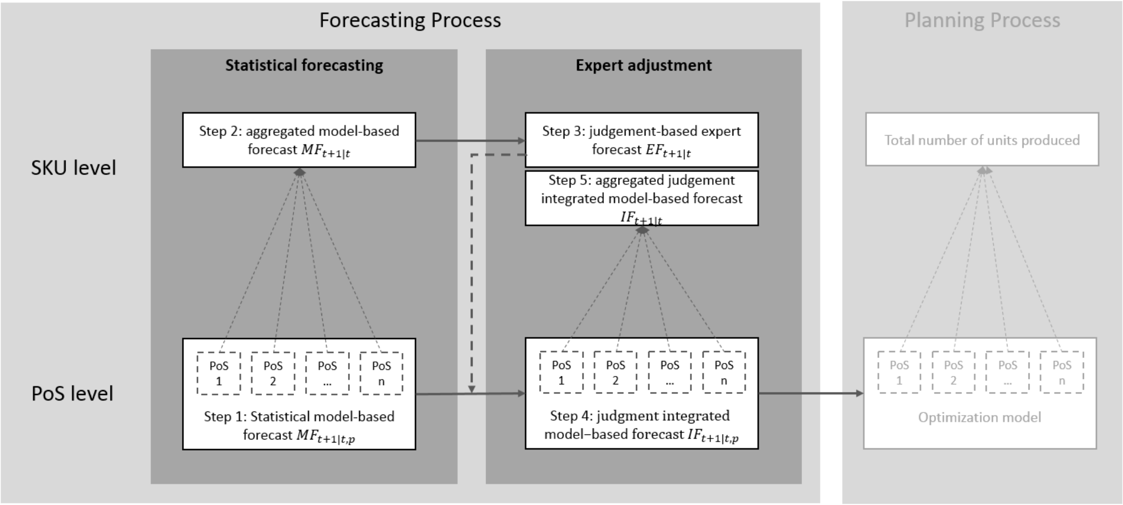
$$EF_{t+1|t,p} \ = (1+\delta)\ MF_{t+1|t,p} \tag{5}$$

Defining

$$\delta = \frac{EF_{t+1|t} - \ MF_{t+1|t}}{MF_{t+1|t}}$$

This forecast on PoS level can be passed on to planning and used as input for an optimization model that maximises the expected profit. This process is further referred to as restrictive judgment, since the judgmental adjustment of the forecaster serves as a restriction on the model output (see Figure 1).

In the integrative model (Figure 2) the beginning of the process is the same as currently used in the company and has been described above: forecasting support systems based on past sales provide a statistically-based forecast in each PoS. This is aggregated to SKU level and presented to the expert for evaluation and adjustment.



Forecasting Process
Planning Process
Statistical forecasting
Expert adjustment
SKU level
PoS level
Step 2: aggregated model-based forecast $MF_{t+1|t}$
Step 3: judgement-based expert forecast $EF_{t+1|t}$
Step 5: aggregated judgement integrated model-based forecast $IF_{t+1|t}$
Total number of units produced
PoS 1
PoS 2
PoS …
PoS n
Step 1: Statistical model-based forecast $MF_{t+1|t,p}$
Step 4: judgment integrated model–based forecast $IF_{t+1|t,p}$
Optimization model

Figure 2. Integrative Judgment Process

However, in contrast with the previous method, the adjusted forecast is not directly drilled down to PoS level afterwards. Rather, the judgment-based expert adjustments now serve as a predictor variable for the statistically-based forecasting model on PoS level, which can be formulated as (Franses & Legerstee, 2011) :

$$D_{t,p} = \alpha_{1,p}D_{t-1,p} + \alpha_{2,p}D_{t-2,p} + \tau_p \beta_t X_{t+1|t} + u_{t,p} \tag{6}$$

Based on equation (4) this can be rewritten as

$$D_{t,p} = \alpha_{1,p}D_{t-1,p} + \alpha_{2,p}D_{t-2,p} + \tau_p(EF_{t|t-1} - \lambda MF_{t|t-1}) + u_{t,p} \tag{7}$$

Defining $\tau_p = \beta_{1,p} + \beta_{2,p}$ and $\tau_p\lambda = \beta_{2,p}$, equation (7) can be rewritten as

$$D_{t,p} = \alpha_{1,p}D_{t-1,p} + \alpha_{2,p}D_{t-2,p} + \beta_{1,p}EF_{t|t-1} + \beta_{2,p}(EF_{t|t-1} - \lambda MF_{t|t-1}) + u_{t,p} \tag{8}$$

Equation (8) shows that this statistically-based forecasting model includes the judgement-based expert forecast on SKU level. The more systematically accurate the expert forecast $EF_{t|t-1}$was in the past, the higher the relevance of this variable will be in the forecasting support system. The model allows parameters $\beta_{1,p}$ and $\beta_{2,p}$to vary depending on the PoS. By this the forecasting support system can distinguish between PoS that are systematically more or less influenced by the unobserved factors taken into account by the expert. Based on this forecasting support system, the judgment integrated model-based forecast $IF_{t+1|t,p}$on PoS level can be expressed as:

$$IF_{t+1|t,p} = \alpha_{1,p}D_{t,p} + \alpha_{2,p}D_{t-1,p} + \beta_{1,p}EF_{t+1|t} + \beta_{2,p}(EF_{t+1|t} - \lambda MF_{t+1|t}) \tag{9}$$

This process is termed integrative judgment (see Figure 2), since the judgment-based expert adjustment is integrated in the forecast equation. Note that in a situation in which no aggregated model forecast is proved (i.e. step 1 until 3 is skipped in the forecasting process), as a result of a lack of time or standardized forecasting process, or when the expert totally ignores this input, a judgment integrated model-based forecast can still be generated. In this specific situation the weight = 0, which reduces equation (9) to:

$$IF_{t+1|t,p} = \alpha_{1,p} D_{t,p} + \alpha_{2,p} D_{t-1,p} + \beta_{1,p} EF_{t+1|t} \tag{10}$$

Finally, in fifth step, these final integrative forecasts on PoS level could still be rolled up to SKU level, defined by $IF_{t+1|t}$, to determine the aggregated demand per SKU:

$$IF_{t+1|t} = \sum_{d=1}^{n} IF_{t+1|t,p} \tag{11}$$

Although the forecast on PoS level expressed in equation (10) will be used for planning, this aggregated forecast is still interesting for comparing the integrative approach with the traditional restrictive approach defined in equation (4).

Since the goal of the forecasting models is to maximize prediction accuracy, this comparison can be expressed in metrics such as Mean Absolute Percentage Error (MAPE) and Median Absolute Percentage Error (MdAPE). However, the forecasting evaluation in business practice goes beyond accuracy, as it serves as a starting point for planning (Kremer, et al., 2015; Mahmoud, et al., 1992; Mentzer, et al., 1999; Moon, et al., 2003). In order to plan the optimal number of units in each PoS, an optimization model is used with the forecasted demand on PoS level as input (see right-hand side on Figures 1 and 2). This algorithm has the goal to maximize profitability taking potential costs into account. More specifically, profitability can be expressed by the following equation:

$$P_{t,p} = R * S_{t,p} - I * Q_{t,p} - O * (Q_{t,p} - S_{t,p}) \tag{12}$$

With $P_{t,p}$ being the operational profit at time t in PoS p (excluding all overhead and fixed costs), R being the revenue earned from selling one magazine, $S_{t,p}$ being the number of magazines sold at time t in PoS p, I the cost to put a magazine in the market (printing + distribution cost), $Q_{t,p}$ the number of magazines distributed to PoS p at time t and O being the cost to get an unsold magazine out of the market.

The outcome of the optimization model is the optimal quantity $Q_{t,p}$ to be delivered at time t per PoS p, that maximises the expected profit taking the expected demand of the forecasting model, the inventory cost and opportunity cost into account. This can be aggregated to give the total number of units to be produced. The focus of this study is on the forecasting process by comparing a restrictive judgment and integrative judgment approach; however, the effect on the planning process will be discussed as well, by expressing the results in a profitability metric.

5. Results

5.1. Descriptive statistics

Judgemental Adjustment, defined as JA, can be measured as followed:

For a restrictive approach:

$$JA = 100 * \frac{EF_{t+1|t} - MF_{t+1|t}}{MF_{t+1|t}} \tag{13}$$

For an integrative approach:

$$JA = 100 * \frac{IF_{t+1|t} - MF_{t+1|t}}{MF_{t+1|t}} \tag{14}$$

Table 1 indicates the relative adjustment sizes for the restrictive and integrative judgment model respectively, with the number of adjustments distributed evenly over four quantiles. The adjustment sizes of the integrative model are noticeably smaller than those of the restrictive model, which shows that the integrative approach is more careful in adapting based on judgemental expertise than the restrictive approach.

	Restrictive Judgment	Integrative judgment
25% Quantile	4.93%	0.00%
Mean	24.51%	4.74%
Median	11.94%	.67%
75% Quantile	31.01%	3.97%

Table 1. Quartiles, Mean and Median of percentage adjustments. N = 1223

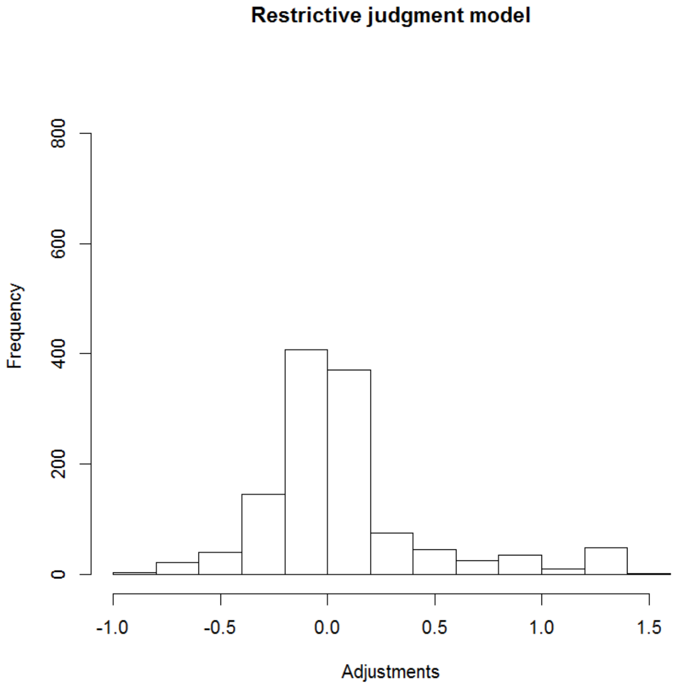
The data was checked against outliers above 250%, similar to Fildes, et al. (2009). No cases needed to be deleted: the largest adjustment made in both the restrictive and the integrative model was 149,55%. Table 2 shows the mean and median sizes of the relative adjustments according to the direction of the adjustment (positive indicating an adjusted forecast that is larger than the system forecast).

	Restrictive judgment			**Integrative judgment**		
	N	**Mean**	**Median**	**N**	**Mean**	**Median**
Downward	587 (48%)	-17.70%	-12.31%	464 (37.94%)	-5.71%	-1.93%
No adjustment	28 (2.29%)	-	-	375 (30.66%)	-	-
Upward	608 (49.71%)	31.57%	12.12%	384 (31.40%)	8.21%	2.48%

Table 2: Mean and median of adjustments, ordered by direction of adjustment

As the table indicates, for the restrictive judgment model, adjustments were made in 1195 or 97.71% of the cases, of which 587 or 48,00% were downward (Mean adjustment size = -17.70%, Median adjustment size = -12.31%) and 49.71% upward (Mean adjustment size = 31.57%, Median adjustment size = 12.12%). Comparatively, in an integrative judgment model, the prediction of the basic model was deemed already optimal in 375 or 30.66% of the cases, resulting in 464 or 37.94% downward adjustments (Mean adjustment size = -5.71%, Median adjustment size = -1.93%) and 384 or 31.40% upward adjustments (Mean adjustment size = 8.21%, Median adjustment size = 2.48%) .

Figure 3 and Figure 4 show the distribution of the adjustment sizes (percentages expressed as decimals) for the restrictive model and the integrative model respectively. The histograms show a clear difference between the integrative and restrictive approach. The adjustments within the restrictive approach are clearly more frequent and larger than in the integrative approach. In addition, the restrictive approach is somewhat skewed towards positive adjustment, which cannot be observed in the histogram of the integrative approach.



:

Figure 3. Distribution of adjustments in the restrictive model

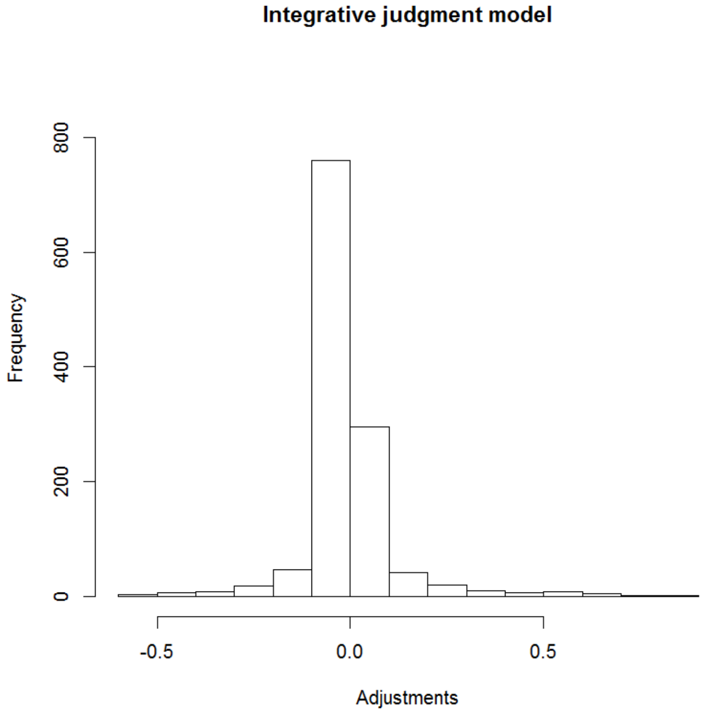


Figure 4. Distribution of adjustments in the integrative model

5.2. Performance

5.2.1. Error metrics

Table 3 indicates the relative performance of the models. Performance is measured as the forecasted demand, compared to the actual sales. The results of MAPE and MdAPE differ such that the median scores are lower than the mean scores, indicating that the data is skewed to the right (similar to Fildes et al., 2009). We therefore report both MAPE and MdAPE as measures of forecasting accuracy. To compare the performance of the restrictive model and the integrative model with the basic statistical model, we use FCIMP (forecast improvement) according to the formula:

For a restrictive approach, FCIMP will be defined as:

$$FCIMP = 100 * \frac{|S_{t+1} - MF_{t+1|t}| - |S_{t+1} - EF_{t+1|t}|}{S_{t+1}} \qquad (15)$$

For an integrative approach, FCIMP will be defined as:

$$FCIMP = 100 * \frac{|S_{t+1} - MF_{t+1|t}| - |S_{t+1} - IF_{t+1|t}|}{S_{t+1}} \qquad (16)$$

This indicates the difference between the APE from the system forecast and the APE of the adjusted forecast (Fildes, et al., 2009), based on a restrictive or on an integrative approach. For those comparisons where the observations were correlated (general comparison, volatility and frequency), we used a bootstrapped paired t-test with 1000 samples for comparing the MAPE, FCIMP and profit. In order to compare the MdAPE, we compared the medians with a Wilcoxon signed rank test. For those comparisons where the observations were partially correlated (direction and size of adjustments), we used a bootstrapped weighed t-test (with 1000 samples) as proposed by Samawi and Vogel (2014). Using a similar method, aWilcoxon Rank Sum Test is combined with a Wilcoxin Signed Rank Test to compare the MdAPEs of partially correlated samples. In general, integrative judgment outperforms restrictive judgment and the basic statistical model in terms of the mean and median absolute percentage error (Table 3).

	Basic statistical model (no judgment)	Restrictive Judgment model	Integrative judgment model
MAPE	25.03%	26.18%$^{\text{BSM: n.s.}}$	21.88% $^{\text{BSM: ***/RM: **}}$
MdAPE	9.95%	10.83% $^{\text{BSM: n.s.}}$	8.24% $^{\text{BSM: *** /RM: ***}}$
FCIMP	-	-1.15%	3.15% $^{\text{RM: **}}$

Note. Proposed models are compared with the basic statistical model (BSM) and each other (RM).

Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 3. Model performance expressed as MAPE, MdAPE, and FCIMP (%)

Paired t-tests indicate that the integrative model (MAPE = 21.88%) outperforms the basic statistical model (MAPE = 25.03%; *t* (1222) = 8.870, *p* < .001), and the restrictive model (MAPE = 26.18%; *t* (1222) = 2.97, *p* = .003). The difference between the restrictive judgment model and the basic model does not appear to be significant (*t* (1222) = -.76, *p* = .445). To compare the MdAPE scores, a Wilcoxon Signed Rank Test was used. Similar to the MAPE scores, the integrative model (MdAPE = 8.24%) outperforms the basic statistical model (MdAPE = 9.95%; *z* = 9.18, *p* < .001), and the restrictive model (MdAPE = 10.83%; *z* = 5.78, *p* < .001). The difference between the restrictive judgment model and the basic model is again not significant (*z* = -1.12, *p* = .261).

Digging deeper into the data, we compare both models with regard to the direction of the adjustments, their size, and the volatility of the data.

	Restrictive judgment				**Integrative judgment**			
	N	**MAPE**	**MdAPE**	**FCIMP**	**N**	**MAPE**	**MdAPE**	**FCIMP**
Downward	587 (48%)	16.79%	9.05%	18.24%	464 (37.94%)	24.67%***	8.98%$^{\text{n.s.}}$	5.01%***
No adjustment	28 (2.29%)	53.84%	32.76%	-	375 (30.66%)	19.97%	6.51%	-
Upward	608 (49.71%)	33.97%	12.63%	-19.91%	384 (31.40%)	20.38%*	9.02%***	1.97%***

Note. Integrative judgment model is compared with the restrictive judgment model.
Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 4. MAPE, MdAPE , and FCIMP, ordered by direction of adjustment.

With regard to the direction of the adjustments (Table 4), our dataset confirms previous literature, which indicates upward adjustments as damaging to forecasting accuracy (FCIMP = -19.91%, compared to the basic model), while downward adjustments are beneficial (FCIMP = +18.24%, compared to the basic model) in a traditional restrictive approach. However, when judgment is included in an integrative way, both upward (FCIMP = +1.97%) and downward (FCIMP = +5.01%) adjustments are beneficial.

The integrative approach thus mitigates the severely damaging effects of restrictive judgmental adjustment. However, while the damaging effect of upward adjustments is neutralized and even translated into a small beneficial effect, the beneficial effects of downward adjustments are tempered (from 18.24% improved accuracy with restrictive judgment, to 5.01% improved accuracy in the case of integrative judgment). These results show that, on average, the integrative approach is able to ignore the adjustments that are consistently over-optimistic and mainly incorporates only useful positive adjustments.

Adjustment size	**Restrictive judgment**				**Integrative judgment**			
	N	**MAPE**	**MdAPE**	**FCIMP**	**N**	**MAPE**	**MdAPE**	**FCIMP**
No adjustment	28	53.85%	32.76%	-	375	19.97%	6.51%	-
Q1	298	12.49%	5.86%	-0.03%	212	28.44%*	8.67%***	0.06%*[1]
Q2	298	13.66%	8.49%	-1.47%	212	18.11%***	7.62%$^{n.s.}$	0.44%$^{n.s.}$
Q3	298	18.55%	12.51%	5.99%	212	21.25%$^{n.s.}$	8.63%*	1.57%***
Q4	298	57.37%	39.29%	-8.89%	212	23.12%**	12.82%***	16.11%***

Note. Integrative judgment model is compared with the restrictive judgment model.
Note. 1 the value was marginally significant at .051.
Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 5. Adjustments ordered by size in four quantiles: MAPE, MdAPE, and FCIMP

With regard to adjustment size for the restrictive model (Table 5), we find a curvilinear relationship similar to Fildes et al. (2009) between adjustment size (expressed in four quantiles) and forecasting accuracy, such that small adjustments (Q1= -0.3% and Q2= -1.47%) and overly large adjustments (Q4= -8.89%) are damaging. Only medium adjustments proved to be beneficial (Q3 = 5.99%). However, when judgment is included in an integrative way, the damaging effects of Q1, Q2 and Q4 are translated into a beneficial effect, showing a concave function with Q1 = 0.06%, Q2 = 0.44%, Q3 = 1.57% and Q4 = 16.11%.

While adjustments are always beneficial in the case of integrative judgment, the large beneficial effect of Q3 with restrictive judgment is limited (a decline in beneficial effect from 5.99% to 1.57%). However, the severely damaging effect of overly large adjustments (-8.89%) is translated in a highly beneficial effect (16.11%). To calculate volatility, we employ two different procedures. First, we calculate volatility as the coefficient of variation of the system forecast absolute error (Fildes, et al., 2009). The resulting volatility scores are divided into four quantiles, ranging from low to high (Table 6).

Volatility (SD)	**Restrictive judgment**				**Integrative judgment**			
	N	**MAPE**	**MdAPE**	**FCIMP**	**N**	**MAPE**	**MdAPE**	**FCIMP**
SD Q1	305	54.25%	31.91%	-28.84%	305	24.81%***	12.95%***	0.60%***
SD Q2	306	26.20%	14.68%	6.40%	306	28.95%$^{n.s.}$	15.46%$^{n.s.}$	3.66%$^{n.s.}$
SD Q3	306	13.16%	7.53%	10.87%	306	18.27%$^{n.s.}$	6.42%$^{n.s.}$	5.76%$^{n.s.}$
SD Q4	306	11.19%	7.31%	6.89%	306	15.51%***	4.90%***	2.57%$^{n.s.}$

Note. Integrative judgment model is compared with the restrictive judgment model.
Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 6: Performance ordered by volatility according to SD: MAPE, MdAPE, and FCIMP

In the case of restrictive judgment, judgmental adjustments are especially troublesome in the lowest quantile, with a reduction of -28.84% in forecasting accuracy. The other quantiles with medium to high volatility on the other hand, show a forecasting improvement of 6.40% (Q2), 10.87% (Q3) and 6.89% (Q4). In the integrative model, there is a slight forecasting improvement in all quantiles and ranges of volatility, displaying a curvilinear effect with Q1 = .6%, Q2 = 3.66%, Q3 = 5.76% and Q4 = 2.57%. The change in FCIMP from the restrictive model to the integrative model is significant for Quartile 1 (*t* (304) = -8.82, *p* < .001), marginally significant for Quartile 3 (*t* (305) = 1.91, *p* = .056), but not significant for Quartile 2 (*t* (305) = 1.30, *p* = .196) or Quartile 4 (*t* (305) = 1.54, *p* = .126). Second, we follow the company's categorization in regular products (considered easy to predict) and exceptional products (considered difficult to predict) as an analogy for low and high volatility (Table 7). The correlation with the previous metric equals r = 0.3468, suggesting a relationship but not a full overlap. For instance, within the low volatility category, there is still an effect of volatility as variation of the system forecast absolute error.

Volatility (category)	**Restrictive judgment**				**Integrative judgment**			
	N	**MAPE**	**MdAPE**	**FCIMP**	**N**	**MAPE**	**MdAPE**	**FCIMP**
Exceptional products	373	26.00%	14.58%	24.42%	373	46.70%***	21.56%***	3.72%***
Regular products	850	26.26%	9.25%	-12.36%	850	10.99%***	5.33%***	2.90%***

Note. Integrative judgment model is compared with the restrictive judgment model.
Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 7. Performance ordered by volatility according to category: MAPE, MdAPE, and FCIMP

The results of the restrictive model confirm that judgmental adjustment in case of low volatility (regular products) is damaging (FCIMP = -12.36%) and is beneficial in the case of high volatility (FCIMP = 24.42%). In the integrative model, the damaging effect of judgment in case of low volatility is eliminated and translated into a positive, albeit small improvement in forecasting accuracy (FCIMP = 2.90%; *t* (372) = 6.10, *p* < .001). In case of high volatility, the forecast improvement is 3.72% (*t* (849) = -11.91, *p* < .001). Logically, for non-volatile products, the statistical method is quite adept at predicting demand. Judgmental adjustment will on average be more harmful (e.g., due to tinkering with the data) than beneficial (e.g., due to having additional information). The parameter of the judgmental component will consequently be low to zero in the case of non-volatile products. However, in situations with high volatility, expert adjustment is beneficial. However, the integrative approach limits the effect.

In sum, the integrative model consistently outperforms the basic statistical model. However, in those cases where the restrictive model has a beneficial effect compared to the basic model, this improvement is larger than with the integrative model.

5.2.2. Profitability

The output of the prediction model serves as input for an optimization model, which calculates the optimal number of magazines for each PoS. The optimization enables an expression of model performance in profatibility. To guarantee anonymity of the company and protect confidential data, we report the measure of profit expressed as a percentage of the hypothetical maximum profit. The hypothetical maximum profit is reached if every PoS would sell the exact amount of predicted magazines. In this situation, no product would have to be returned or none would have to be delivered additionally, maximizing profit. The results of the optimization model for restrictive judgment and integrative judgment respectively, are compared to this perfect situation. The result is a percentage which indicates how close the optimization model comes to the ideal situation. This amounts to a profit percentage of 82,00% for the basic statistical model (see Table 8). This implies that there is still a potential to improve profitability with 18% if the forecast in each PoS would be perfectly accurate.

	Basic statistical model (no judgment)	Restrictive Judgment model	Integrative judgment model
MAPE	25.03%	26.18% $^{\text{BSM: n.s.}}$	22.54% $^{\text{BSM: ***/RM: **}}$
MdAPE	9.95%	10.83% $^{\text{BSM: n.s.}}$	8.24% $^{\text{BSM: ***/RM: ***}}$
FCIMP	-	-1.15%	2.49% $^{\text{RM: **}}$
Profit	82.00%	77.17% $^{\text{BSM: ***}}$	82.80% $^{\text{BSM: ***/RM: ***}}$

Note. Integrative judgment model is compared with the restrictive judgment model.
Note. Significance levels: * p < .05, ** p < .01, *** p < .001, n.s. not significant

Table 8: Model performance expressed as MAPE, MdAPE, FCIMP and profit (%)

The restrictive model demonstrates a damaging effect on profit, reducing the profit percentage to 77,17% (a decline of 4.83% compared to the statistical model; *t* (1222) = 6.91, *p* < .001). The integrative model however, displays a positive effect on profit compared to the statistical model (an improvement of .80%), with a profit percentage of 82.80% (*t* (1222) = -9.58, *p* < .001). This shows that by simply adapting the forecasting process, significant profits can be obtained by the organization.

6. Discussion

6.1. Summary of findings

This study builds further on previous judgmental forecasting research, which indicates that restrictive human judgment proves to be valuable only in specific cases. The method of incorporating human judgement predictions in an integrative way suggests a way to counter the harmful effects of upward adjustments, small or overly large adjustments and adjustments made in the case of low volatility. Previous research has extensively proven that people have a tendency to adjust forecasts, even if they have no additional information (Lawrence, et al., 2006). Forecasters tend to have increased confidence in forecasts they have adjusted (Kotteman, et al., 1994) and tinker unnecessarily with the outcomes of the statistical model, simply to show they are paying attention to the task (Fildes, et al., 2009). These unnecessary adjustments lead to a general decline in forecasting accuracy. The integrative approach counters the harmful effects of these adjustments. Indeed, the methodology has a positive effect on accuracy in all scenarios. However, in those cases where the restrictive approach proved to be beneficial (medium sized and big adjustments, downward adjustments and adjustments in case of high volatility), the integrative approach limited these beneficial effects.

6.2. Managerial contributions

The integrative approach can prove beneficial for business practice on four different aspects: accuracy, process, profit and people management. First, the integrative approach has proven to be able to counter any damaging effects that existed in the forecasts because of judgmental factors. While the traditional restrictive approach demonstrated the same pitfalls (damaging small adjustments, positive adjustments) as found in previous literature (e.g., Fildes et al., 2009), the integrative approach cancelled out all effects that were detrimental for forecasting accuracy. Classic forecast accuracy metrics such as MAPE and MdAPE show a reduction in error rates and thus an improvement in general forecasting accuracy.

Second, the integrative approach can lead to process improvement. The model enables a tailored drill-down of judgmental adjustment to the different PoS. Manual adjustment per PoS would require a labour intensive way of adjusting forecasts. Using business rules may not always prove accurate. By using the integrative approach, the model parameters automatically indicate the importance of judgmental adjustment for each PoS separately.

Third, this dataset provided the opportunity to look beyond established theoretical measures of accuracy and provided a picture of profitability by distinguishing between the forecasting system and the optimization system. The integrative approach thus has an even more tangible positive consequence, in that it heightens not only accuracy measures but also heightens profit. While the classic, restrictive approach damaged profit compared to the basic statistical forecasting model, the integrative approach not only mitigated this damaging effect but caused an increase in gained profit on top of the statistical forecasting model. While the percentages expressed in the results section may be seen as small numbers, it is important to note that these correspond to large differences in absolute profit numbers. The translation from error measures to profit margins provides a unique opportunity for a better communication between researchers and practitioners on the importance of improved forecasting accuracy.

Fourth, the integrative approach has a high chance of being seen as acceptable by the forecasters, since it does not take away their input and thus sense of ownership. Previous efforts aimed at reducing error due to faulty judgment in forecasting have focussed on biases associated with judgment, and on de-biasing approaches. A popular technique has been de-biasing via consciously attempting to correct judgment via advice or explicitly pointing towards the declining accuracy. While this technique has the potential to be applicable to biases that can vary over time, judgmental forecasters however, seem to persist in their damaging adjustments (Lim & O'Connor, 1995). Önkal and Gönül (2005) interviewed forecasters and found that they not only adjust to integrate their knowledge, but also to own the forecasts, contribute to the forecasts and gain a sense of control over them. Consequently, when forecasters are advised to leave the forecast alone, this sense of ownership is taken away from them, possibly leading to resistance to this way of working. With an integrative approach on the other hand, the possibility remains for the forecasters to provide their input, while simultaneously correcting for possible damaging adjustments. Indeed, forecasters continue to be asked for their input. Additionally, the integrative model takes the judgment variable into account according to its predictive power in past forecasts. Increased judgmental accuracy in the past will therefore lead to increased impact of the judgment of the forecasters in the present. If the forecasters performs badly, e.g. consistent under-forecasting by a sales forecaster to easily achieve its target, judgment will no longer have a significant influence on the forecast result. This method would thus motivate forecasters to perform accurately and more objectively. Additionally, forecasters can be informed on when to rely on their own judgment (restrictive) and when to rely on the integrative model for better results.

6.3. Limitations and further research

The study has some limitations. First, the dataset contains information about forecasts, adjustments, sales numbers and profit from a single company. However, our analysis finds the same pattern as previous studies (e.g., Fildes et al., 2009), providing an indication that the data is comparable. Similar to previous literature (Fildes et al., 2009), negative adjustments were more profitable than positive adjustments. Similar to Sanders and Ritzman (1992), judgmental adjustment was found to be beneficial in high volatility series, measured both by the forecast error and the classification by the company. In low volatility series, restrictive judgment damaged accuracy. Given the large similarities in data patterns, results should hold in further research with the integrative model in other companies, to further test the robustness of this solution. Additionally, qualitative inquiry is needed to test the acceptance of this new method by the forecast users. Previous research has indicated trust issues with forecast support systems that restrict judgment (Goodwin, et al., 2011).

The proposed method should counter the negative effects of judgmental adjustments, while retaining the positive feelings of the forecaster concerning their input, value and ownership. Last, in this study, including judgment in an integrative way improves performance consistently. Unfortunately, while the integrative judgment was able to translate all damaging judgmental adjustments into beneficial adjustments, it also tempered the magnitude of the beneficial judgmental adjustments. Future research should further look into optimizing the effects of integrative judgment, such that the beneficial effects of judgment remain equally large. A possible avenue is the establishment of application rules with regard to the use of restrictive judgment or integrative judgment.

7. Conclusion

This study looked into a forecasting model which integrated expert adjustment into the model itself. The results show that there is a beneficial effect of integrative judgment, compared to the restrictive approach and the basic statistical model. The dataset was tested extensively according to size of the adjustments, their direction, and the volatility of the data. The latter was tested in two ways: by using the Standard Deviation (Fildes et al., 2009) and by using a category as defined by the company. Results showed similar patterns to the extensive dataset of Fildes et al. (2009), indicating a level of comparability between our dataset and others. Additionally, this study takes into account the call for judgmental forecasting researchers to pay attention to hierarchical levels of forecasts (Kremer, et al., 2015). The analyses show the added value of the integrated approach on both levels of granularity. Lastly, this study was able to provide an insight into the direct financial consequences of improved forecasting accuracy. The relationship with direct financial gain is close to non-existent in judgmental forecasting literature, while it plays a pivotal role for the practitioner. Improving forecasting accuracy in practice is an important task for researchers in our field (Sanders & Manrodt, 2003). This study thus responds to a call for more research with company data (Sanders, 2009). The integrative approach can de-bias judgmental forecasting without negatively affecting feelings of ownership from the forecaster. It improved forecasting accuracy and profitability in a dataset with exceptional events.

8. References

Alvarado-Valencia, J., Barrero, L. H., Önkal, D., & Dennerlein, J. T. (2016). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting, in press*.

Armstrong, J. S., Green, K. C., & Graefe, A. (2013). *Golden Rule of Forecasting: Be Conservative*. Working paper.

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

Clarke, S. (2006). Transformation Lessons from Coca-Cola Enterprises Inc.: Managing the Introduction of a Structured Forecast Process. *Foresight: The International Journal of Applied Forecasting*(4), 21 - 25.

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*(6), 570-576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 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*(1), 3 - 23.

Fildes, R., Goodwin, P., & Onkal, D. (2016). *Information use in supply chain planning*. Lancaster University Dept. Management Science.

Franses, P. H., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting, 25*(1), 35 - 47.

Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts 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?

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega 30, 30*(2), 127 - 135.

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*(1), 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*(3), 242 - 253.

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

Jones, D. R., & Brown, D. (2002). The division of labor between human and computer in the presence of decision support system advice. *Decision Support Systems, 33*, 375 - 388.

Kerkkänen, A., Korpela, J., & Huiskonen, J. (2009). Demand forecasting errors in industrial context: Measurement and impacts. *International Journal of Production Economics, 118*(1), 43 - 48.

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

Kourentzes, N., & Petropoulos, F. (2015). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*.

Kremer, M., Siemsen, E., & Thomas, D. J. (2015). The Sum and Its Parts: Judgmental Hierarchical Forecasting. *Management Science*.

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

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

Mahmoud, E., DeRoeck, R., Brown, R., & Rice, G. (1992). Bridging the gap between theory and practice in forecasting. *International Journal of Forecasting, 8*(2), 251 - 267.

Mentzer, J. T., Bienstock, C. C., & Kahn, K. B. (1999). Benchmarking sales forecasting management. *Business Horizons, 42*(3), 48-56.

Merzifonluoglu, Y. (2015). Risk averse supply portfolio selection with supply, demand and spot market volatility. *Omega, 57A*, 40 - 53.

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). Judgemental forecasting in times of change. *International Journal of Forecasting, 9*, 163 - 172.

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*(2), 138 - 151.

Önkal, D., & Gönul, M. S. (2005). Judgmental adjustment: A challenge for providers and users of forecasts. *Foresight: The International Journal of Applied Forecasting, 1*(1), 13-17.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Ransbotham, S., Kiron, D., & Prentice, P. K. (2016). Beyond the Hype: The Hard Work Behind Analytics Success. *Mit Sloan Management Review, March*.

Sanders, N. R. (2009). Comments on "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning". *International Journal of Forecasting, 25*, 24 - 26.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantiative forecasting methods in practice. *Omega, 31*, 511 - 522.

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.

Scarpel, R. A. (2015). An integrated mixture of local experts model for demand forecasting. *International Journal of Production Economics, 164*, 35 - 42.

Shan, J. Z., Ward, J., Jain, S., Beltran, J., Amirjalayer, F., & Kim, Y. (2009). Spare-parts forecasting: A case study at Hewlett-Packard. *Foresight: The International Journal of Applied Forecasting, 14*, 40-47.

Silver, M. S. (1991). Decisional Guidance for Computer-Based Decision Support. *MIS Quarterly, 15*(1), 105 - 122.

Sinha, A. P., & Zhao, H. (2008). Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems, 46*, 287 - 299.

Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively Categorical Variables: Revealing the Information in Zip Codes. *Marketing Science, 22*(1), 40 - 57.

Syntetos, A., Kholidasari, I., & Naim, M. M. (2016). The effects of integrating management judgement into OUT levels: In or out of context? *European Journal of Operational Research, 249*, 853-863.

Syntetos, A., Nikolopoulos, K., & Boylan, J. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting, 26*(1), 134 - 143.

Syntetos, A., Nikolopoulos, K., Boylan, J., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgement into intermittent demand forecasts. *International Journal of Production Economics, 118*, 72 - 81.

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*(2), 234 - 243.

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

Webby, R., & O'Connor, M. (1996). Judgmental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting, 12*, 91 - 118.

Worthen, B. (2003). Future results not guaranteed; contrary to what vendors tell you, computer systems alone are incapable of producing accurate forecasts. *CIO, 16*(19), 1 - 4.

Zotteri, G., & Kalchschmidt, M. (2007). A model for selecting the appropriate level of aggregation in forecasting processes. *International Journal of Production Economics, 108*(1-2), 74 - 83.

Chapter 3

Enhanced anchoring effects produced by the presence of statistical forecasts: Effects on judgmental forecasting

Enhanced anchoring effects produced by the presence of statistical forecasts: Effects on judgmental forecasting

Abstract

Are judgmental forecasters more accurate when they are given statistical forecasts? People saw graphs of past sales figures, together with information about promotions that affected them. They made forecasts for normal periods and for periods in which promotions were planned. The presence or absence of statistical forecasts was manipulated. In Experiment 1, these forecasts impaired forecasting accuracy relative to that obtained with unaided judgment. To test whether this effect arose because the line on the graph showing the history of past statistical forecasts produced an enhanced visual anchor, Experiment 2 included statistical forecasts only for the future periods that were to be predicted. The detrimental effect of providing statistical forecasts was maintained. In both Experiments 1 and 2, 40% of past sales periods contained promotions. With fewer promotions, the statistical forecasts would be much closer to sales level of the periods without promotions in the displayed series. As a result, forecasting errors due to anchoring should be much smaller on normal periods – and the enhanced anchoring effects produced by the presence of a statistical forecast should matter less. To test this, the proportion of promotions in the past data was reduced to 10% in Experiment 3. This removed the detrimental effect of statistical forecasts for normal periods but not for promotional ones. Our findings provide a cautionary tale in the application of statistical forecasting in time series with disturbances.

Keywords: anchoring, judgmental forecasting

1. Introduction

It is generally accepted that the presence of a statistical forecast has the potential to improve on a forecast made by judgment alone. Both judgment and statistical models have their strengths and weaknesses. Formal models can outperform judgment when it comes to consistency or the processing of large amounts of data. However, formal models tend to have difficulties with discontinuities, unexpected events and external influences (Armstrong & Collopy, 1998; Goodwin, 2002; Hughes, 2001; Taleb, 2007). Judgment is deemed especially useful for dealing with special events such as these. Promotional investments and their effects on future sales numbers are examples of them that have been studied experimentally (Goodwin & Fildes, 1999).

Increasingly sophisticated methods have been developed to enable better promotional forecasting (e.g., Trapero, Kourentzes, & Fildes, 2015; Trapero, et al., 2013). However, a number of difficulties have come to light. People make insufficient use of provided forecasts (e.g., Lim & O'Connor, 1996) and tend to discount 'advice' from a formal model (Goodwin, et al., 2007; Önkal, et al., 2009). Also, surveys have shown that practitioners are slow in adopting formal methods (Lawrence, 2000; Sanders & Manrodt, 2003) and that statistical tools in organizations tend to be distrusted and quickly fall into disuse (Zbaracki, 1998). Because of these problems, much of the literature on judgmental forecasting literature and on forecast support systems has focussed on methods designed to facilitate the acceptance of statistical forecasts by forecasters (e.g., Fildes, et al., 2006; Goodwin, et al., 2007).

The role of judgment cannot be underestimated. Surveys indicate that it plays an important role in practice (e.g., Sanders & Manrodt, 1994; Sanders & Manrodt, 2003), either because practitioners make forecasts using unaided judgment or because they combine their judgment with forecasts derived from a formal model. It is often assumed that the latter produces better results and that our efforts should therefore be geared to increasing its use. However, to date, there is only limited research that has directly compared unaided judgmental forecasting and combined forecasting within the same dataset and group of forecasters. Here, we address this fundamental question: Does provision of statistical forecasts improve the accuracy of judgmental forecasting?

A number of factors provide context to our investigation. First, as mentioned above, formal models can have difficulty in allowing for the effects of sporadic external events. Second, practitioners tend to resist the introduction of sophisticated formal models (Asimakopoulos, 2013; Fildes, et al., 2006). Third, judgment is pervasive in forecasting practice (Fildes & Goodwin, 2007; Goodwin, 2002). Fourth, practitioners with causal information make better forecasts than statistical methods or practitioners without that information (e.g., Armstrong, 1983; Edmundson, Lawrence, & O'Connor, 1988; Lawrence, O'Connor, & Edmundson, 2000).

To develop our hypotheses, we review papers that report experiments or analyse organisational data that include a) judgmental forecasts made with the access to statistical forecasts and b) past and future periods with and without promotions. Thus, we do not cover papers in which no distinction is made between promotion and non-promotional periods (e.g., Fildes, et al., 2009), papers in which forecasting is purely statistical (e.g., Trapero, et al., 2015), or papers in which the size of future promotional effects is estimated without forecasts being made (e.g., Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007).

2. Literature background and development of hypotheses

Lim and O'Connor (1996) first required people to forecast soft drink sales from a time series of past sales. Then they presented them with a) causal information (air temperature) relevant to the forecasting period, b) a statistical forecast that did not take this causal information into account, or c) both the causal information and the statistical forecast. Participants adjusted their original forecast to take this new information into account. It was found that, though forecasts were improved after receiving temperature information, this improvement was small and no greater when a statistical forecast was provided. This may have been because people are conservative when making sequential adjustments (Harvey & Fischer, 1997), because forecasters were presented only with temperature information relevant to only the current forecast period and would have had to mentally integrate information from past periods to judge its effect on sales, or because temperature information was always present and so forecasters could not compare outcomes when it was available with when it was not.

Goodwin and Fildes (1999) used an experimental design that was not affected by these problems. They employed a simple extrapolative forecasting task rather than an adjustment task; no learning was necessary because information about the size of causal factors affecting past data points was presented as vertical bars (labelled 'promotional expenditure') on the same graph as the time series that had to be forecast; promotions were sporadic, affecting about half the time periods in the presented series. In their experiment, series were either simple (independent points scattered around a constant mean) or complex (linear trend with a multiplicative seasonal pattern superimposed) and contained either high or low noise. The relation between the size of the promotion and its effect in elevating sales was always linear but was either weak or strong. Participants were asked to made successive forecasts for the same time series. The outcome of the series was updated before each new forecast was made. Thus, all participants received immediate outcome feedback. Some participants received statistical forecasts based on exponential smoothing of past data cleaned of promotional effects. When these forecasts were given, they were shown for the whole of the presented series as well as for the upcoming required forecast.

Goodwin and Fildes (1999) found that forecasts for non-promotional periods were worse when promotions had a stronger effect on sales: this was presumably because effects of promotions appeared to add a higher level of noise to the series when they were stronger and this impaired forecasting for periods without promotions. They also found that provision of statistical forecasts had no effect on forecasts for periods with promotions and that they improved forecasts for periods without promotions only when series were complex or contained high noise. Thus, despite using a different experimental design from that of Lim and O'Connor (1996), Goodwin and Fildes' (1999, p.

49) conclusions were similar: “The main finding of this study is that, while judgmental forecasters benefited from the availability of statistical forecasts under certain conditions, they almost always made insufficient use of these forecasts. … In Lim and O’Connor’s study subjects had already made an initial forecast before they were presented with the statistical forecast. Our study suggests that this underweighting prevails even when the statistical forecast is presented before the judgmental forecast has been formed”.

Goodwin, Fildes, Lawrence and Stephens (2011) developed Goodwin and Fildes’ (1999) research further. Series were again either relatively simple or more complex and contained either high or low noise. Presentation of the time series and the promotional events was the same as before, though it appears that promotions were less frequent. In this study, participants’ forecasts made with statistical forecasts were not compared to those made without them. Instead the aim was to investigate whether better use would be made of statistical forecasts when large changes away from those forecasts were restricted (i.e., forbidden) or when guidance was given about the appropriateness of making changes away from those forecasts. Thus, participants could use statistical forecasts without restriction or guidance, with restriction but no guidance, or with guidance but no restriction. Analyses showed that guidance failed to improve performance and that restriction impaired it. Generally, judgmental forecasts for promotion periods were better than raw statistical forecasts whereas those for non-promotion periods were worse. These effects did not interact with series type.

In summary, Goodwin and Fildes (1999) found that providing statistical forecasts to support judgment (J +S) produced no increase in accuracy over judgment alone (J) on either promotional or non-promotional periods when series were ‘simple’. (They improved accuracy for non-promotional periods only when series that contained high noise levels and/or complex pattern information.) Goodwin et al (2011) found that judgment combined with a statistical forecast (J+S) outperformed the raw statistical forecast (S) only on promotional periods and that this finding did not differ according to noise level or complexity of the data series. Combining both studies, we hypothesize that, for relatively simple series:

H1: the accuracy of different types of forecast will depend on the period type, such that

H1a: For promotional periods, unaided judgment (J) and judgment combined with a formal model (J+S) will outperform the raw statistical forecast (S).

H1b: For normal periods, the raw statistical forecast (S) and judgment combined with a formal model (J+S) will outperform unaided judgment (J).

Webby, O'Connor and Edmundson's (2005) gave participants a time series that was subject to sporadic perturbations in both upward and downward directions arising from factors such as promotions and competitors' promotions. Perturbations affecting the 48 periods of the presented time series occurred in both directions whereas those affecting the 12 forecast periods were all in the upward direction. Participants were asked a) to remove the effects of the perturbing events from the presented series, b) to make extrapolations from the resulting underlying series for the 12 periods requiring forecasts, and c) to add the effects of the expected perturbations to these extrapolations to produce final forecasts. Results showed that, when removing or adding effects of perturbations, participants were conservative: they made insufficient adjustment to allow for the effects of the perturbations. Thus, it appears that the trend line of the underlying time series acted as a mental anchor and that adjustments relative to it showed the insufficiency that is typically observed when the anchor-and-adjustment heuristic is used (Tversky & Kahneman, 1974). However, these anchoring effects were superimposed on an optimism bias that affected extrapolation of the underlying time series. This optimism bias was greater when more events perturbed the presented and forecast sections of the series. Webby et al (2005) suggest that more events increase forecasters' cognitive load and this leaves them more susceptible to cognitive biases, a notion that is consistent with Kahneman's (2011) two-system theory of cognition.

Optimism (the tendency to over-forecast desirable quantities, such as sales or profits) is a well-established phenomenon in the forecasting literature (Eggleton, 1982; Harvey & Bolger, 1996; Lawrence & Makridakis, 1989). Trapero, Pedregal, Fildes and Kourentzes (2013) found evidence for it in their analyses of data obtained from a manufacturing company. What marked their work out was that they were able to obtain statistical forecasts and final forecasts for both promotional and non-promotional periods: “this is the first case study to employ organizational data for verifying whether judgmental forecasts during promotional periods achieve lower forecasting errors than their statistical counterparts” (Trapero, et al., 2013, p. 235). The dataset comprised 18,096 data triplets (i.e., statistical forecast, final forecast, outcome) from 169 SKUs. Eight percent of the triplets were for promotional periods. Statistical forecasts were based solely on time series information and so took no account of the effect of promotions. In other words, they were not ‘cleaned’ of promotions. Because of this and because promotional periods were relatively rare, accuracy of statistical forecasts was lower for promotional periods than for non-promotional ones. This was at least partly because statistical forecasts were slightly too high for non-promotional periods but much too low for promotional ones.

Analyses showed that final, adjusted forecasts were less accurate than statistical ones, particularly for promotional periods. Overall, adjustment produced final forecasts that were overestimates and considerably higher than the statistical forecasts. The authors attributed this to optimism. More detailed analysis showed that small negative adjustments improved and large positive adjustments impaired accuracy on non-promotional periods whereas small positive adjustments improved and other adjustments impaired accuracy on promotional periods. These patterns are to be expected given the under-forecasting of the statistical model for promotional periods and the over-forecasting for normal periods.

While Webby et al (2005) showed that when forecasters are explicitly presented with time series information, adjustments for events such as promotions were insufficient – presumably due to mental anchoring on the mean or trend line of the presented time series. This was not found by

Trapero et al’s (2013) in their analysis of organizational data. One possible explanation is that forecasters in the manufacturing company did not consider time series information that was long enough to produce the anchoring effect. Goodwin and Fildes’ (2011) survey of company forecasting behaviour showed that forecasts are often based on very short data series (e.g., six points). Thus, our second hypothesis is that when forecasters are explicitly presented with time series information:

H2a: Mental anchoring on the series mean will produce under-forecasting for promotional periods but over-forecasting for non-promotional ones.

Would a statistical forecast help alleviate any such bias? To emulate the real-life forecasts examined by Trapero et al (2013) and Fildes et al (2009) as well as possible, statistical forecasts in this study did not distinguish between promotional and normal periods. It is therefore unlikely that they would lessen an anchoring bias based on this distinction.

H2b: The presence of a statistical model will not affect the mental anchoring bias outlined in H2a.

We expect that the optimism effects identified by Trapero et al (2013) will be overlaid on the anchoring bias. Assuming that the effects of these two biases are additive, over-forecasting of normal periods will be larger than under-forecasting of promotional periods. Thus, our next hypothesis is as follows:

H3a: the positive directional error of normal periods will be greater than the negative directional error for promotion periods.

Statistical forecasts are not subject to optimism. Thus, if people are influenced by statistical forecasts, the optimism bias should be reduced.

H3b: the asymmetry described in H3a will be less in the presence of a statistical forecast

To successfully allow for the effects of promotions when adjusting statistical forecasts, forecasters must appreciate the relation between the size of promotions and the size of their effects. For convenience, we term this relation the 'promotion function'. Our remaining hypotheses focus on biases related to the forecasters' perception of this function.

Range contraction effects (Poulton, 1989) are one of the foundations of range-frequency theory (Parducci, 1965, 1973), a psychophysical model of context effects. When people respond to a range of values on some scale, the range of their responses on that scale tends to be less than that of the range of the stimuli to which they respond. In other words, their response range is contracted, compared to the stimulus range. The effect may arise because people mentally anchor their judgments on the centre of the range of values and under-adjust away from this anchor when producing their judgments. In our experiments, elevations in sales produced by promotions are linearly related to the size of those promotions. Because of range contraction, regression of the elevations implied by people's forecasts on to the size of planned promotions should reveal a slope that is less than the actual slope characterizing the relation between sales elevation and promotion size. Naturally, we expect that participants realize that a planned promotion with zero investment will have not raise sales at all: the zero point should therefore be fixed. As a result, forecasters who treat the slope of the promotion function slope as too flat (because they are influenced by range contraction) will increasingly underestimate the effects of promotions as those promotions increase in size:

H4a: under-forecasting will be greater for larger promotions

As our statistical forecasts take no account of the difference between promotional and non-promotional periods, we do not expect them to influence range contraction effects.

H4b: the presence of a statistical effect will not influence the effect hypothesized in H4a.

How do people allow for promotions when adjusting statistical forecasts? One possibility is that they first use the past sales figures from periods without promotions to produce a baseline forecast. Then they search back through the presented series for an instance in which the promotion is the closest in size to the one planned for the forecast period. They estimate the sales elevation associated with this past promotion and then add it to their baseline forecast. Goodwin and Fildes (1999) argued that forecasters in their task used this instance-based strategy. However, forecasters may use a rule-base strategy instead. They could inspect the series, together with the various promotions that perturb it, and extract a rule that relates sales elevation to promotion size by performing a mental 'regression'. They would then just need to insert the size of the promotion planned for the forecast period into this rule to obtain their estimate of the appropriate sales elevation to add to their baseline forecast.

Researchers into function learning have carried out experiments to determine the conditions under which people use instance-based strategies and the conditions under which they adopt rule-based ones (McDaniel, Dimperio, Griego, & Busemeyer, 2009). In our task, forecasters do not have to learn the promotion function. What they need to do is better characterized as function extraction. Nevertheless, we can use the strategies adopted by those working on function learning to distinguish between instance-based and rule-based strategies in our forecasting task. A rule-based strategy enables people to make estimates beyond the range of their experience more easily than an instance-based strategy. The latter, being an associative model, tends to fail when applied to a testing range that is outside of the stimulus range. Hence, we compare the absolute error of forecasts for periods within and outside the range of promotions displayed in the presented series. If Goodwin and Fildes (1999) are correct in assuming that an instance-based strategy is used, we expect errors in the latter case to be larger than those of the former one.

H5: forecasters use an instance-based strategy that results in the forecasting error being smaller when the planned promotion is within the range of those presented with the displayed series than when it is outside that range.

3. Experiment 1

We report an experiment based on the paradigm developed by Goodwin and Fildes (1999) and Goodwin et al (2011). Forecasters were presented with time series that were sporadically perturbed upwards by promotions and were required to make forecasts for periods with and without planned promotions. Our primary focus is on identifying the conditions under which performance is better with a statistical forecast than without one.

3.1 Method

3.1.1. Participants A total of 41 prospective students from University College London participated in the study. Their mean age was 18.15 (*SD* = 1.86) and 28 of them were female.

3.1.2 Stimulus materials Forty 50-point sales series were simulated. For each one, a grey line graph represented the sales history of a product of the past 50 weeks. R statistical software was used to generate 40 time series with a mean of 300 and an error level of 21 (7% of the mean). In half of the resulting series, the sales were independent (M = 299.56, SD = 21.53) and, in the other half, sales were sequentially dependent (M = 298.79, SD = 19.48, ρ = 0.7).

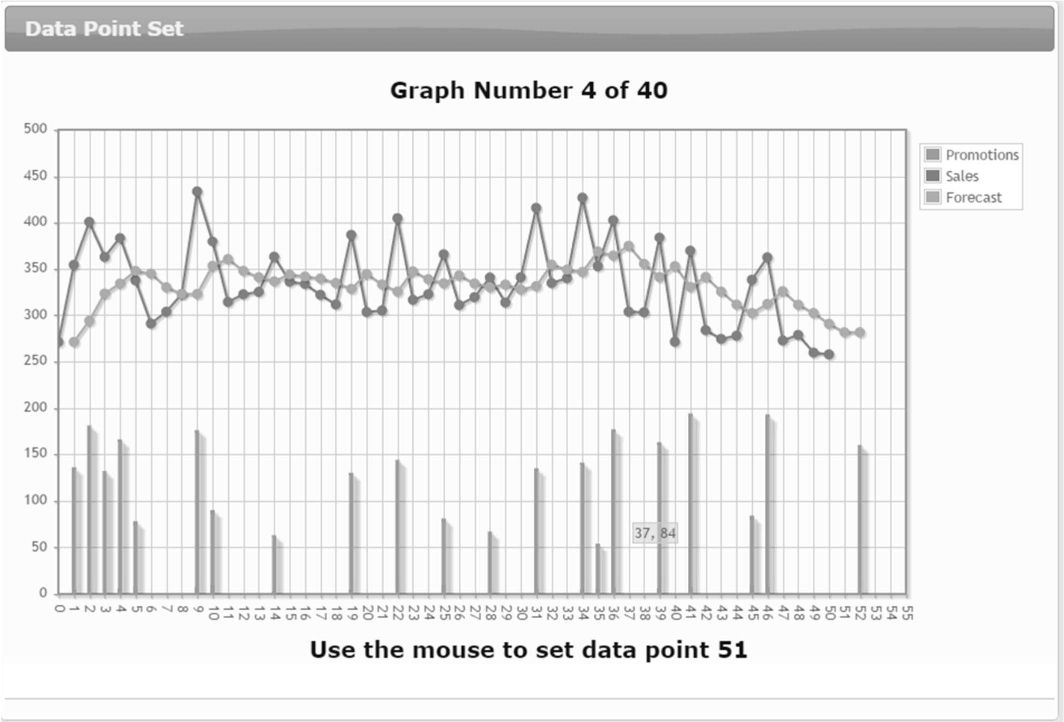
Half of both types of series contained a statistical forecast (an additional line graph) and half did not. The statistical forecast was calculated via the Holt-Winters exponential smoothing method. A line graph represented the statistical forecast history from week 2 to week 52 (Figure 1).

Bars indicated the presence of a promotional expenditure for the product involved. Out of the 50 weeks, 20 weeks were randomly chosen to have a promotion. Promotion size was randomized over these 20 locations with one promotion for every tenth value between 50 and 200. The size of the promotion had a same-week effect on the sales number according to the following formula:

$$PI_t = \frac{Pt}{5} * S_t$$

This indicates a percentage increase (*PI)* on period *t* over the regular sales (*S)* for period *t* that is equal to one fifth of the promotional expenditure (*P)* on period *t*.

For each series, participants were asked to forecast one step ahead and two steps ahead. A promotion was present either on time period 51 or time period 52. The size of this promotion was randomly selected across trials from a range between 30 and 220 (every tenth value) for every participant. Thus, participants had to take account of four promotion sizes (30, 40, 210, 220) that were not within the range of promotions (50 - 200) that were included in the displayed sales series.



Data Point Set

Graph Number 4 of 40

Use the mouse to set data point 51

Figure 1. Example of the screen display for Experiment 1.

3.1.3. Design The study was conducted as a 2 (statistical forecast presence) x 2 (promotion presence) within-subjects experiment. A series of 40 graphs were created, each consisting of 50-point time series.

3.1.4 Procedure Prior to the start of the experiment, participants were given a brief explanation about the different aspects of the graphs that would be presented to them: a line representing the sales history of a product, bars indicating the occasional presence and size of promotions and in half of the cases, an additional line indicating the statistical forecast history. Participants were instructed that this statistical forecast was based on a simple model of past sales data that took no account of whether or not promotions were present. They were told they could choose whether or not to follow the statistical forecast. They were further informed that the goal of the experiment was to forecast as accurately as possible. Series were randomized for each participant. They completed the set of 40 forecasting trials, were thanked, and de-briefed.

3.2 Results

Error measures were calculated relative to the ideal forecast (i.e., the time series signal excluding the noise on non-promotional periods and the signal plus the promotion effect on promotional periods). Mean absolute error (MAE) provided a measure of overall forecast accuracy. Mean error (ME), provided a measure of directional error. It was calculated as the ideal forecast minus participant's forecast, so that negative ME indicates under-forecasting and positive directional error indicates over-forecasting.

Outlier analysis indicated two participants scored more than two standard deviations away from the mean in at least half of the trials and they were therefore excluded.

3.2.1. Mean absolute error Results for MAE are shown in Table 1. Overall, its mean value was 33.93 (*SD* = 8.53). To investigate the effects of the two independent variables (statistical forecast presence, promotion presence), a repeated-measures analysis of variance (ANOVA) was run. This showed a main effect of the presence of a statistical forecast (*F*(1, 38) = 5.29, *p* = .027), with MAE higher for series with a statistical forecast (*M* = 35.08, *SD* = 7.31) than for series without a statistical forecast (*M* = 32.77, *SD* = 10.52). Neither the main effect of the presence of a promotion nor the interaction of this variable with the presence of a statistical forecast approached significance.

Table 1

Descriptive statistics for Experiment 1: MAE and ME for periods with and without promotions and for trials with and without statistical forecasts

Independent Variables		MAE	SD	ME	SD
Raw statistical model	Promotion	51.42	7.59	49.48	8.64
	No promotion	26.80	70	-24.99	0
Aided judgment	Promotion	34.80	10.69	-15.98	18.12
	No promotion	35.37	10.32	20.77	14.84
Unaided judgment	Promotion	32.93	14.20	-14.23	20.60
	No promotion	32.62	10.33	12.36	15.08

Statistical forecasts (*MAE* = 39.11, *SD* = 3.79) were outperformed by both unaided judgment (*t* (38) = 3.68, *p* = .001) and combined judgment (*t* (38) = 3.30, *p* = .002). The statistical model performed better on normal periods (*MAE* = 26.80, *SD* = 0) than on promotional periods (*MAE* = 51.42, *SD* = 7.59): this difference was significant (*t* (38) = -20.26, *p* < .001). These results are not consistent with Hypothesis 1. First, MAE for unaided (J) and combined forecasts (J+S) did not depend on whether a promotion was planned. Second, unaided judgment (J) outperformed both statistical forecasts (S) and combined forecasts (J + S).

3.2.2. Mean error Results for ME are shown in Table 1. A repeated-measures ANOVA revealed a main effect of the presence of a statistical forecast (*F* (1, 38) = 4.96, *p* = .032), with ME for series with a statistical forecast equal to -2.40 (*SD* = 11.76) and for series without a statistical forecast equal to .93 (*SD* = 9.55). There was also a main effect for the presence of the promotion (*F* (1, 38) = 58.52, *p* < .001), with ME for a period with a promotion present equal to -15.10 (*SD* = 17.98) and ME for a normal period equal to 16.56 (*SD* = 14.03). Finally, there was a significant interaction effect between the two variables (*F* (1, 38) = 13.78, *p* = .001). Tests for simple effects of the effect of the presence of a statistical forecast showed that the effect of this variable was significant when there was no promotion was present (*F* (1, 38) = 25.77; *p* < .001) but was not significant when one was present (*F* (1, 38) = .564; *p* = .457).

One-sample t-tests showed that ME for series without promotions (Mean = 16.56; *SD* = 14.03) was significantly above zero (*t* (38) = 7.37, *p* < .001). In contrast, ME for series with promotions (Mean = -15.13; *SD* = 17.98) was significantly below zero (*t* (38) = -5.25, *p* < .001). These results are consistent with Hypothesis 2a: we obtained under-forecasting for periods with promotions but over-forecasting for periods without them. They are not consistent with Hypothesis 2b. This stated that the presence of a statistical forecast would not influence the effect identified in Hypothesis 2a. In fact, it did influence this effect but only when no promotion was present.

According to Hypothesis 3, optimism results in over-forecasting for non-promotional periods being greater than under-forecasting for promotional periods. In fact, tests for simple effects showed that this occurred only when statistical forecasts were present.

3.2.3. Analyses of the implied promotion function To test Hypothesis 4, a regression line was calculated for every participant separately. For each planned promotion value, the mean elevation in the participant's forecast relative to the case when there was no promotion was calculated as a proportion of sales. This elevation was then regressed on to the sales value to obtain the promotion function implied by that person's forecasts (linear, quadratic, and cubic terms were included in these regressions). The mean linear slope of these regression lines in trials without a statistical forecast

(Figure 2a) was .51 (*SD* = .24), significantly different from the slope (.60) of the actual promotion function (*t* (38) = -2.423; *p* = .020). Their mean slope when there was a statistical forecast (Figure 2b) was .48 (*SD* = .19), again significantly different from the slope of the actual promotion function (*t* (38) = -4.38; *p* < .001). Mean values of the slopes, intercepts, Fischer-transformed R² values, and residual errors are shown are shown in Table 2.

Table 2

Mean values of regression parameters for the implied promotion functions when statistical forecasts were present and absent.

Independent Variables	Slope	Intercept	R² (Fischer transformed)	Residual error
Statistical forecast	0.48	.106	0.446	1458.52
No statistical forecast	0.51	-1.77	0.823	668.30

Separate one-way ANOVAs on these data sets indicated that the presence of a statistical forecast reduced the linear fit: significant effects were obtained for *R*² (*F* (1, 76) = 19.04; *p* < .001) and residual error (*F* (1, 76) = 26.83; *p* < .001). However, as Hypothesis 4 states, there was no effect of the presence of a statistical forecast on slope: the range contraction effect was unaffected by the independent variable (*F* (1, 76) = .340; *p* = .561).

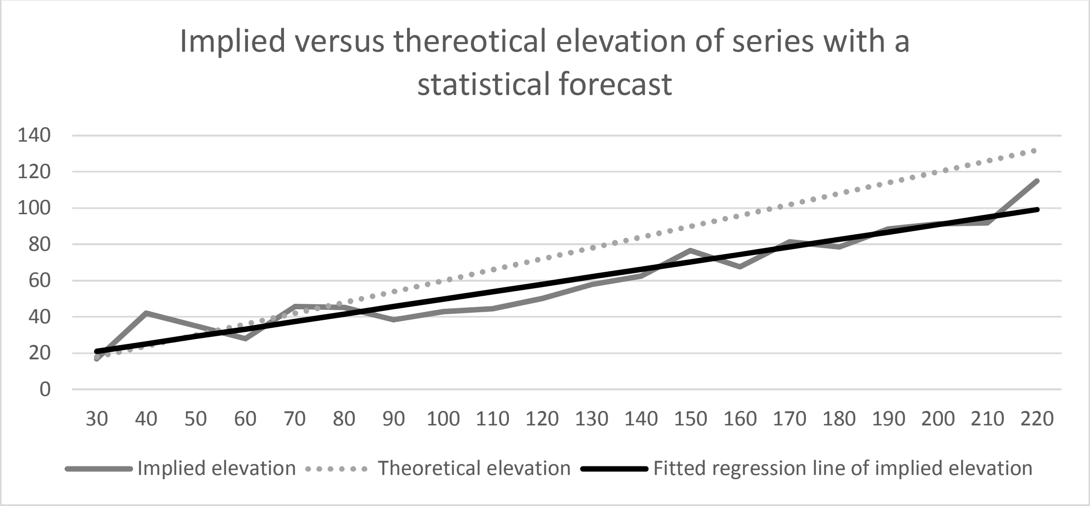


Figure 2a. Implied versus theoretical elevation of the promotional increase in the presence of a statistical forecast

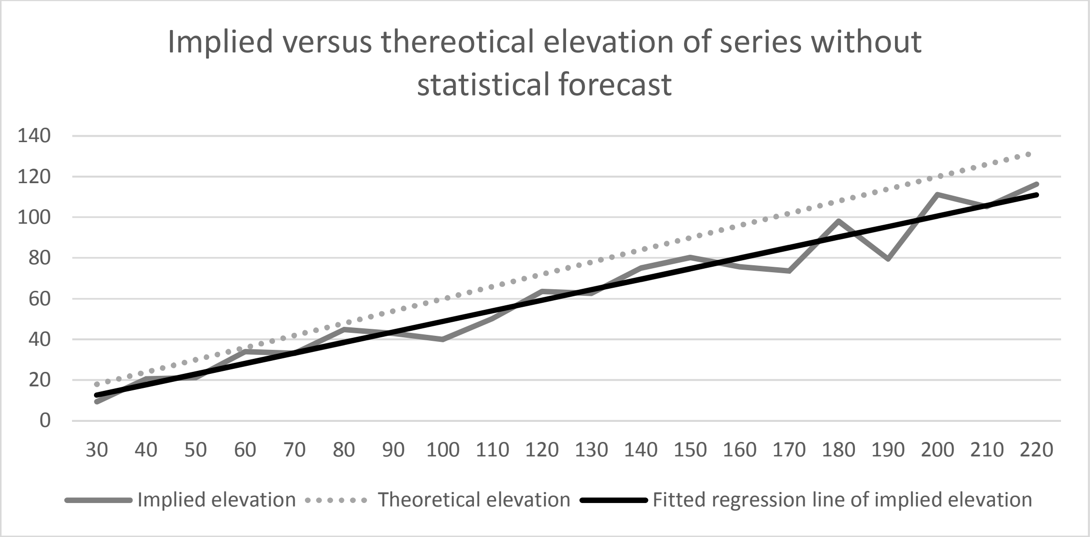


Figure 2b. Implied versus theoretical elevation of the promotional increase without a statistical forecast

More specifically, for trials without a statistical forecast, data for the majority of the participants (31 out of 39) were best fitted by a linear model (Figure 3). Data from the remaining eight participants did not fit any of the tested models (linear, quadratic, cubic). Thus the linear model fitted more often than would be expected by chance (χ^2 (1, 39) = 21.55, *p* < .001). For trials with a statistical forecast, data were fitted best by a linear model in 20 cases, by a quadratic model in five cases, and by no model in 14 cases. Overall, no model provided the best fit significantly more often than the other two models (χ^2 = (1, 39) = 5.08, *p* = .079). These results suggest that the presence of the statistical forecast made it difficult for people to appreciate the linear nature of the promotion function.

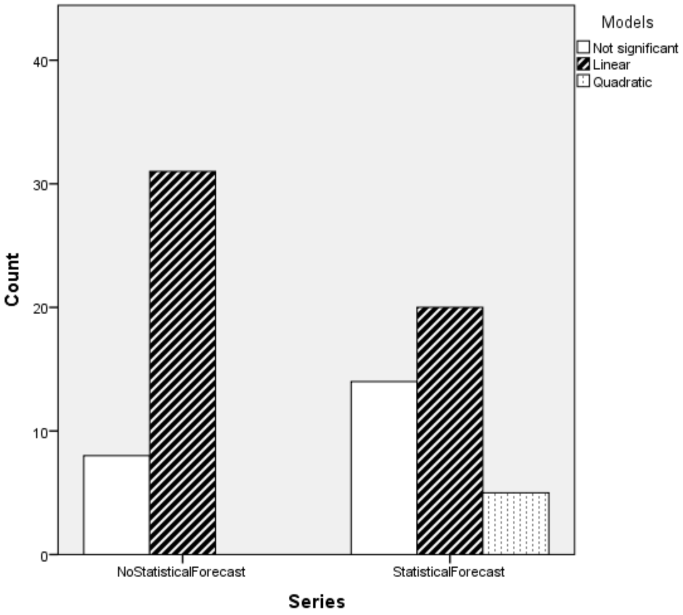


Figure 3. Categorization of models according to presence of a statistical forecast

Promotions affecting past data points in the time series ranged in size between 50 and 200. In contrast, the planned promotions that participants had to take into account when making their forecasts ranged in size between 30 and 220. Forecast errors that are no greater for promotions outside of the example range (30, 40, 210, 220) than for those within it (50-200) would be consistent with participants' use of a rule-based strategy. However, as Hypothesis 5 states, higher forecast errors for promotions outside the example range than within the example range would be more consistent with use of an instance-based strategy.

First, we compared forecast error when planned promotions were inside the example range (*MAE* = 34.59, *SD* = 10.87) with that obtained when planned promotions were 30 and 40 and, hence, beneath the example range (*MAE* = 27.88, *SD* = 15.68). This revealed that error was significantly *higher* in the former case (*t* (38) = -3.71, *p* < .001). Next, we compared forecast error when planned promotions were the inside range with that obtained when planned promotions were 210 and 220 and, hence, beyond the example range (*MAE* = 39.69, *SD* = 19.70). Now error was significantly higher in the latter case (*t* (38) = 3.75, *p* < .001). Figure 4 shows MAE across all promotion sizes. The graph suggests an increase in error with increasing promotion size. A regression analysis confirms a linear relationship between promotional size and MAE *(F* (1, 18) = 19.95, *p* < .001, R^2 = .53). This is consistent with the Weber's Law: error is proportional to stimulus magnitude (Fechner, 1858; Weber, 1934).

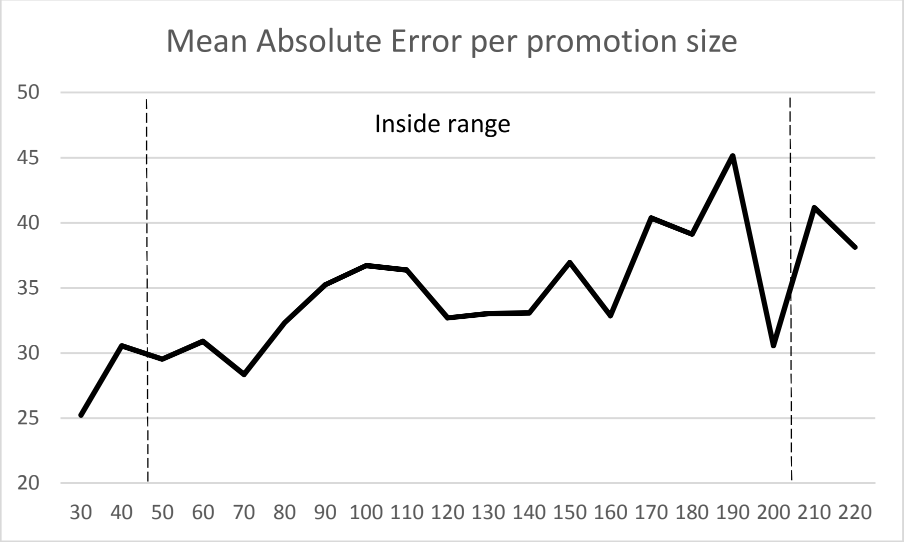


Figure 4. Mean absolute error per promotion size: lower outside range (MAE = 27.88, SD = 15.68), inner range (MAE = 34.59, SD = 10.87), upper outside range (MAE = 39.69, SD = 19.70)

3.3 Discussion

Unaided judgment significantly outperformed both a combination of judgment and statistical forecasting and pure statistical forecasting. The detrimental effect of providing a statistical forecast implied by this result was unexpected and occurred irrespective of whether a promotion was planned or not. In Goodwin and Fildes (1999) and Lim and O'Connor's (1996) studies, providing judges with statistical forecasts did not improve their performance (except in the cases of complex or very noisy series[1]). However, there was no evidence that those forecasts had a detrimental effect. In our study, statistical forecasts impaired accuracy. Why did this occur?

There were strong anchoring effects in our study. These produced under-forecasting on promotional periods and over-forecasting on normal ones. Furthermore, we found that, in normal periods at least, these anchoring effects were stronger when statistical forecasts were provided. This increased level of bias could be responsible for the lower accuracy that we observed when participants had access to such forecasts. It is possible that the additional line on the graph of the data series that displayed the history of the statistical forecasts strengthened the anchoring effect. Lawrence and O'Connor (1992) argued that forecasters anchor on the mean level of the series. Providing the history of the statistical forecast in the form of additional line on the graph may have made this mean more salient and therefore have strengthened the anchoring effect. In the next experiment, we test this possibility.

4. Experiment 2

In this experiment, participants were provided with statistical forecasts for only the periods for which sales had to be forecast. This approach is one that has been adopted in the past by Önkal and her colleagues (e.g., Gönül, Önkal, & Lawrence, 2006; Önkal, Gönul, & Lawrence, 2008). If the line displaying the history of previous statistical forecasts does indeed increase anchoring biases, removing it should eliminate or lessen the detrimental effect of providing statistical forecasts.

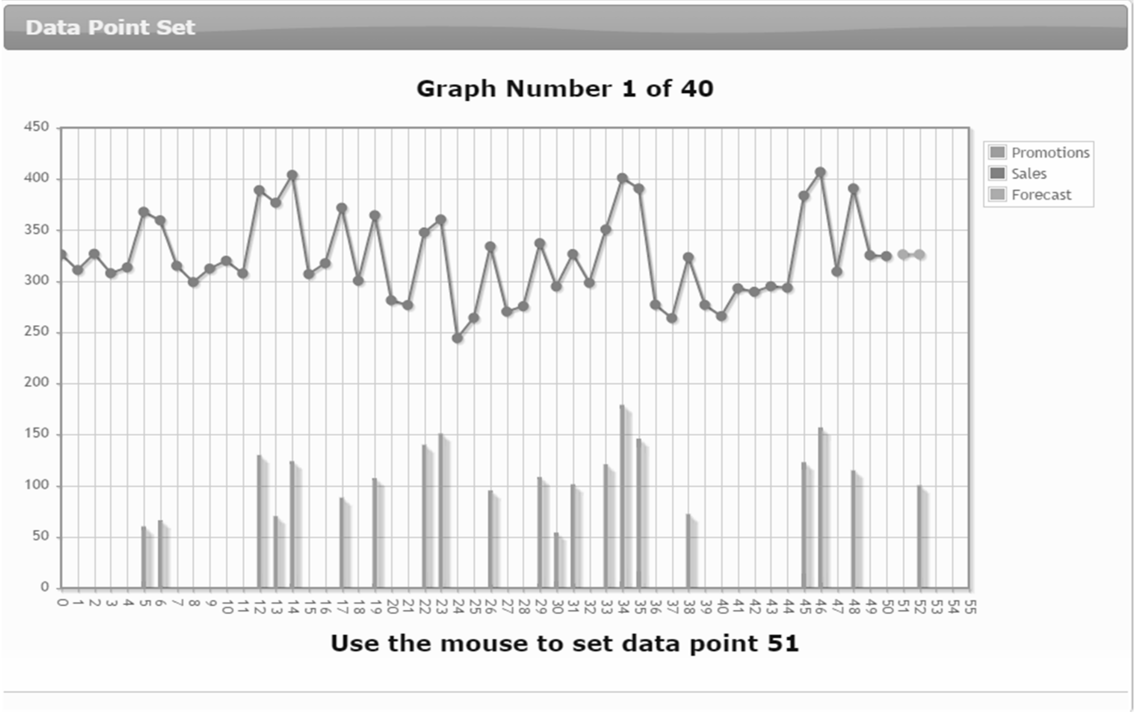
4.1 Method

The experiment was identical to Experiment 1, except that statistical forecasts were displayed only for the period for which forecasts were required.

[1] Though we included autoregressive and independent series to ensure that results generalize over more than one series type, this was not an independent variable in our design. However, post-hoc test analysis showed that series type did not interact with the presence of a statistical forecast in our study (*F* (1, 38) = .59, *p* = .447). This may have been because, compared to the complex series studied by Goodwin and Fildes (1999), both types of series employed here were relatively simple.

4.1.1. Participants Forty-one participants were recruited online as participants in the study. Their mean age was 37.33 years (*SD* = 12.97 years) and 23 of them were female.

4.1.2. Stimulus materials, design and procedure In the instructions and in the experiment, the statistical forecasts were presented as shown in Figure 5. In all other respects, the experiment was identical to the first one.



Data Point Set

Graph Number 1 of 40

Use the mouse to set data point 51

Figure 5. Example of the screen display for Experiment 2.

4.2 Results

Overall level of performance of this experiment (*MAE* = 31.29, *SD* = 9.24) and the previous one (*MAE* = 33.99, SD = 9.30) were not significantly different. MAE scores for periods with and without promotions and for trials with and without statistical forecasts are shown in Table 3. A repeated-measures ANOVA using presence of a statistical forecast and presence of a promotion as within-participant variables revealed only a main effect showing that the presence of a statistical forecast impaired accuracy (*F* (1, 40) = 17.55, *p* < .001).

Purely statistical forecasts had a mean MAE of 40.92 (*SD* = 3.48) and were inferior to both unaided judgmental forecasts (*t* (40) =6.61; *p* < .001) and combined forecasts (*t* (40) = 5.50 ; *p* < .001).

Table 3

Descriptive statistics for Experiment 2: MAE and ME for periods with and without promotions and for trials with and without statistical forecasts

Independent Variables		MAE	SD	ME	SD
Raw statistical forecast	Promotion	55.04	6.96	53.74	7.33
	No Promotion	26.80	0	-26.73	0
Aided judgment	Promotion	32.24	9.33	-14.96	17.27
	No promotion	34.71	11.44	24.73	14.44
Unaided judgment	Promotion	29.25	11.22	-13.25	15.83
	No promotion	28.94	12.56	10.18	13.14

ME scores are shown in Table 3. A repeated-measures ANOVA using presence of a statistical forecast and presence of a promotion as within-participant variables revealed a main effect of the presence of a statistical forecast (*F* (1, 40) = 20.05, *p* < .001), a main effect of presence of a promotion (*F* (1, 40) = 91.67, *p* < .001), and an interaction effect between these two variables (*F* (1, 40) = 91.67, *p* < .001). These findings closely replicate those obtained in Experiment 1. Under-forecasting occurred on promotion periods whereas over-forecasting occurred on normal ones. These effects are consistent with under-adjustment from an anchor (represented by the mean of the series). Presence of a statistical forecast did not affect under-adjustment when there was a promotion but more than doubled the size of the anchoring effect on normal periods.

4.3 Discussion

As Experiment 1, the presence of a statistical forecast significantly impaired performance. An ANOVA using experiment as a between-participants variable showed that there was no significant difference in the size of this effect between the two experiments. Eliminating the forecast history as a source of mental anchoring did not have the expected effect: there was no evidence that the line representing the history of the statistical forecasts increased anchoring in the manner that we hypothesized.

The Holt-Winters exponential smoothing model that was used to produce the statistical forecasts processed the data without distinguishing between normal periods and those affected by promotions[2]. Because of this, the high proportion of periods in the displayed series that were perturbed by promotions would have produced a high level of noise in the data input to the statistical model. The data series may have been too noisy for the simple statistical model to prove useful. By combining the judgmental forecasts with these low quality statistical ones, participants reduced their accuracy below that produced by their unaided judgment.

Because the statistical forecast did not distinguish between promotional and normal periods, its long-term expected value would correspond to a weighted average of the mean sales on promotional periods and on normal periods. Given the high frequency of promotions in the time series (40%), this weighted average would be positioned almost half-way between the mean of the

[2] This was also true of the statistical forecasts used by company that produced the data analysed by Fildes et al (2009) and Trapero et al (2013).

normal periods and the mean of the promotional periods. As a consequence, forecasters would have to adjust upward for a promotion and downward for a normal period. As we have seen, adjustment is difficult and produces errors that can be explained in terms of anchoring.

This line of reasoning implies that the detrimental effect of statistical forecast would lessen if the proportion of promotions in the series were reduced. A less disturbed series would have three advantages. First, it would be visually less complex and should therefore be easier for forecasters to process. Second, with very few promotions, little downward adjustment of the statistical forecast would be needed on normal periods. Third, a time series with fewer perturbations would be treated as less noisy by the statistical model and, as a result, this model would produce more accurate forecasts[3].

5. Experiment 3

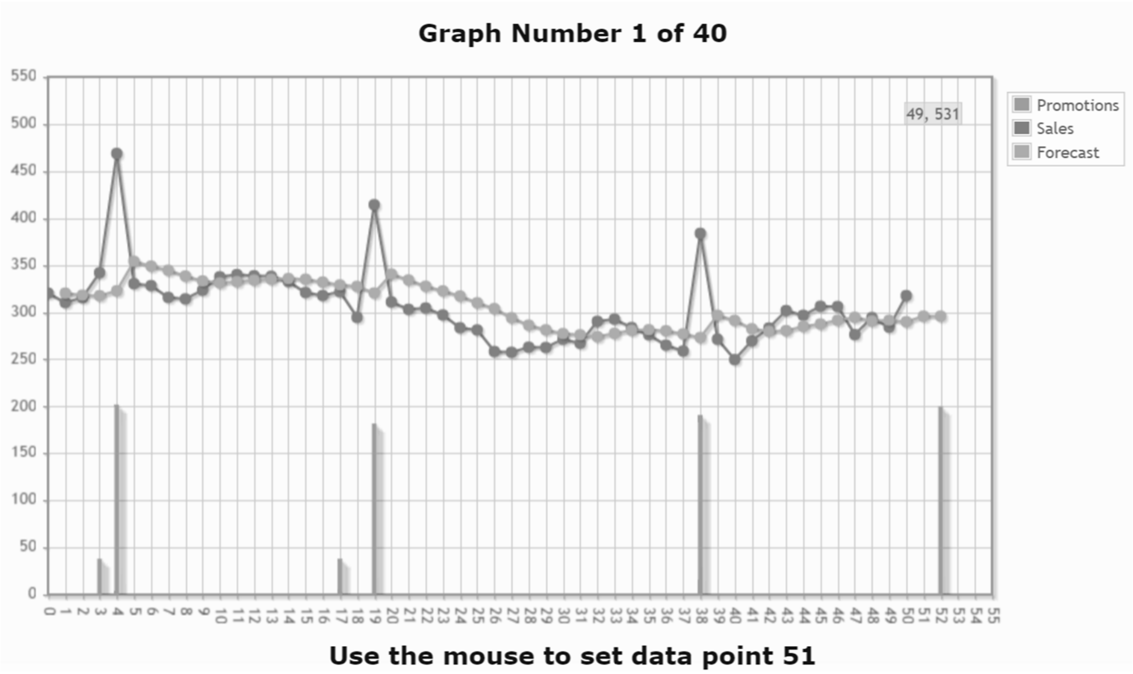
In this third experiment, we hypothesize that the detrimental effect of the statistical forecast will be lower in time series in which only 10% of the periods are subject to perturbations (compared to previous experiments in which time series contained 40% of periods that were subject to perturbations). However, we predict that the mean error of the promotional forecasts will be higher in this new experiment than in Experiment 1. In all other respects, the experiment was the same as Experiment 1.

5.1 Method

5.1.1. Participants Forty participants were recruited online to take part in the study. Their mean age was 20.92 (*SD* = 1.54) and 21 of them were female.

5.1.2. Stimulus materials, design and procedure The method was identical to that of Experiment 1, with the exception that the proportion of promotions was reduced from 40% to 10% of the data points (Figure 6).

[3] On the other hand, fewer perturbations would mean that the promotion function is less well-defined: this could counteract the beneficial effects of fewer promotions.



Graph Number 1 of 40

Use the mouse to set data point 51

Figure 6. Example of the screen display for Experiment 3.

5.2 Results

Outlier analysis revealed that two participants produced MAE scores that were more than two standard deviations away from the mean in more than half of the trials. They were removed from the dataset.

Overall, unaided judgment (*MAE* = 24.70, *SD* = 6.81) performed better than combined judgment (*MAE* = 26.45, *SD* = 7.57). The difference between them was significant: *t* (37) = 2.24, *p* = .031. Pure statistical forecasts (*MAE* = 41.11, *SD* = 4.19) were worse than either unaided judgmental forecasts (*t* (37) =-11.63, *p* < .001) or combined forecasts (*t* (37) = -10.12, *p* < .001).

Data for MAE with and without promotions and with and without statistical forecasts are shown in Table 4. A repeated-measures ANOVA using presence of a statistical forecast and presence of a promotion as within-participant variables revealed a main effect of presence of a statistical forecast (*F* (1, 37) = 4.99, *p* = .032), a main effect of presence of a promotion (*F*(1, 37) = 20.88, *p* < .001), and interaction between these two variables (*F* (1, 37) = 21.17, *p* < .001). Tests for simple effects showed that the statistical forecast had no effect on normal periods (*F* (1, 37) = 1.53; *p* = .224) but impaired performance on promotional periods (*F* (1, 37) = 19.70; *p* < .001).

Data for ME are shown in Table 4. A repeated-measures ANOVA using the same within-participant variables as before revealed a main effect of presence of a promotion (*F* (1, 37) = 7.07, *p* = .012) and for an interaction between this variable and presence of a statistical forecast (*F* (1, 37) = 15.99, *p* < .001). Simple effect analysis shows that provision of a statistical forecast made ME for promotion periods more negative (increased under-forecasting; *F* (1, 37) = 7.75, *p* = .008) but made ME for normal periods more positive (increased over-forecasting; *F* (1, 37) = 9.04, *p* = .005).

Table 4

Descriptive statistics for Experiment3: MAE and ME for periods with and without promotions and for trials with and without statistical forecasts

Independent Variables		MAE	SD	ME	SD
Raw statistical forecast	Promotion	70.26	8.38	71.68	8.50
	No promotion	11.97	0	-6.48	0
Aided judgment	Promotion	30.79	9.50	-4.69	18.05
	No promotion	22.10	8.21	6.74	9.38
Unaided judgment	Promotion	26.13	6.98	-.04	13.37
	No promotion	23.28	8.77	2.92	10.15

5.3 Discussion

Lowering the proportion of perturbations reduced both overall error (MAE) and directional error (Figure 7a and 7b). For all combinations (Statistical forecast yes/no, Promotion yes/no), these errors were significantly lower (at the .01 level) than in the other two experiments. Additionally, the lower proportion of perturbations increased the difference between error on normal periods and error on periods with planned promotions. Whereas the effect of a promotion was not significant in the MAE analysis of Experiment 1 (*F* (1, 38) = .01, *p* = .948), it was in Experiment 3 (*F* (1, 37) = 20.88, *p* < .001).

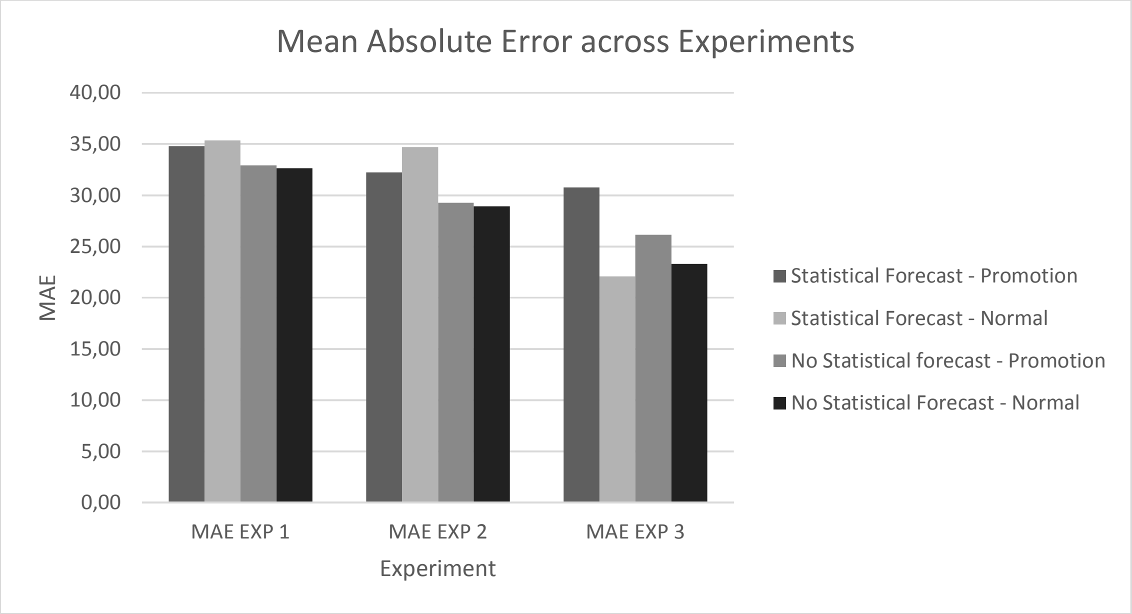


Figure 7a. Mean Absolute Error across Experiments

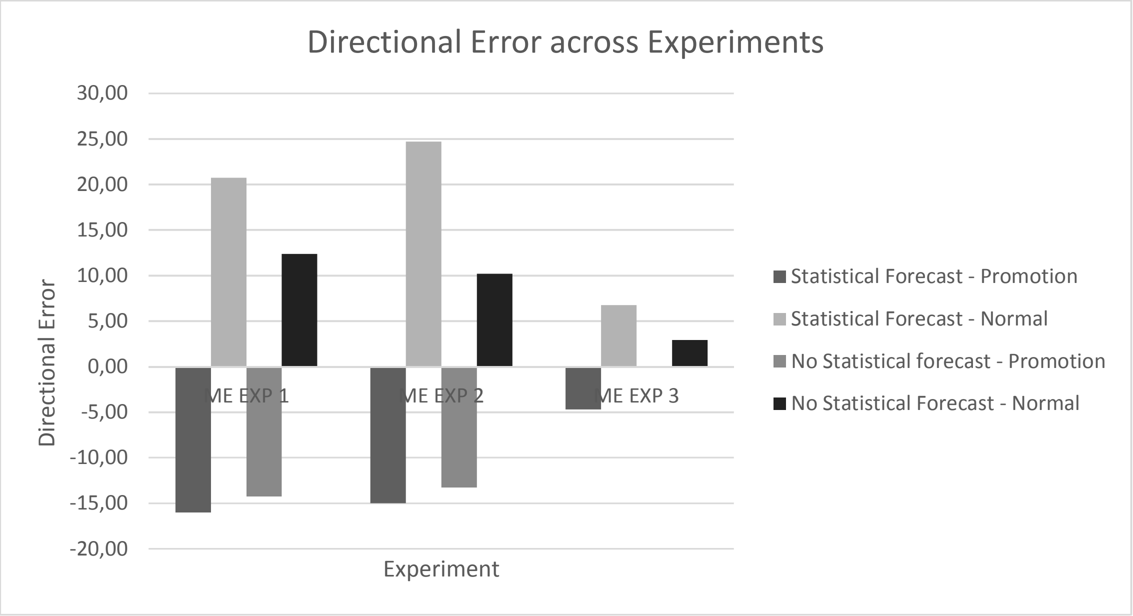


Figure 7b. Directional Error across Experiments

Importantly, in Experiments 1 and 2, the provision of a statistical forecast damaged the accuracy on both normal periods and promotional periods. However, in Experiment 3, the presence of the statistical forecast was no longer detrimental to forecasting accuracy on normal periods. Due to the lower frequency of promotions, the appropriate forecast for normal periods was much closer to the mean of the series and so very little was adjustment necessary. Even with no adjustment away from the mean of the series, error in unaided judgmental forecasts would have been low. Furthermore, statistical forecasts for normal periods were usually close to the mean of the series. (The only exceptions would have been when a promotion perturbed the series on period 50 or 49 and, hence, deflected the statistical forecast produced by the smoothing algorithm upwards.) Hence, in general, the appropriate forecast, the statistical forecast, and an unaided judgmental forecast based on minimal adjustment from the mean of the displayed series would have been closely aligned on normal periods: error in the unaided judgmental forecast would have been low and this error would not have been magnified by taking account of the statistical forecast.

In contrast, when promotions were planned, unaided forecasters would have had to adjust upwards from the mean of the series to take them into account. As we have seen, such adjustments are subject to anchoring and, hence, typically insufficient. The statistical forecasts would again usually be close to the mean of the series. Hence, people who considered them worth taking into account would have inhibited their (insufficient) tendency to adjust upwards and, as a result, their adjustments would have been even more insufficient than they were in the absence of statistical forecasts.

6 General Discussion

The primary aim of this study was to investigate the effect of providing a statistical forecast on judgmental forecasting from time series affected by perturbations (promotions). This investigation was warranted by a number of factors: judgment is pervasive in practice (Fildes & Goodwin, 2007; Goodwin, 2002), formal models can have difficulty dealing with perturbed time series, practitioners resist introduction of these models (Asimakopoulos, 2013; Fildes, et al., 2006) (Asimakopolous & Dix, 2013; Fildes, Goodwin & Lawrence, 2006) and, importantly, practitioners with causal information are often able to make better forecasts than statistical methods or practitioners without that information (e.g., Armstrong, 1983; Edmundson, et al., 1988; Lawrence, et al., 2000). Based on the studies of Lim and O'Connor (1996), Goodwin and Fildes (1999) and Goodwin et al. (2011), we predicted that, for normal periods, judgmental forecasts would be more accurate with the provision of a statistical forecast. For promotional periods, both unaided judgment and combined judgment were expected to outperform the statistical forecast.

The first experiment showed that promotional periods are under-forecast and normal periods are over-forecast, confirming the findings of Webby et al (2005). Such effects imply anchoring on the series mean. These effects occurred regardless of the presence of a statistical forecast. We found no evidence of optimism bias, neither in unaided judgment trials nor in trials with a statistical forecast. Range contraction (Poulton, 1989) affected people's appreciation of the promotion function; in particular, large promotional effects were strongly under-estimated. This effect was not influenced by the presence of a statistical forecast. However, statistical forecasts made it more difficult for forecasters to appreciate the linear relationship that existed between the promotional expenditure and the promotional increase. There was also evidence that MAE was larger with larger promotions, as Weber’s Law (1934) implies it should be.

The most surprising finding was the detrimental effects of statistical forecasts. This effect has not been found in previous studies. These studies found lack of use or insufficient use of such forecasts (e.g., Goodwin & Fildes, 1999; Lim & O'Connor, 1996). In our experiments, statistical forecasts increased the overall error and the directional error for normal periods. We hypothesized that this was due to the line representing the history of the statistical forecast increasing the salience of the mean of the series and so magnifying the anchoring effect. However, results of Experiment 2 did not support this. Experiment 3 showed that reducing the proportion of promotions in the historical series significantly lowered the error. Importantly, whereas the statistical forecast still had a higher overall and directional error for promotional periods, the detrimental effect for normal periods vanished. As very little adjustment away from the series mean was needed for normal periods when there are relatively few promotions in the displayed series, this implies that the statistical forecast somehow interferes with the adjustment process.

In our discussion of Experiment 3, we outlined how this interference is likely to occur. Unaided forecasters anchor on the mean of the series and adjust upwards when a promotion is planned and downwards when none is planned. When the proportion of promotional and normal periods in the displayed series is about equal, the required adjustment in these two cases is also about equal. Because of anchoring effects, unaided forecasters tend to over-forecast for normal periods and under-forecast when promotions are planned. When a statistical forecast is provided (based on the ‘uncleaned’ data series), it is usually close to the mean of the series. Hence, forecasters who consider such forecasts worth taking into account are influenced by them. This influence increases the insufficiency of the already insufficient adjustment away from the mean of the series and, as a result, increases the absolute size of the directional error (Figure 7a).

We suspect that the increase in overall error when a statistical forecast is provided is largely due to this increase in directional error. However, it could also reflect an increase in random error: forecasters may be more inconsistent when a statistical forecast is provided because they are faced with a conflict: the information in the series suggests that adjustments should be made whereas the statistical forecast suggests that they should not.

6.1 Limitations, further research and practical implications

While the findings were broadly robust across experiments and manipulations, replication in other situations is needed. The within-participants design that we used does not generally reflect what practitioners do. For example, someone making sales forecasts for a set of products is unlikely to make use of a statistical forecast for some of them but not others. Thus, it would be useful to investigate whether the effects reported here are replicated using a between-participants design.

Another limitation of this study could be the type of statistical forecast used. Our approach was based on that used by the practitioners in the organizations studied by Trapero et al. (2013) and Fildes et al. (2009). Those companies used forecasts that did not distinguish between promotional and non-promotional periods. However, in the experiment reported by Goodwin and Fildes (1999), statistical forecasts were based on time series that had been cleaned of promotional effects. It could be that the detrimental effects of the statistical forecasts were due to their inadequacy: they provided starting points that were under-forecasts for promotional periods and over-forecasts for normal ones.

It would also be possible to move a step further and provide statistical forecasts for promotional periods that take evidence about the promotion function (obtained from past data) into account. Promotional modelling has been explored in recent years. Trapero et al. (2013) suggest a hybrid model for forecasting promotions that combines judgmental adjustment and transfer function forecasts. Huang, Fildes and Soopramanien (2014) developed a two-stage method, involving variable selection and an Autoregressive Distributed Lag (ADL) model. Most recently, Kourentzes and Petropoulos (in press) suggested an extended multivariate Multiple Aggregation Prediction Algorithm (MAPA) for forecasting. While these models hold much promise, practice is notoriously slow in adapting more advanced formal methods (Lawrence, 2000; Sanders & Manrodt, 2003). Many organizations lack the necessary expertise and resources to employ sophisticated models (Hughes, 2001; Trapero, et al., 2015). Thus, the difficulty here lies not only in the methods themselves but also in getting people to use them (Lawrence, 2000). A possible framework is the Technology Acceptance Model (Davis, 1989; Davis, Bagozzi, & Warschaw, 1989), which suggests that the behavioral intention to use a new system is a combination of perceived usefulness and perceived ease of use. Venkatesh and Davis (2000) propose an extension of this model, and suggest a number of other influencing factors, including experience, subjective norms, image, job relevance, output quality, and the demonstrability of results.

6.2 Conclusion

Previous research has already established that people make insufficient use of statistical forecasts and tend to discount advice from formal models (Goodwin, et al., 2007; Önkal, et al., 2009). Our study provides an even more cautionary tale: the provision of a statistical model holds the potential of producing a worse forecast, compared to unaided judgment, in time series disturbed by special events such as promotions. While we should be careful in formulating practical conclusions before the finding is tested further, one key take-away of this study would be that researchers need to be cautious in viewing the statistical model as being universally better than unaided judgment.

7. References

Armstrong, J. S. (1983). Relative accuracy of judgemental and extrapolative methods in forecasting annual earning. *Journal of Forecasting, 2*, 437 - 447.

Armstrong, J. S., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. In G. Wright & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 269 - 293). New York: John Wiley & Sons.

Asimakopoulos, S. (2013). Forecasting support systems technologies-in-practice: A model of adoption and use for product forecasting. *International Journal of Forecasting, 29*(2), 322.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*, 319 - 339.

Davis, F. D., Bagozzi, R. P., & Warschaw, P. R. (1989). User acceptance of computer technology: a comparison of two theoretical models. *Management Science, 35*, 982 - 1002.

Edmundson, R., Lawrence, M., & O'Connor, M. (1988). The use of non time series information in sales forecasting: a case study. *Journal of Forecasting, 7*, 201 - 211.

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

Fechner, G. T. (1858). Über ein wichtiges psychophysiches Grundgesetz und dessen Beziehung zur Schazung der Sterngrössen. Abk. k. . *Ges. Wissensch. Math.- Phys., K1*(4).

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37*(6), 570-576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 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*(1), 3 - 23.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems, 42*, 1481–1493.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega 30, 30*(2), 127 - 135.

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*(1), 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting, 23*(3), 391 - 404.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*(3), 242 - 253.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: effects of data presentation format on judgemental forecasting. *International Journal of Forecasting, 12*, 119 - 137.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *organizational Behavior & Human Decision Processes, 70*(2), 117-133.

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research, 237*(2), 738 - 748.

Hughes, M. C. (2001). Forecasting practice: organisational issues. *Journal of the Operational Research Society, 52*, 143 - 149.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kourentzes, N., & Petropoulos, F. (in press). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*.

Lawrence, M. (2000). Editorial: What does it take to achieve adoption in sales forecasting? *International Journal of Forecasting, 16*, 147 - 148.

Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior & Human Decision Processes, 43*, 172-187.

Lawrence, M., & O’Connor, M. (1992). Exploring judgemental forecasting. *International Journal of Forecasting, 8*, 15 - 26.

Lawrence, M., O’Connor, M., & Edmundson, R. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research, 122*(2), 151 - 160.

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

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

McDaniel, M. A., Dimperio, E., Griego, J. A., & Busemeyer, J. R. (2009). Predicting Transfer Performance: A Comparison of Competing Function Learning Models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(1), 173-195.

Önkal, D., Gönul, S., & Lawrence, M. (2008). Judgmental adjustments of previously adjusted forecasts. *Decision Sciences, 39*(2), 213 - 238.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*, 407 - 418.

Parducci, A. (1973). A range-frequency approach to sequential effects in category ratings. In S. Kornblum (Ed.), *Attention and performance symposium*. New York: Academic Press.

Poulton, E. C. (1989). *Bias in quantifying judgments*. Hillsdale, US: Taylor & Francis.

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

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantiative forecasting methods in practice. *Omega, 31*, 511 - 522.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Trapero, J. R., Kourentzes, N., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society, 66*(2), 299 - 307.

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*(2), 234 - 243.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124 - 1131.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science, 46*(2), 186 - 204.

Webby, R., O'Connor, M., & Edmundson, R. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting, 21*(3), 411 - 423.

Weber, E. H. (1934). *De pulsu, resorptione, audita et tactu. Annotationes anatomicae et physiologicae*. Leipzig: Koehler.

Zbaracki, M. J. (1998). The rhetoric and reality of total quality management. *Administrative Science Quarterly, 43*, 602 - 636.

Chapter 4

Taking account of the effects of promotions: A comparison of unaided judgmental forecasting and judgmental adjustment of statistical forecasts

Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support

Abstract

How effective are different approaches to providing forecasting support? In a first experiment, forecasters made predictions from time series data (past sales) that were subject to sporadic perturbations (promotions). Some received statistical forecasts that took no account of the fact that some past data points were affected by promotions. Others received statistical forecasts based on data cleansed of the effects of promotions. Forecasts were made for periods with and without planned promotions. Overall accuracy levels in these groups did not differ but both were higher than accuracy in a third group that had no forecasting support. All groups showed under-forecasting on promotional periods but over-forecasting on normal ones. Relative size of these biases depended on the proportion of promotions in the data series. Forecasting support helped not because it reduced them but because it decreased random error (scatter). In a second experiment, forecasters received optimal statistical forecasts that took effects of promotions fully into account. Overall accuracy was higher than in the groups that received statistical support in the first experiment. This was because biases were almost eliminated and because of a further reduction in random error. However, this random error remained high at over 80% of its previous level.

Key words: Forecasting support; judgmental adjustment; time series; promotions; sales

1. Introduction

Business forecasters use both pure (i.e. unaided) judgmental forecasting and forecasting aided by formal statistical forecasts (Sanders & Manrodt, 2003). The latter approach may be increasing as users become more familiar with software that provides forecasting support. As a result, forecast support systems have great potential for improving forecast performance. However, there are factors that prevent this potential being fully realised. Forecasters tend to ignore the 'advice' provided by a formal forecast or take too little account of it (Goodwin, et al., 2007; Lim & O'Connor, 1996; Önkal, et al., 2009). When they do take some account of it, the resulting improvements are generally small, albeit somewhat greater when series are complex and the formal forecasts are of higher quality (Goodwin & Fildes, 1999; Goodwin, et al., 2011; Lim & O'Connor, 1995; Trapero, et al., 2013).

The picture is more complex in the case of series with sporadic perturbations, such as those associated with promotions. Goodwin and Fildes (1999) showed that, in this situation, statistical forecasts tend to help on normal periods but not on those subject to promotions. However, the statistical forecasts used in this research did not take effects of promotions into account: they were based on the baseline time series cleansed of the effects of promotions. Recently, forecasting models that do allow for the effects of promotions have been developed (Huang, et al., 2014; Kourentzes & Petropoulos, in press; Trapero, et al., 2013). However, given that there is considerable lag between development of more sophisticated statistical models and their implementation by practitioners (Lawrence, 2000; Sanders & Manrodt, 2003), it is likely to be some time before they impact business practice.

Even in the case of relatively simple models, there appears to be a gap between the formal forecasts used in experimental studies and those used in business practice. In experimental studies, the formal forecast is based on non-promotional periods only (e.g., Goodwin & Fildes, 1999). In other words, the forecast is calculated from the baseline series cleansed of promotion effects. In non-experimental studies, on the other hand, formal forecasts take no account of whether past periods contain promotions (Fildes, et al., 2009; Trapero, et al., 2013). Hence, if we are interested in the relevance of experimental results to business practice, we need to ask whether any advantage of using judgmentally adjusted statistical forecasts over unaided judgment depends on the type of statistical forecast used.

Goodwin and Fildes (1999) have argued that the benefit of providing statistical forecasts should be greater when they are based on data that have been cleansed of promotional effects. Referring to the estimated level of sales when a promotion does not run as the *baseline* value, they point out that this is because the baseline values provided by that type of statistical forecast can be accepted without any adjustment when no promotions are planned. Moreover, past differences between promotional and non-promotional periods can be directly used as a basis for assessing the size of the adjustment needed when promotions are planned.

In what follows, we address the following questions. First, is there an advantage of using a judgmentally adjusted statistical forecast over using unaided judgment? Second, is any such advantage greater when statistical forecasts are based on past data cleansed of promotional effects? Third, does any benefit derived from provision of statistical forecasts depend on features of the data series (i.e., ratio of promotional to non-promotional periods) or of the periods to be forecast (i.e., whether a promotion is planned)? Finally, can people make good use of ‘ideal’ statistical forecasts that include allowance for the effects of promotions (cf., Huang et al, 2014; Kourentzes and Petropoulos, In Press; Trapero et al, 2013)? In other words, do they adopt these forecasts without making any adjustment?

2. Development of hypotheses

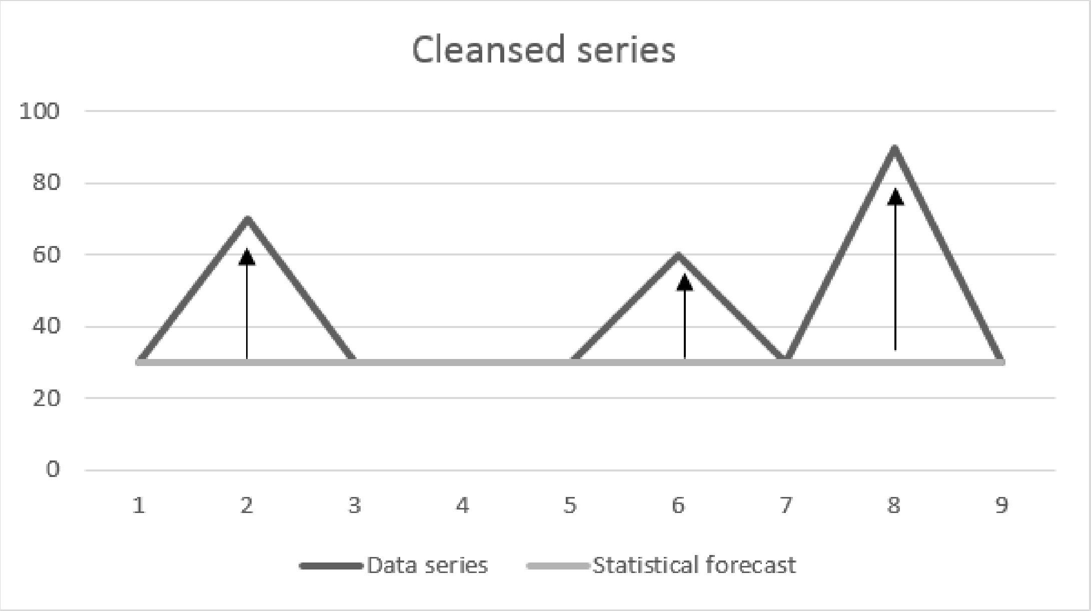
In their survey, Fildes and Goodwin (2007) found that 75% of respondents indicated that they used judgment when making forecasts. Of these, 25% said that they used unaided judgment and 50% said that they used a combination of judgment and statistical forecasting (averaging, judgmental adjustment). Over recent years, use of statistical software has become more pervasive in business settings and so the proportion of forecasters using a combinatorial approach may have increased.

Judgmental adjustment does not always improve statistical forecasts. People tend to make *unnecessary* adjustments even when they have no additional information (Goodwin, 2000; Lawrence, et al., 2006). This may be because they discern patterns in noise (Fildes, et al., 2009), because they are too optimistic and place excess weight on positive signals (Bovi, 2009; Durand, 2003; Kotteman, et al., 1994), or because they want to feel ownership of their forecasts (Önkal & Gönul, 2005). They also tend to be overconfident in the accuracy of their forecasts (Arkes, 2001; Bovi, 2009; Lawrence, et al., 2006), perhaps because a self-serving attribution bias causes them to overestimate the importance of their own judgment relative to that of the statistical forecast (Hilary & Hsu, 2011; Libby & Rennekamp, 2012).

All these studies have focused on whether judgmentally adjusted forecasts are better or worse than raw statistical forecasts. The underlying issue was whether forecasters should be allowed to make adjustments to statistical forecasts and, if they should, whether anything can be done to ensure that their adjustments are beneficial (Goodwin, et al., 2011). In contrast, our primary aim here is to investigate the value of providing a formal forecast to increase forecasting accuracy. Thus, our focus is on whether judgmentally adjusted statistical forecasts are better or worse than unaided judgmental forecasts. For us, the underlying issue is to quantify the benefit of providing forecasters with forecasting support systems. These systems have been assumed to be beneficial (Alvarado-Valencia & Barrero, 2014) because they reduce the processing demands imposed on forecasters (Fildes & Goodwin, 2013). Furthermore, combining forecasts from more than one source outperforms the results of a single forecasting method (Armstrong, 2001a), particularly when the two methods are independent and rely on different information. Given the complementary nature of judgment and statistical methods, their combination should be especially beneficial (Blattberg & Hoch, 1990). Therefore:

H1: Providing forecasters with statistical forecasts improves forecasting accuracy compared to unaided judgment.

Önkal, Sayim, & Lawrence (2012) noted that some differences exist between the characteristics of forecasts examined in experimental research and those prevalent in business practice. As mentioned above, one such difference is in the nature of the statistical forecast provided when series are subject to perturbations of the sort typically produced by promotions: in experimental work, statistical forecasts have been cleansed of promotional effects (Goodwin and Fildes, 1999; Goodwin et al, 2011) whereas, in business data analysed by researchers, they have not (Fildes, et al., 2009; Trapero, et al., 2013). As we mentioned above, Goodwin and Fildes (1999) expected the former approach to produce better results. Specifically, they argued: "This has the benefit of clearly separating the underlying time series from the promotion effects. Moreover, some commercial forecasting packages like *Forecast Pro* now allow observations for special periods to be separated out so that they cannot contaminate forecasts for normal periods. … With access to a statistical time series forecast of the 'baseline value' the judge has only to estimate the effect of the cue and make an appropriate adjustment to the statistical forecast " (p 41).



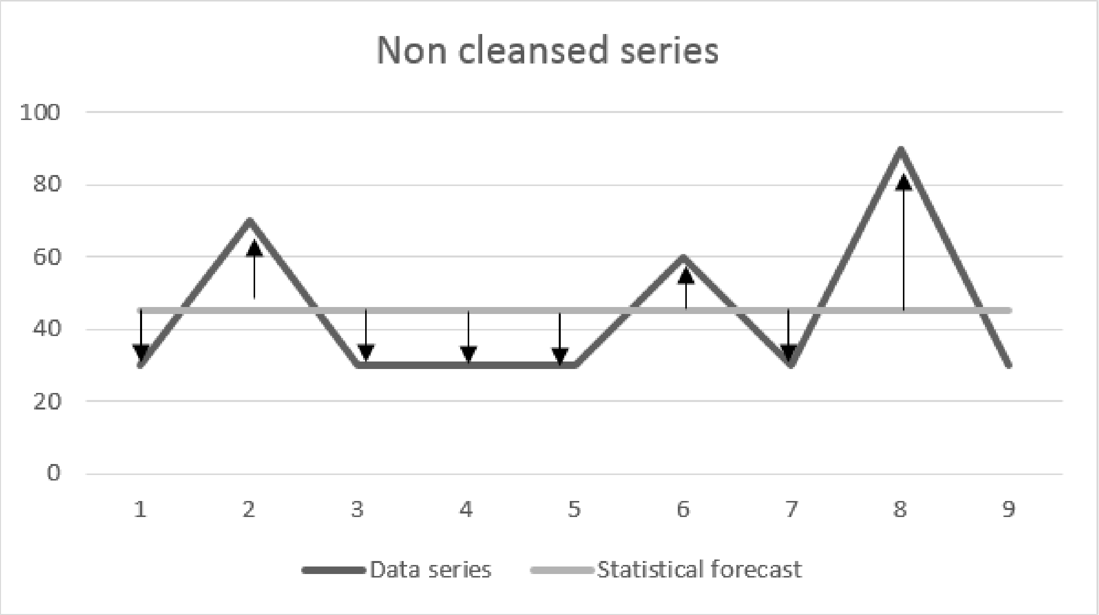


Figure 1. Adjustments necessary for a statistical forecast based on non-cleansed series (upper panel) and cleansed series (lower panel).

As an example, consider a promotion of a given size that has elevated sales by 100 units above the baseline in the past. If a promotion of the same size is planned for the future, 100 units can simply be added to the statistical forecast of the baseline forecast (Figure 1, upper panel). On the other hand, if no promotion is planned, the baseline forecast can be adopted without adjustment. In contrast, statistical forecasts based on non-cleansed data always have to be adjusted. When no promotion is planned, the forecast must be adjusted downwards and, when one is planned, it must be adjusted upwards (Figure 1, lower panel). Forecasters need to know how much the statistical forecast has been influenced by the presence of past promotions in the data series. Without that knowledge, it is difficult for them to know how much to adjust upwards when promotions are planned and how much to adjust downwards when they are not. Thus, the forecasting process is more complex than when the statistical forecast is based on cleansed data series.

In fact, few studies have compared the effects of different types of statistical forecast on the accuracy of judgmental forecasters provided with those forecasts. One exception is Lim and

O'Connor's (1995) experimental study of forecasting from time series without disturbances. They manipulated the accuracy of the statistical forecasts; they varied from low (naïve forecast) to medium (damped) to high (average of damped forecast and the actual value). Participants were asked to make an initial forecast based on their own judgment and were then presented with one of the three types of statistical forecast. After every trial, they were able to see their final forecast, the statistical forecast and the actual value, thus facilitating learning over trials. There was an overall beneficial effect of providing statistical forecasts, consistent with our first hypothesis. Additionally, more accurate statistical forecasts provided greater improvements in accuracy.

Thus, based on Goodwin and Fildes (1999) reasoning and on Lim and O'Connor's (1995) findings:

Hypothesis 2: Formal forecasts based on cleansed series are more beneficial than those based on non-cleansed series.

Judgmental forecasting from time series appears to depend on use of anchoring heuristics (Lawrence and O'Connor, 1992). Given an un-trended data series that includes both normal and promotional periods, unaided forecasters are likely to anchor on the mean of that series. Then they adjust upwards to allow for the presence of a planned promotion in the forecast period and adjust downwards to allow for the absence of a planned promotion. Given that adjustment is typically insufficient when anchoring heuristics are used (Tversky and Kahneman, 1974), we expect under-forecasting for promotional periods but over-forecasting for normal ones. As statistical forecasts based on non-cleansed series follow the mean of the data series, we expect the same mental anchor to be used as for unaided forecasting. Thus, where directional error is given by the outcome minus the forecast:

H3a: For forecasting that is unaided or aided by statistical forecasts based on non-cleansed data series, directional error will be positive for normal periods and negative for promotional ones.

When statistical forecasts are based on cleansed data series, the mean of the statistical forecast history will approximate the mean of the non-promotional periods. Hence, to predict sales for a period when no promotion is planned, forecasters do not need any adjustment. However, for promotional periods, they still need to adjust upwards (Figure 1, lower panel) and this adjustment will be insufficient. Hence:

H3a: For forecasting that is aided by statistical forecasts based on cleansed data series, the directional error will be zero for normal periods and negative for promotional periods.

Statistical forecasts based on non-cleansed series tend to lie between the sales level associated with non-promotional periods and the average sales level associated with promotional periods. When the ratio of promotional to non-promotional periods is low (e.g., 10%), the historical mean of statistical forecasts will be much closer to the actual baseline of the series than when it is high (e.g., 40%). This should benefit forecasting for periods without promotions as minimal adjustment is required. However, when this ratio is low, there is less information on which to estimate the relation between promotional size and its effect. This is likely to impair forecasting for promotional periods. Thus, when statistical forecasts are based on non-cleansed series:

H4: A lower proportion of promotions in the data series will benefit forecasts for non-promotional periods but impair those for promotional periods.

When there are relatively few promotional periods in the data, statistical forecasts based on non-cleansed data series are closer to the baseline and, as a result, they approximate statistical forecasts based on cleansed data. In contrast, when the proportion of promotional periods is high, statistical forecasts based on non-cleansed series are well above the baseline and the difference between them and statistical forecasts based on cleansed-series is larger. Hence:

Hypothesis 5: Any difference in the benefits derived from the two types of statistical forecast will be greater when the proportion of promotional periods in the data series is higher.

3. Experiment 1

A mixed design was used to test these hypotheses. Type of task (unaided judgmental forecasting/forecasting aided by statistical forecasts based on non-cleansed series/forecasting aided by statistical forecasts based on cleansed series) was varied between participants and proportion of promotions in the presented data (40% versus 10%) and forecasting for promotional versus non-promotional periods were varied within participants.

3.1. Method

3.1.1. Participants A total of 153 students from University College London participated in the study. Their mean age was 18.56 years (*SD* = 1.03 years) and 127 of them were female.

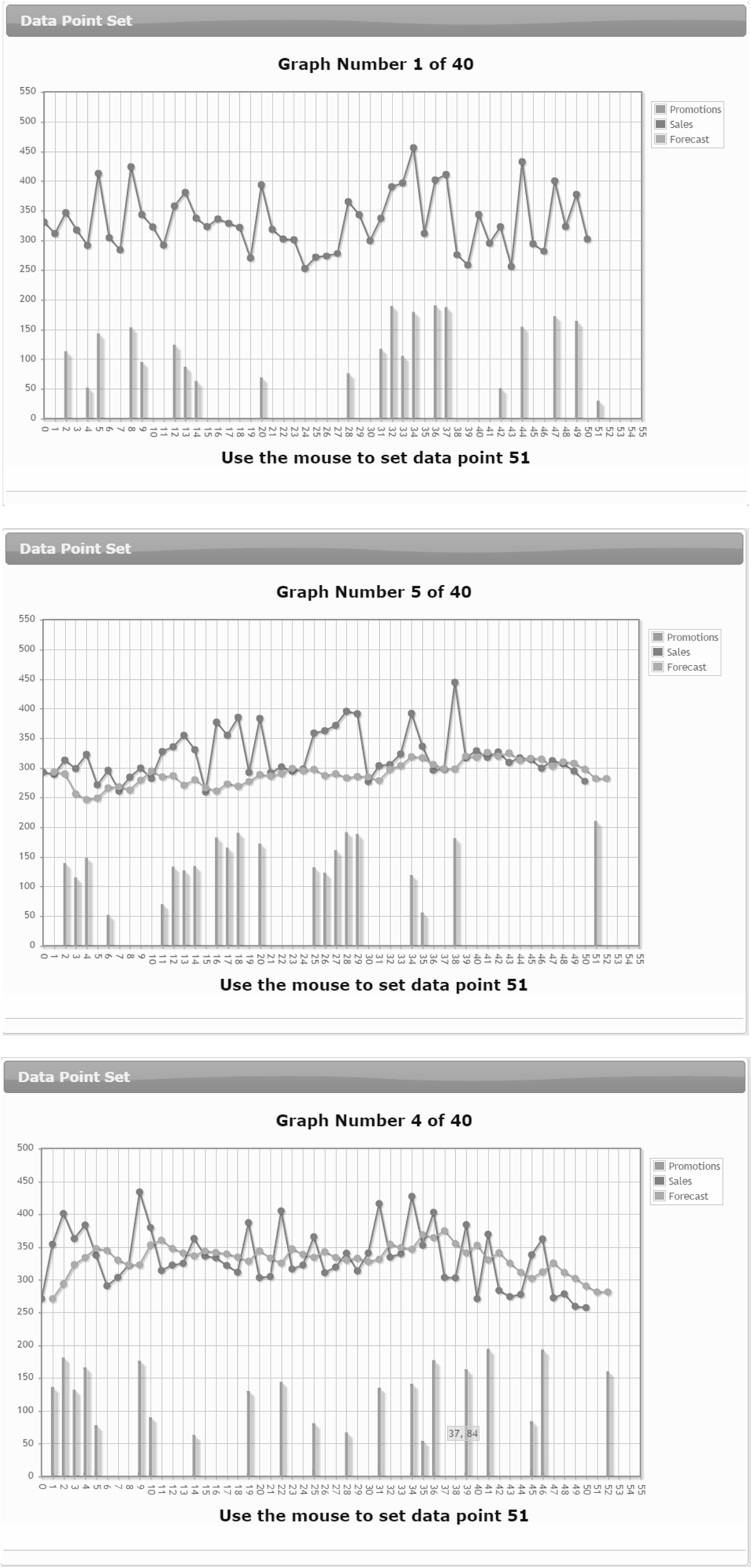
3.1.2 Design and stimulus materials

Forty series, each consisting of 50 data points, were generated with R statistical software. Half of the series were independent (mean = 300, error = 7%) and half were autoregressive (mean = 300, ρ = 0.7, error = 7%). Series were displayed as a grey line and were labelled 'sales'. The graphs also contained vertical blue bars that indicated promotional expenditure on either five or 20 of the 50 periods. Both location and size of these promotions were assigned randomly. Promotion size was selected at random without replacement from a list of every tenth value between 50 and 200. The size of the promotion had a same-week effect on the sales number according to the following formula:

$$PI_t = \frac{Pt}{5} * S_t$$

This indicates a same-week percentage increase *PI* at time *t* (over the regular sales *S* at that time) equal to one fifth of the promotional expenditure *P*.

Participants were asked to forecast one step ahead and two steps ahead. A promotion was present either on time period 51 or time period 52. The size of this promotion was randomized for every participant across trials. Over the experimental session, it included every tenth value between 30 and 220. Thus, participants were required to forecast four promotion sizes (30, 40, 210, 220) that were not included in the range presented in the data series (i.e., 50–200).



Data Point Set

Graph Number 1 of 40

Use the mouse to set data point 51

Data Point Set

Graph Number 5 of 40

Use the mouse to set data point 51

Data Point Set

Graph Number 4 of 40

Use the mouse to set data point 51

Figure 2. Experiment 1: Example of the screen display. Unaided judgment (upper panel), aided judgment based on cleansed series (middle panel) and based on non-cleansed series (lower panel)

The presence and type of statistical forecasts was manipulated between participants. The first group (A) did not receive a statistical forecast (Figure 2; Upper panel). Two other groups received a statistical forecast calculated using the Holt-Winters exponential smoothing method. A line graph represented the statistical forecast history for week 2 to week 52. One of these groups (B) received a statistical forecast based on the total sales: in calculating it, no distinction was made between normal and promotional periods (Figure 2; Middle panel). The other group (C) received a statistical forecast based on the baseline data series, cleansed of promotional effects (Figure 2; Lower panel).

3.1.3. Procedure Participants of group A (unaided judgment, no statistical forecast) were given the following text on an instruction sheet: "*Please read this document carefully before you start with the first graph! In this experiment, you will receive a number of graphs such as the one depicted below. On the X-axis, you will find the time period, ranging from 0 to 55. On the Y-axis, you will find the sales number, ranging from 0 to 500. The grey line indicates the sales data of a product in the past 50 time periods. The blue bars indicate the promotional investment (e.g., an advertisement campaign) made for that product. The number of promotions can vary: some graphs will have 5 promotions, others will have 20. It is your job to predict the sales number of the following two time periods (51 and 52), as accurately as possible. Pay attention, because sometimes there is a promotion present and sometimes there isn't. You can make your prediction by clicking with your mouse on the graph. An information box with your mouse's location appears next to your cursor. First click on your prediction for time period 51 and only then for time period 52. Afterwards, a box 'next graph' will appear on the bottom of the page."*

Participants in group B (statistical forecast based on cleansed series) saw the following additional text: *"The orange line is a forecast from a statistical model. The model is based on the cleaned sales data: the promotion effects have been taken out of the data until only the baseline remained. The model uses these baseline data to produce the statistical forecasts. You can see the* *predictions it made in the past and what it predicts for time period 51 and 52. You can choose whether or not to follow the statistical forecast"*.

For those in group C (statistical forecast based on non-cleansed data), the additional text was as follows: "*The orange line is a forecast from a statistical model. We have fed the sales data to a statistical model. You can see the predictions it made in the past and what it predicts for time period 51 and 52. This statistical model is a simple model that ignores whether or not a promotion took place: it is just based on the value of the sales figures. You can choose whether or not to follow the statistical forecast.*"

In addition, participants were orally instructed to pay close attention to the explanation of the graphical components on their instruction sheet and were given a short demonstration of two example trials.

3.2. Results

We present analyses of three error scores: mean absolute error (MAE), mean error (ME), and variable error (VE).

3.2.1. Mean absolute error. MAE was used to measure overall error level. Errors were calculated relative to the ideal forecast. This was provided by the *signal* (excluding the noise) of the time series on non-promotional periods and by the signal plus the promotion effect on promotional periods. For the independent time series, the ideal forecast for a non-promotional period was 300 (the mean). For the autoregressive series, it was 0.2 (1-ρ) of the distance from the last data point towards the mean. Outlier analysis indicated two participants had scored more than two standard deviations away from the mean on half of the trials or more; they were therefore excluded from the analyses.

Table 1a shows MAE values for each combination of the three independent variables. The overall mean value of MAE was 30.20 (*SD* = 9.65). Table 1b shows the scores of the raw statistical forecasts for the different trial types.

Table 1a

Condition Means (MAE) of Promotion Frequency and Presence of Promotion

Independent Variables		Statistical forecast			
		None	Cleansed	Not cleansed	Means
40% promotions	Promotion	33.48	30.13	30.24	31.28
	No promotion	35.06	28.66	31.67	31.79
10% promotions	Promotion	34.54	31.11	29.43	31.69
	No promotion	29.94	24.44	23.73	26.03
Means		33.26	28.59	28.77	30.20

Table 1b

Error Scores of the Raw Statistical Forecasts according to Promotion Frequency and Presence of Promotion

Independent Variables		Cleansed		Not cleansed	
		MAE	SD	MAE	SD
40% promotions	Promotion	76.04	7.51	52.27	7.51
	No promotion	3.90	0	2.21	0
10% promotions	Promotion	77.14	7.55	68.05	7.98
	No promotion	7.71	0	7.69	0

An analysis of variance with statistical forecast as a between-participants variable and promotion frequency and promotion presence as within-participant variables revealed a main effect of statistical forecast (*F* (2,150) = 3.99, *p* = .021, η_p^2 = .050). Hypothesis 1 stated that the provision of a statistical forecast would be beneficial to forecasting accuracy, such that unaided judgment would result in higher error than unaided judgment. One-tailed t-tests confirm that the MAE for the unaided judgment group (*MAE* = 33.25, *SD* = 9.84) was significantly higher from that of the cleansed forecast group (*MAE* = 28.58, *SD* = 9.29; *t* (100) = 2.47, *p* = .008) and significantly higher than that of the non-cleansed forecast (*MAE* = 28.77, *SD* = 9.26; *t* (100) = 2.37, *p* = .010).

The MAE scores of the cleansed forecast group and the non-cleansed forecast group were not significantly different from one another (*t* (100) = -.10, *p* = .921). Thus we failed to obtain support for Hypothesis 2, which stated that participants given a statistical forecast based on cleansed series would be more accurate than those given a statistical forecast based on non-cleansed series.

There was a main effect of frequency of promotions in the data series (*F* (2,150) = 28.21, *p* < .001, η_p^2 = .158), a main effect of the presence of a promotion in the period to be forecast (*F* (2,150) = 7.35, *p* = .008, η_p^2 = .047), and an interaction between these two variables (*F* (2,150) = 743.82, *p* < .001, η_p^2 = .226). Analysis of simple effects showed that these effects arose because lower error with less frequent promotions occurred when forecasts were made for non-promotional periods (*F* (1, 150) = 61.66, *p* < .001) but not when they were made for promotional ones.

We failed to obtain support for Hypothesis 5: there was no significant interaction between the type of statistical forecast provided and frequency of promotions in the data series.

3.2.2. Mean error MAE is a measure of overall error. Following Thurstone (1926), we can regard overall error as being made up of directional error or bias (ME) and scatter or variable error (VE). Taking D as the Actual – Forecast, ME is defined as ΣD/n and VE as $\sqrt{}$ ([Σ (D – ME)^2]/n). Thus, overall error could theoretically comprise a) bias but no scatter (all forecasts are a fixed distance from the optimal forecast with no distribution around that point), b) scatter but no bias (forecasts are distributed around a central point but that central point is the optimal forecast), or c) bias and scatter (forecasts are distributed around a central point that is a fixed distance from the optimal forecast). In practice, both bias and scatter contribute to overall error but their relative contributions depend on contextual factors.

To investigate the reasons for the differences in MAE reported above and to test hypothesis 3 – 5, we report analyses of ME (Table 2) in this section and of VE (Table 3) in the following one.

Table 2

Condition Means (ME) of Promotion Frequency and Presence of Promotion

Independent Variables		Statistical forecast			
		None	Cleansed	Not cleansed	Mean
40% promotions	Promotion	-8.07	-6.47	-3.85	-6.13
	No promotion	18.21	14.08	18.38	16.89
10% promotions	Promotion	-10.93	-11.39	-7.04	-9.79
	No promotion	12.63	8.07	9.84	10.18
Means		2.96	1.07	4.33	2.79

Table 3

Means of Promotion Frequency and Presence of Promotion for the three Conditions for Variable Error (VE)

Independent Variables		Statistical forecast			
		None	Cleansed	Not cleansed	Mean
40% promotions	Promotion	7.72	7.15	7.05	7.31
	No promotion	7.02	6.63	6.78	6.81
10% promotions	Promotion	7.60	6.86	6.73	7.06
	No promotion	7.02	6.27	5.62	6.30
Means		7.34	6.73	6.55	6.87

There was a main effect of whether the forecast was for a period with or without promotions (*F* (2,150) = 126.69, *p* < .001, η_p^2 = .458). Mean Error was negative when forecasts were for periods on which promotions were planned but positive when they were for periods with no promotions planned. There was also a main effect of the proportion of promotions in the data series (*F* (2,150) = 67.17, *p* < .001, η_p^2 = .309): overall, ME was lower when there was a low proportion of promotions in the data series than when there was a high one. There was also a significant interaction between these two variables (*F* (2,150) = 6.49, *p* = .012, η_p^2 = .041). Analysis of simple effects showed that this arose because a lower proportion of promotions in the data series decreased the positive ME of forecasts for non-promotional periods (*F* (1, 150) = 63.77, *p* < .001) but increased the negative ME of forecasts for promotional periods (*F* (1, 150) = 16.54, *p* < .001). This result is consistent with Hypothesis 4 that stated that fewer promotions would benefit forecasts for non-promotional periods but impair those for promotional ones.

Hypothesis 3a predicted that, when the statistical forecast was based on *non-cleansed* series, under-forecasting for promotional periods would occur, and over-forecasting for normal periods. One-sample t-tests confirmed that ME was significantly below zero on promotional periods (*t* (50) = -2.27, *p* = .028) and significantly above zero (*t* (50) = 8.20, *p* < .001) on normal ones.

Hypothesis 3b predicted that, for the forecasts based on *cleansed* series, the ME for normal periods would be zero and the ME for promotional periods would be negative (i.e., under-forecasting). While the ME for promotional periods in the cleansed series condition was indeed significantly below zero (*t* (50) = -3.51, *p* = .001), the ME for normal periods was positive and significantly different from zero (*t* (50) = 7.37, *p* < .001). This over-forecasting on normal periods was greater when there were 40% promotions in the data series than when there were 10% promotion in the data series (*t* (50) = 4.25, *p* < .001).

3.2.3. Variable Error. There was a significant effect of group on VE (*F* (2,150) = 3.10, *p* = .048). We hypothesized that the error of the unaided judgment group would be higher than that of the aided judgment groups. One tailed t-tests confirm that the VE of unaided judgment group was indeed larger than that of the group who received cleansed forecasts (*t* (100) = 1.98, *p* = .025), and that of the group that received cleansed forecasts (*t* (100) = 2.34, *p* = .011). (VE scores in the latter two groups were not significantly different from one another.)

Forecasts from data series with 40% promotions had higher VE scores than those from data series with 10% promotions (*F* (1,150) = 7.38, *p* = .007, η_p^2 = .047). In addition, forecasts for promotional periods had higher VE than those for non-promotional ones (*F* (1,150) = 17.98, *p* < .001, η_p^2 = .107).

3.3. Discussion

The experiment produced two separate groups of effects. The first concerns the effects on MAE and VE of whether participants made unaided forecasts, made forecasts after being given non-cleansed statistical forecasts, or made forecasts after being given cleansed statistical forecasts. The second concerns effects on MAE, ME, and VE of the proportion of promotional periods in the data series and of whether forecasts were made for promotional or normal periods. As there were no interactions between these two groups of effects, we will discuss them separately. Once we have done so, we will summarise a unitary account of the cognitive processes underlying performance that explains both types of effect.

3.3.1. Effects of providing forecast support

Provision of statistical forecasts reduced overall error (MAE). However, further analysis showed that this was not because they reduced the directional error or bias (ME) in forecasts. Instead, it was because they reduced random error or scatter (VE): they made forecasts more consistent.

We anticipated that the cleansed statistical forecasts would improve forecasting more than the non-cleansed ones. However, our rationale for this was based on our expectation that the cleansed forecasts would lower bias by reducing the under-adjustment from the mean of the series – the salient anchor in the unaided and non-cleansed statistical forecast conditions. It was on this basis that we generated hypotheses 2, 3a, and 5. However, no differences in the effectiveness of the cleansed and non-cleansed statistical forecasts were evident, either as main effects or as interactions in our analyses of MAE, ME and VE. They did not affect degree of under-adjustment from the mean of the series.

Provision of cleansed and of non-cleansed statistical hypotheses both improved overall accuracy but there was no difference in the degree to which they did so. This was because there was no difference in the extent to which they reduced VE.

3.3.2. Effects of promotions in the data series and in the periods to be forecast

Proportion of promotions in the data series and whether the forecast was for a normal or for a promotional period interacted in their effects on overall forecast accuracy (Table 1): a greater proportion of non-promotional periods in the data specifically helped forecasts for non-promotional periods. To understand why this was, we need to consider the separate analyses of ME and VE.

Forecasters are likely to anchor on the overall mean of the data series (Lawrence and O'Connor, 1992). Fewer promotions meant that that overall mean was closer to the mean value of the non-promotional periods but further from the mean value of the promotional periods. So, with fewer promotions in the data series, a larger adjustment from the overall mean of the series was needed to forecast promotional periods but a smaller adjustment was needed to forecast non-promotional periods. The data show that under-adjustment was greater when a larger adjustment was needed. This is to be expected. In psychophysics, the Weber-Fechner Law (Baird and Noma, 1978; Fechner, 1860; Weber, 1834) summarizes many findings showing that errors in discrimination are proportional to the overall size of the stimulus being judged. Hence, because under-adjustment was proportional to the size of the required adjustment, ME became less positive on normal periods but more negative on promotional ones as the proportion of promotions in the data series decreased (Table 2).

A greater proportion of promotions in the data series increased its variability. If people used their estimate of the overall mean of the series as a judgment anchor, this estimate would have been more variable when the proportion of promotions in the data series was higher. As a result, VE was also higher (Table 3).

To allow for the absence or presence of a promotion in the period to be forecast, people would have had to adjust away from this initial judgment anchor. When there was no promotion planned, this would require forecasters merely to estimate from the data series the mean value of sales when no promotion had occurred (and to move their judgment away from the initial anchor towards that mean value). However, when a promotion was planned, they would have to do more than just estimate the mean value of sales when a promotion occurred: they would also have to take into account the relation between the size of a promotion and the elevating effect it had on sales. This could be done in various ways (e.g., via some kind of mental regression). However, it is reasonable to assume that this additional process would be imperfect and so add some random error to the forecasts. As a result, VE was higher in forecasts for promotional periods (Table 3).

The reasons for the relatively low value of MAE when forecasts for normal (rather than promotional) periods were made from series with 10% (rather than 40%) promotions are now clear. VE is reduced by forecasting for normal rather than for promotional periods. Additionally, VE is reduced with fewer promotions in the data series. Finally, fewer promotions in the data also result in a reduction of the size of the positive ME associated with forecasts that are made for normal periods. This combination of two factors reducing VE (normal periods, fewer promotions) and one factor reducing ME (fewer promotions) results in a particularly low MAE value. MAE is higher in all other cases because factors that lower VE and those that lower ME do not combine in the same felicitous manner. For example, consider the case in which forecasts are made for a promotional period from data series containing 40% promotions. Here, the higher proportion of promotions in the data series reduces the size of the negative ME associated with making forecasts for promotional periods. However the beneficial effects of this are counteracted by the fact that VE is higher when forecasts are made for promotional periods and when the proportion of promotions in the data series is higher.

Why did the presence of a statistical forecast lower VE and, hence, MAE? We have argued that forecasters first estimate the overall mean of the data series and that this acts as an initial judgment anchor. Furthermore, this is an error-prone process: VE is higher when the data series is more variable. Both types of statistical forecast act to make it less error-prone. Forecasters could reduce the amount of random error in their estimate of the series mean simply by averaging it with the non-cleansed statistical forecast or by averaging it with the cleansed statistical forecast plus some increment specific to the proportion of promotions in the data series.

3.3.3. Summary

In summary, we can explain all the patterns in the data by assuming that people produce their forecast in two steps. First, they estimate the overall mean of the data series in order to use it as an initial judgment anchor. The size of the random error associated with this estimate is higher when data series are more variable but it can be reduced by provision of a statistical forecast. Second, forecasters adjust away from this initial anchor to allow for whether a promotion is planned or not. Under-adjustment results in under-forecasting on promotional periods and over-forecasting on normal ones. The size of the under-adjustment is greater when a larger adjustment is required: hence, a greater proportion of promotions in the data series results in greater (positive) ME on normal periods but smaller (negative) ME on promotional periods. Adjustments are based on just the mean value of sales on non-promotional periods when normal periods are forecast but they must take into account the relation between size of promotions and the size of their effects on promotional periods. This additional process is error-prone and hence results in higher VE on promotional periods.

4. Experiment 2

Unexpectedly, Experiment 1 failed to reveal any difference in forecast accuracy between participants who received the cleansed statistical forecasts and those who received the non-cleansed ones. Non-cleansed statistical forecasts are cruder: they require less processing of the data series but always require some adjustment. In contrast, cleansed forecasts provide a clearly defined baseline series and, as a result, they can be accepted without adjustment for non-promotional periods. Despite this, participants made large upward adjustments on these periods (Table 2).

It is possible that people who see that the cleansed statistical forecast does not account for promotions falsely infer that cannot be 'trusted' for normal periods either. As a result, they make adjustments for both types of period. Forecasts need to be relevant, justifiable and valuable in dealing with future uncertainties in order for them to be acceptable (Gönül, et al., 2006). The clear unacceptability of the cleansed statistical forecasts on promotional periods may have been inappropriately generalized to affect the acceptability of those forecasts for both types of period (promotional and normal).

This possibility prompted us to carry out Experiment 2. We provided participants with 'optimal' forecasts. Each forecast for a promotional period was based on the cleansed statistical forecasts but with the appropriate increase in sales produced by the promotion in the promotional period added to it. While it is not completely realistic to obtain such forecasts in business practice, it is an approach that is now approximated by recently developed forecasting methods that include promotional modelling (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, in press; Trapero, et al., 2013).

We suggested above that cleansed forecasts for non-promotional periods are not accepted without adjustment because it is clear to forecasters that cleansed forecasts for promotional periods are unacceptable without adjustment and this leads to a lack of trust in all forecasts. As a result, all forecasts are adjusted. In the present experiment, it was made clear to forecasters that forecasts for promotional as well as for normal periods were acceptable without adjustment. If our suggestion is correct, then forecasters would have no reason not to trust the statistical forecasts. As a result, they should be judged acceptable and adopted without adjustment.

4.1. Method

The experiment was identical to Experiment 1, except that statistical forecasts for promotional periods were elevated by an amount that was appropriate to the size of the planned promotion.

4.1.1. Participants Fifty students from University College participated in the study. Their mean age was 17.77 years (*SD* = 0.87 years) and 40 of them were female.

4.1.2. Stimulus materials, design and procedure In the instructions and in the experiment, the statistical forecasts were presented as shown in Figure 3. In all other respects, the experiment was identical to the first one. The instructions with regard to the statistical forecast were adapted as follows: "*The orange line is a forecast from a statistical model. The model is based on the cleaned sales data: the promotion effects have been taken out of the data until only the baseline remained. The model uses these baseline data to produce the statistical forecasts and then adds the promotion effects on top of this forecast. You can see the predictions it made in the past and what it predicts for time period 51 and 52. You can choose whether or not to follow the statistical forecast.*"

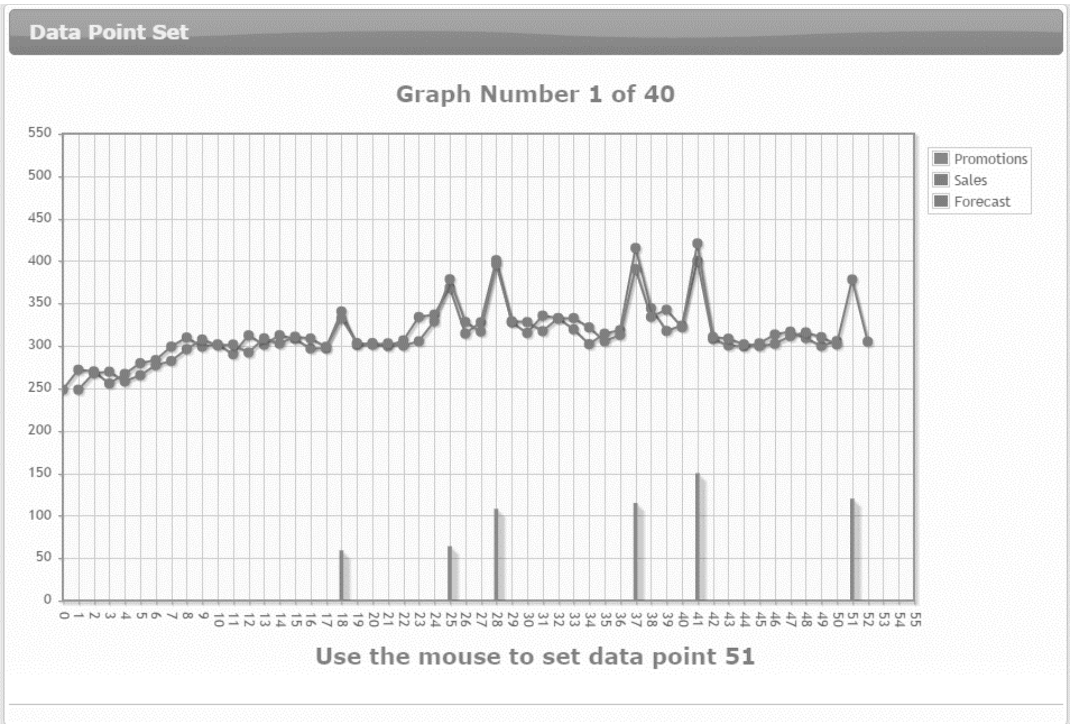


Figure 3. Experiment 2: Example of the screen display.

4.2. Results

Data for MAE, ME, and VE of forecasts for periods with and without promotions and from data series with low and high frequency of promotions are shown in Table 4. A repeated-measures ANOVA of the MAE revealed a main effect of the frequency of promotions (*F* (1, 49) = 30.15, *p* < .001), indicating that forecasts were more accurate with fewer promotions in the data series, and a main effect of presence of a promotion in the period to be forecast (*F* (1, 49) = 18.62, *p* < .001), showing that forecasts were more accurate for normal than for promotional periods. Analysis of ME revealed only a main effect of the frequency of promotions (*F* (1, 49) = 26.03, *p* < .001) indicating slight over-forecasting when data series contained 40% promotional periods but slight under-forecasting when they contained 10% promotional periods. The Variable Error indicated a main effect of the frequency of promotions (F (1, 49) = 11.37, p = .001) and a main effect of the presence of a promotion (F (1, 49) = 9.53, p = .003). The direction of these effects mirrored that of those obtained for MAE.

Table 4

Condition Means (MAE and ME) of Promotion Frequency and Presence of Promotion

Independent Variables		MAE	ME	VE
40% promotions	Promotion	25.61	1.99	6.28
	No promotion	23.09	2.54	5.74
10% promotions	Promotion	23.02	-2.8	5.70
	No promotion	19.50	-.59	5.19
Means		22.81	.29	5.73

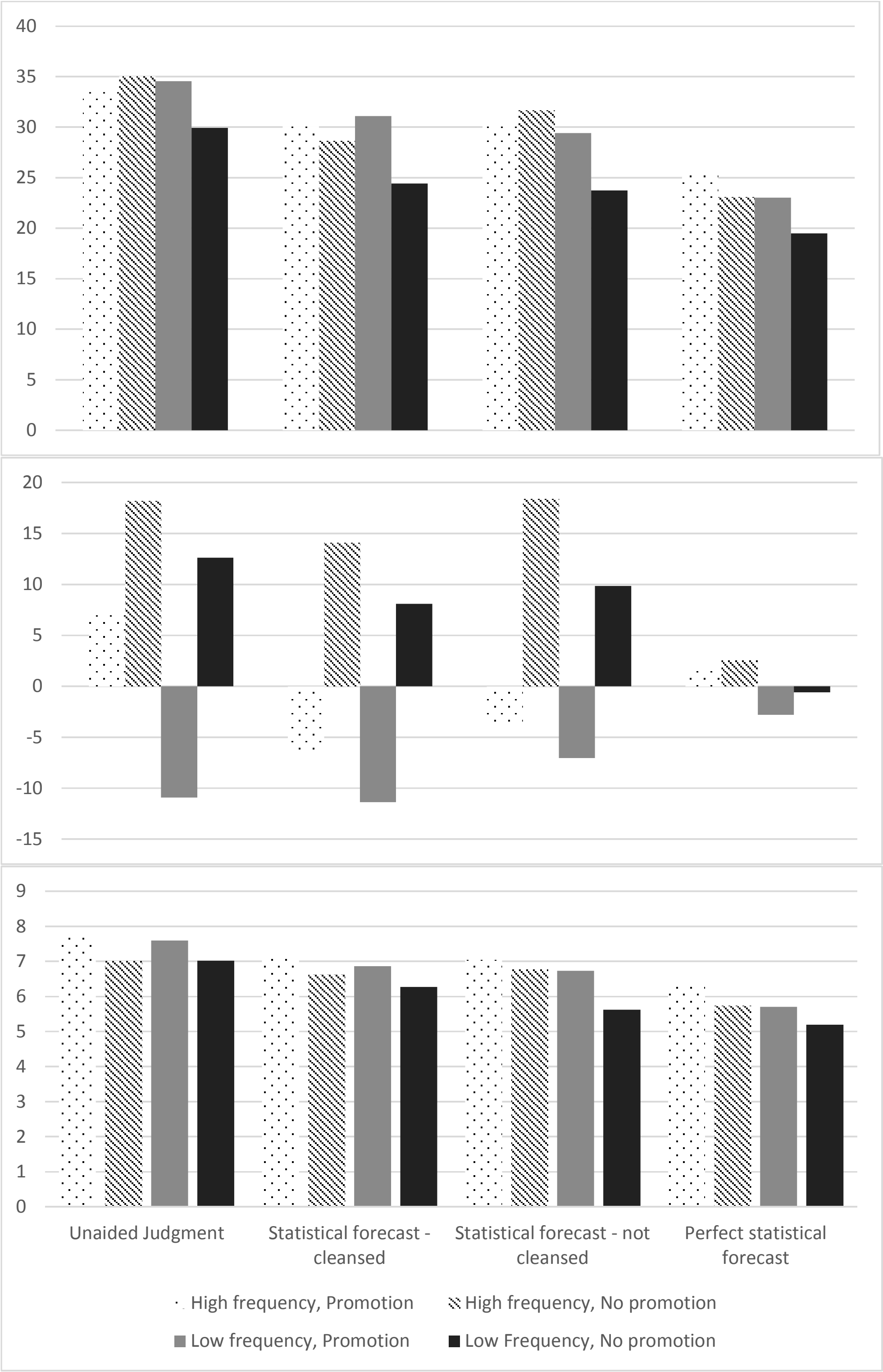


Figure 4. Error scores associated with different forecasting conditions studied in the two experiments:

MAE (upper panel); ME (central panel); VE (lower panel)

4.2.1. Comparison of performance with that obtained in Experiment 1

Did the enhanced statistical forecast provided in this experiment lead to better forecasting than that obtained by using unaided judgment or by using judgment aided by the provision of other types of statistical forecast? To find out, we compared forecast accuracy with that obtained in the three conditions of Experiment 1 (Figure 4). With regard to overall error (MAE), performance was significantly better than unaided judgment (F (1, 99) = 27.77, $p < .001$), aided judgment with a non-cleansed-series statistical forecast (F (1, 99) = 14.71, $p < .001$) and aided judgment with a cleansed-series forecast (F (1, 99) = 23.13, $p < .001$).

To compare the size of ME scores across experiments, we analyzed their absolute values. As hypothesized, those in the present experiment were lower than those in all conditions of the previous experiment: unaided judgment (t (58) = -6.12, $p < .001$); judgment aided with non-cleansed statistical forecasts (t (63) = -5.46, $p < .001$): judgment aided with cleansed statistical forecasts (t (60) = -4.56, $p < .001$)[1]. Similarly, the VE was significantly lower in the current experiment than it was in all conditions of the previous experiment: unaided judgment (t (77) = -6.84, $p < .001$), judgment aided with a non-cleansed statistical forecasts (t (67) = -2.81, $p = .003$); judgment aided with cleansed statistical forecasts (t (73) = -3.93, $p < .001$)[4].

4.3. Discussion

Provision of optimal statistical forecasts significantly reduced all types of error relative to the corresponding error levels observed in all conditions of Experiment 1. In particular, the absolute size of the directional error reduced very considerably. This implies that under-adjustment from the initial anchor was strongly attenuated. However, MAE scores show that a fair amount of error still persisted (Figure 4). This was primarily driven by VE. Although this type of error was significantly lower than it was in all the conditions of Experiment 1, it remained high at 83% of its size in that

[4] In cross-experimental comparisons of ME and VE, Levene's test indicated unequal variances and so degrees of freedom were adjusted accordingly.

experiment. As before, it was larger when there were more promotions in the data series and when promotions were planned for a forecast period. These influences on VE are likely to explain their re-appearance in the analyses of MAE.

Lim and O'Connor (1995, Experiment 3) obtained similar findings to ours. They found that forecasters made insufficient use of near-perfect statistical forecasts that were generated by taking the average of a highly reliable statistical forecast and the actual outcome. Forecasters put too much weight on their own views and not enough on the statistical forecast. Similarly, Gardner and Berry (1995) found that people performing a control task who were freely offered perfectly correct advice decided to obtain it on only 44% of occasions. Furthermore, those who obtained it acted in accordance with it on only 73% of occasions. One interpretation of both of these results is that people tend to be overconfident in their own abilities. As a result, they do not take sufficient account of good advice.

According to the account that we provided of the results from Experiment 1, forecasts are produced in two stages. First, forecasters (even those who are provided with statistical forecasts) make their own assessment of the mean of the data series to use as an initial judgment anchor. This assessment is subject to random error that is reflected in the VE scores. This random error is greater when data series are more variable. They are more variable when they contain a higher proportion of promotions and, hence, VE is greater when the proportion of promotions in the data series is higher. This same effect was found in the present experiment and so it is reasonable to assume that forecasters initially processed the series in a similar way in the present experiment.

The statistical forecasts examined in Experiment 1 were beneficial because they reduced VE. The statistical forecasts used in the present experiment also reduced VE. We suggested that this reduction occurs because forecasters can obtain estimates of the series mean both from the raw data series and from the series of past statistical forecasts. (Unaided judgmental forecasters can use only the data series.) A weighted average of these two estimates then provides the initial judgment anchor. If people are less confident in the statistical forecasts, they may put insufficient weight on the estimate obtained from them. As a result, VE may be reduced but not by as much as it could be. In the present experiment, the reduction in VE was greater than that produced by the statistical forecasts provided in Experiment 1. This may have been because the description of how statistical forecasts were generated provided in the instructions gave forecasters greater confidence in them: as a result, they put more weight on them and thereby generated a more accurate estimate of the series mean to use as an initial judgment anchor.

In the second stage, forecasters adjust away from the initial judgment anchor to take account of the presence or absence of a promotion in the period to be forecast. We saw in Experiment 1 that adjustment is typically insufficient (Tversky and Kahneman, 1974). As a result, promotional periods are under-forecast whereas normal periods are over-forecast. Adjustments for normal periods are based just on the mean value of normal periods in the data series but those for promotional periods have to take account of the relation between the size of promotions and the elevation in sales that they produce. This additional process is error-prone and therefore increases VE of forecasts for promotional periods relative to forecasts for normal periods.

This same effect (higher VE on promotional periods) was found in the present experiment. However, in contrast to Experiment 1, analyses of ME showed that there was no evidence of under-forecasting on promotional periods or of over-forecasting on normal ones. Thus, including an element allowing for promotions in statistical forecasts is beneficial not just because it reduces VE but also because it reduces the absolute size of ME. However, VE was still higher for forecasts for promotional periods than for those for normal ones. This implies that people do not merely accept the statistical forecast. Their low ME scores show that, on average, the mean value of their forecasts for both normal and promotional periods is very close to those provided by the statistical forecasts. However, there is considerable scatter around these mean values and this scatter is greater for forecasts for promotional periods. We attribute this greater scatter to additional error-prone cognitive processing that is needed to allow for the promotion function (i.e., the relation between promotion size and its effect).

Statistical forecasts that include an element to allow for the effects of promotions are beneficial because they reduce both bias and random error in forecasts. However, forecasters do not accept them automatically. This is clear not just from the high levels of VE that persist when statistical forecasts are provided but also from the fact that VE levels are affected by variables concerned with the nature of both the data series (proportion of promotions) and the periods to be forecast (normal or promotional). Furthermore, because the way that VE levels are affected by these variables when statistical forecasts (of whatever type) are provided is the same as the way in which they are affected in unaided judgmental forecasting, our view is that the provision of statistical forecasts does not fundamentally alter the cognitive processes that forecasters employ to perform their task. Instead, they facilitate these processes and do so more for some of them (e.g., the 'de-biasing' observed in Experiment 2) than for others (e.g., extracting an initial mental anchor from the data series). In other words, forecasters still used an anchoring-and-adjustment heuristic when given optimal statistical forecasts but their estimate of the appropriate anchor is somewhat more consistent and their adjustment from that anchor is almost free of bias.

5. General discussion

We provided forecasters with different types of statistical forecast to investigate how effective they are in improving forecasters' accuracy. We also varied the type of period to be forecast and the proportion of promotional periods in the data series because we expected these factors to influence the benefits that statistical forecasts bestow on forecasting performance. Finally, we developed an account of how forecasts are made from time series that are perturbed by sporadic events (i.e. promotions) and of how those forecasts are affected when forecasters have access to statistical forecasts. Here we discuss each of these aspects of our work in turn. All conclusions on error decomposition can be found in Table 5.

Table 5

Comparison of findings on MAE, ME and VE across factors

	MAE	ME	VE
Unaided judgment compared to aided judgment	Higher	No significant difference	Higher
Cleansed statistical forecast compared to non-cleansed statistical forecast	No significant difference	No significant difference	No significant difference
Optimal statistical forecast compared to suboptimal statistical forecast	Lower	Lower	Lower
Promotions compared to normal periods	Higher	Higher	Higher
10% compared to 40% promotions	Lower for normal periods; no significant difference for promotional periods	Lower for normal periods; higher for promotional periods	Lower for normal periods; lower for promotional periods

5.1. Effects of statistical forecasts on forecast accuracy

Statistical forecasts that take no account of whether periods in the data series were affected by sporadic events, such as promotions, provide the most common form of forecasting support for practitioners (e.g., Fildes, et al., 2009; Trapero, et al., 2013). However, in experimental research (e.g., Goodwin and Fildes, 1999; Goodwin et al, 2011), researchers have investigated the usefulness of statistical forecasts based only on normal periods not subject to promotions. We expected that the latter approach would be more effective in improving forecasting accuracy (Hypothesis 2).

While both of these types of statistical forecast improved accuracy relative to that observed with unaided judgmental forecasting (Hypothesis 1), there was no difference in the degree to which they did so. Given previous work by Lim and O'Connor (1995) and the persuasiveness of the arguments in favour of using statistical forecasts based on cleansed data series, this finding was unexpected. However, the rationale for Hyporthesis 2 was based on the assumption that statistical forecasts reduce bias: we anticipated that the anchoring bias for normal periods would be removed when statistical forecasts are based on cleansed rather than uncleansed series. In fact, our data show that statistical forecasts were effective because they reduced scatter (VE) rather than bias (ME) and there is no reason to expect scatter to be reduced more by statisical forecasts based on cleansed series than by statistical forecasts based on non-cleansed data series.

It appears that statistical forecasts that are clearly inadequate for promotional periods affect the degree to which forecasters feel able to trust them for normal periods (even when they are, in fact, optimal for those periods). We reasoned that statistical forecasts that are optimal for both promotional and normal periods should be seen as trustworthy and therefore be capable of reducing the anchoring biases. Experiment 2 demonstrated that this was so: ME values very close to zero showed that anchoring biases were virtually eliminated. However, VE values remained high at 83% of the level observed in the aided conditions of Experiment 1. Nevertheless, the marked drop in overall error (MAE) levels indicates that efforts to incorporate promotional effects into statistical forecasts (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, in press; Trapero, et al., 2013) hold great promise for increasing the effectiveness of forecasting support systems.

5.2. Effects of promotions in the periods to be forecast

We expected participants to anchor on the mean level of the data series and to adjust upwards/downwards from this to take account of the presence/absence of a promotion planned for the forecast period. As adjustment is typically insufficient (Tversky and Kahneman, 1974), we expected under-forecasting on promotional periods but over-forecasting on normal ones when forecasting was unaided or supported by a statistical forecast based on non-cleansed data series (Hypothesis 3a). This is indeed what we found, thereby confirming forecasters use of the anchoring heuristic. We expected that this anchoring bias would not be present on normal periods when statistical forecasts were based on cleansed data series as forecasters would realise that the statistical forecast could be accepted without adjustment (Hypothesis 3b). However, as we discussed in the previous section, these forecasts appear not to have been trusted (perhaps because those for promotional periods obviously needed adjustment). Forecasters continued to use the mean of the series as a judgment anchor and adjust down from it (insufficiently) to make forecasts for normal periods. Hence, over-forecasting for those periods persisted.

5.3. Effects of proportion of promotions in the data series

We expected that a lower proportion of promotional periods in the data series would reduce overall forecasting error on normal periods but increase it on promotional ones (Hypothesis 4). In fact, lowering the proportion of promotions resulted in a lower MAE on normal periods but promotional ones were unaffected. Decomposing overall error showed why this was so. On promotional periods, the absolute size of the under-forecasting bias increased when the proportion of promotions in the data series was reduced but scatter decreased. These two effects cancelled one another out and so there was no resultant effect on overall error. (For normal periods, reducing the proportion of promotions in the data series decreased both the over-forecasting bias and scatter: hence, the predicted effect occurred.)

When there were fewer promotional periods in the data series, statistical forecasts derived from non-cleansed series were closer to the baseline forecasts provided by the statistical forecasts derived from cleansed data series. Hence we expected any accuracy advantage of the statistical forecasts based on cleansed series (over the statistical forecasts based on non-cleansed series) would be greater when the proportion of promotions in the data series was higher (Hypothesis 5). However, there was no evidence of an interaction between proportion of promotions in the data series and type of statistical forecast. As we have seen, forecasters in Experiment 1 appear to have made their judgments in a similar way whether they were unaided or supported by either type of statistical forecast. The only reason that statistical forecasts helped was that they enabled them to make these judgments more consistently.

5.4. Forecasting from time series subject to sporadic perturbation

We have suggested that the cognitive processes underlying forecasting from time series subject to sporadic perturbation are broadly the same whether or not forecasting is aided by provision of statistical forecasts. This is particularly true for the two types of statistical forecast in current use: those that take no account of whether periods in the data series are normal or promotional and those that are based only on the normal periods. As Experiment 1 showed, anchoring effects and effects of proportion of promotions in the data series were unaffected by the presence of a statistical forecast or by its type when present. This implies that the way that the judgments were made was the same across all conditions of Experiment 1. The provision of statistical forecasts did improve accuracy but this was because they made judgment processes more consistent rather than because they changed the nature of those processes.

The optimal statistical forecasts provided in Experiment 2 virtually eliminated under-adjustment. However, VE values remained high. Furthermore, they were still affected by variables that affected VE in Experiment 1. We suspect that similar cognitive processes were responsible for performance in the two experiments. A mental anchor based on the mean of the data series was first extracted. This process was based on noisier data when the series contained more promotions, thereby explaining the effect of that variable on VE. The optimal forecasts ensured that, on average, the adjustments from this anchor were appropriate. However, VE was still higher when forecasts had to be made for promotional periods. To us, this implies that the adjustment process was more complex on promotional periods than on normal ones (because of the additional processing stage involved in extracting and using the relation between the size of a promotion and the effect that it had). Clearly, optimal statistical forecasts are not accepted automatically. They influence judgment but do not supersede it.

Why was there a considerable level of variable error, regardless of the presence and type of statistical forecast? Human forecasters introduce inconsistency or random error into forecasts. At least in part, this error is likely to arise from the noise that is inherent in cognitive processing. Since

Thurstone (1926), it has been known that judgment contains a random element. When people make a series of judgments about a criterion variable (e.g., salary levels of a number of different people) from information they are given about cue variables imperfectly correlated with the criterion (e.g., the weight, age, and nationality of those people), the relation between their judgment and the cues contains a random element (Brehmer, 1978) that decreases but does not disappear with practice and feedback. There are many hypotheses about why this occurs (Harvey, 1995). For example, Hammond and Summers (1972) referred to a failure of cognitive control: just as hand tremor causes inconsistency in the execution of fine motor skills, so some analogous process is affects judgment. Modern computational modelling of cognition is based on the notion that each component process contributes some random error to the total observed in the data (Lewandowsky and Farrell, 2011).

Noise inherent in cognitive processing is unlikely to be the only reason for high VE levels. Lawrence et al (2006, p 501) suggest that small damaging adjustments of the sort reported by Fildes et al (2009) may reflect “a tendency to tinker at the edges”. In other words, forecasters *intend* to introduce these small changes that do not, overall, lead to (greater) over-forecasting or under-forecasting but do increase scatter. But why would forecasters do this?

There are various possibilities. One is that the changes that they make provide them with a way in which to assert their ‘ownership’ of the forecasts (Önkal & Gönul, 2005). Another concerns people’s responses to automation. Whenever tasks become partially automated, concerns tend to arise among those responsible for performing them that they risk becoming de-skilled (Bainbridge, 1983). Without feedback about the effects of their own actions, they will not be able to acquire or maintain the abilities that they need to perform their tasks autonomously (something that may be needed if the automated system suddenly becomes unavailable). Hence, to ensure they receive such feedback, operators may occasionally over-rule or interfere with the output produced by the automatic system. (For forecasters, receiving feedback about statistical forecasts is no substitute for receiving it about the forecasts that they have generated themselves: only in the latter case is the rationale for the forecasts known.)

5.5. Practical implications

Our main message for practitioners is that the provision of statistical forecasts reduces forecast error but whether those statistical forecasts are based on data cleansed of promotional effects does not matter. This knowledge could save time and money because it implies that cleansing the data (which in itself is subject to biases; Webby, et al., 2005) is unnecessary. Even a relatively simple statistical forecast can be of value for a company. Hence, companies that wish to improve their forecasting accuracy but do not currently have a large budget or manpower to spare can still benefit from a simple approach that requires minimal effort.

The results of the second experiment demonstrated that highly sophisticated statistical forecasts that explicitly take account of the effects of promotions benefit forecasters considerably more than those that do not. Efforts made to develop ways of producing such forecasts (e.g., Huang, et al., 2014; Kourentzes & Petropoulos, in press; Trapero, et al., 2013) are clearly worthwhile. However, this second experiment also showed that even forecasters who are given optimal statistical forecasts make adjustments that impair accuracy. As we have seen, there are different ways of explaining this finding but they all imply that, for one reason or another, forecasters are not good at taking 'advice' from a statistical model. Such discounting of advice has been reported before (e.g., Goodwin, 2000; Lim & O'Connor, 1995) and factors that have been proposed to account for it include concerns about the credibility of a statistical model rather than a human being as a source of advice (Önkal, et al., 2008; Önkal, et al., 2009), and people's beliefs that their own opinions are better founded than those of others (Harvey & Harries, 2004).

Preventing damaging adjustments has been an important topic in judgmental forecasting. Goodwin et al. (2011) found that neither restriction nor guidance improved accuracy. Indeed guidance was met with resistance by forecasters. Such resistance is consistent with Bainbridge's (1983) views about responses to automation. However, as we pointed out, reasons that forecasters make damaging adjustments may not be purely volitional (i.e., arising because, for one reason or another, they *want* to make those adjustments) but may also be at least partly cognitive (i.e., noise may be inherent in the cognitive processes that underlie forecasting).

5.6. Limitations

Our study is potentially subject to limitations. The first of these, use of student participants, appears to be straightforward. It could be argued that experts have more insight into how statistical forecasts should be used. However, previous work has shown that experts are subject to similar errors in reasoning as those that afflict novices. Indeed, in some cases, it has even revealed inverse expertise effects (Önkal & Muradoğlu, 1994; Yates, McDaniel, & Brown, 1991). Advice discounting may be even greater in experts because they value their own opinion even more than novices do. In fact, Önkal and Muradoğlu (1994) demonstrated that experts exhibited even more over-confidence in their forecasts than those who were less expert. This situation is typical of what happens when experience at a task (e.g., forecasting) fails to produce learning as quickly as people expect it to (Harvey and Fischer, 2005).

An experiment is a simulation or model of a task performed by practitioners. As with any model, some features of the real world task are excluded. Thus we do not expect to see all characteristics of practitioner performance reflected in experimental results. Analysis of data obtained from organizations has revealed that forecasters are often subject to optimism effects: inappropriate upward adjustments of statistical forecasts are greater or made more often than inappropriate downward ones (e.g., Fildes et al, 2009). We did not observe such optimism in our experiments. They were not designed to study or reveal it. All the same, it is possible to argue that optimism would have produced less under-forecasting on promotional periods than over-forecasting on normal periods. We did not find this pattern in the data. However, this prediction does not compare like with like. As we have emphasized, processes underlying forecasting on promotional periods are different from those that underlie it on normal ones. To research into optimism experimentally, studies should be specifically designed with that aim in mind. One approach is to compare two groups performing exactly the same forecasting task but to label the variable being forecast as ‘profits’ in one case but ‘losses’ in the other. Forecasts are systematically higher in the former case (Harvey and Reimers, 2013).

5.7. Conclusions

Provision of statistical forecasts, even crude ones, can improve forecasting accuracy by reducing variable error. When forecasts are made from time series perturbed by sporadic exogenous events, the effort needed to produce forecasts cleansed of their effects appears not be warranted. However, current efforts to develop methods to incorporate effects of these events into statistical forecasts are worthwhile and are likely to result in improved forecast accuracy.

7. References

Alvarado-Valencia, J., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior, 36*, 102 - 113.

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In J. S. Armstrong (Ed.), *Principles of forecasting*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a* handbook for researchers and practitioners (pp. 417 - 439). New York: Kluwer.

Bainbridge, L. (1983). Ironies of automation. Automatica, 19, 775-779.

Baird, J. C., & Noma, E. (1978). Fundamentals of scaling and psychophysics. New York: Wiley.

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

Bovi, M. (2009). Economic versus psychological forecasting. Evidence from consumer confidence surveys. *Journal of Economic Psychology, 30*, 563 - 574.

Brehmer, B. (1978). Response consistencty in probabilistic inference tasks. *Organizational Behavior and Human Decision Processes, 22*, 103-115.

Durand, R. (2003). Predicting a firm's forecasting ability: the roles of organizational illusion of control and organizational attention. Strategic Management Journal, 24(9), 821 - 838.

Fechner, G. T. (1860). Elemente der psychophysik [Elements of psychophysics] (Volume 1). Leipzig: Breitkopf und Harterl.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces, 37(6), 570-576.

Fildes, R., & Goodwin, P. (2013). Forecasting support systems: What we know, what we need to know. *International Journal of Forecasting, 29*(2), 290 - 294.

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*(1), 3 - 23.

Gardner, D. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology*, **9,** 555–579.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems, 42*, 1481–1493.

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*(1), 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting, 23*(3), 391 - 404.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*(3), 242 - 253.

Hammond, K. R. & Summers, D. A. (1972). Cognitive control. *Psychological Review, 79,* 58-67.

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

Harvey N and Fischer, I. (2005). Development of experience-based judgment and decision-making: The role of outcome feedback. In T. Betsch and S. Haberstroh (Eds). *The Routines of Decision-Making.* Erlbaum: Mahwah, NJ, pp. 119–137.

Harvey, N., & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the same outcomes. *International Journal of Forecasting, 20*(3), 391 - 409.

Harvey, N., & Reimers, S. (2013). Trend damping: under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology, 39*(2), 589 - 607.

Hilary, G., & Hsu, C. (2011). Endogenous overconfidence in managerial forecasts. *Journal of Accounting and Economics, 51*(3), 300 - 313.

Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research, 237*(2), 738 - 748.

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

Kourentzes, N., & Petropoulos, F. (in press). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics.*

Lawrence, M. (2000). Editorial: What does it take to achieve adoption in sales forecasting? *International Journal of Forecasting, 16,* 147 - 148.

Lawrence, M. and O'Connor, M. (1992). Exploring judgmental forecasting. *International Journal ofForecasting, 8 ,* 15-26.

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

Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice.* London: Sage.

Libby, R., & Rennekamp, K. (2012). Self-Serving Attribution Bias, Overconfidence, and the Issuance of Management Forecasts. *Journal of Accounting Research, 50*(1), 197 - 231.

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.

Önkal, D., & Gönul, M. S. (2005). Judgmental adjustment: A challenge for providers and users of forecasts. *Foresight: The International Journal of Applied Forecasting, 1*(1), 13-17.

Önkal, D., Gönul, S., & Lawrence, M. (2008). Judgmental adjustments of previously adjusted forecasts. *Decision Sciences, 39*(2), 213 - 238.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Önkal, D., & Muradoğlu. (1994). Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research, 74*(2), 350 - 358.

Önkal, D., Sayim, K. Z., & Lawrence, M. (2012). Wisdom of group forecasts: Does role-playing play a role? *Omega, 40*(6), 693 - 702.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantiative forecasting methods in practice. *Omega, 31*, 511 - 522.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology, 17*, 446 - 457.

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*(2), 234 - 243.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124 - 1131.

Webby, R., O'Connor, M., & Edmundson, R. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of* Forecasting, 21(3), 411 - 423.

Weber, E. H. (1834). De pulsu, resorptione, auditu et tactu [On stimulation, response, hearing and touch]. Annotationes, anatomical et physiological. Leipzig: Koehler.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior & Human Decision Processes, 40*, 60 - 79.

Chapter 5

Conclusion

Concluding Chapter

1. General conclusions

This doctoral thesis set out to investigate the benefits of combining judgmental and statistical forecasting. When does judgment add value on top of statistical models? Or vice versa, when does a statistical model add value when combined with unaided judgment? Can we provide a forecast support system to increase the accuracy of this combination of judgment and statistics? This thesis focussed specifically on forecasts disturbed by external events (promotions). This type of time series is especially relevant for the combination of judgment with statistical methods. Statistical methods are known for their consistency and ability to handle large amounts of data (Blattberg & Hoch, 1990; Hoch & Schkade, 1996). Yet, they cannot deal well with exogenous events (Armstrong & Collopy, 1998; Goodwin, 2002; Hughes, 2001; Taleb, 2007). A forecaster's judgment can provide valuable information in addition to the system forecast: its strengths lie in the interpretation of data, the detection of unusual events and the integration of non-quantifiable information. However, judgment is notoriously double-sided: it has a much wider application range than a statistical forecast, but it is also heavily subject to bias (Kahneman, 2011; Tversky & Kahneman, 1974). To dig deeper into the circumstances surrounding judgmental forecasting in time series with disturbances, this doctoral thesis investigated the issue by means of two different methods.

The first involved employing a dataset from a publishing company, with real forecasts made by expert forecasters. Our first research objective was to design a forecast support system that can improve on judgmentally adjusted forecasts. Importantly, this system had to increase accuracy while simultaneously allowing judgmental input. Why was the latter so important? Goodwin, et al. (2011) tested a forecast support system which was based on restriction. Based on the finding that small adjustments tend to be damaging (a result replicated in the first study of this thesis; Fildes, et al., 2009), they designed the forecast support system so that small adjustments were prohibited. This restriction turned out to be counter-productive and damaged accuracy: forecasters no longer adjusted when it was necessary, and made more overly large adjustments. So, when developing a forecast support system, we have to take the acceptance by the user into account. Forecasters should be able to retain their sense of ownership (Silver, 1991). Thus, we tested a forecast support system which allowed for the integration of judgment into the model itself. The properties of the dataset and the different methods were analysed in-depth (direction and size of adjustments, volatility of the data measured in two ways). We linked the model to demand planning by the use of an optimization model, distinguished between different hierarchical levels (as recommended by Kremer, et al., 2015), and provided information on profitability. To the best of our knowledge, this is the first study to provide such data. We found that the model eliminated the harmful effects of judgment when the latter was applied wrongfully. However, in some cases the restrictive model, or the classic case of judgmental adjustment, proved to be beneficial. We confirmed previous literature, in that medium sized and big adjustments appear to be beneficial, as well as downward adjustments and adjustments in case of high volatility. Presumably, this is because these adjustments stem from knowledge that is not yet integrated in the statistical forecasting model. This is in contrast with small adjustments, adjustments in the case of low volatility and positive adjustments, which are more likely to be the result of the illusion of control effect, or an inherent optimism bias (Fildes, et al., 2009).

Second, we employed the experimental method for fundamental research into the biases associated with providing forecasts from a statistical model to a judgmental forecaster. In our experiments, we set out to compare unaided judgment and judgment aided by (different types of) statistical forecasts. In Chapter 3, we asked whether judgmental forecasts would improve when we provided a statistical forecast. We found that performance was in fact impaired, compared to unaided judgmental forecasting. While previous studies found lack of use or insufficient use of such forecasts (e.g., Goodwin & Fildes, 1999; Lim & O'Connor, 1995), we found a damaging effect when compared to unaided judgmental forecasting. Further testing of the experiment, but with a lower proportion of promotions, resulted in a lowering of the overall error and, importantly, in a disappearance of the detrimental effect of a statistical forecast on predicting the value of normal periods. Given that very little adjustment from the mean was necessary for normal periods in the experiment with a low proportion of promotions, we concluded that the presence of the statistical forecast interfered with the adjustment process. Importantly, this suggests that we need to be careful in viewing statistical forecasts as universally better than unaided judgment when forecasting from time series disturbed by exogenous events. As this was a novel experimental finding, but one that was quite robust across our three experiments, we set out to investigate this further in the fourth Chapter.

Rather than varying the forecasting task (with or without provision of a statistical forecast) within participants, this variable was manipulated between participants. One group of participants never received a statistical forecast and had to rely solely on their judgment. Another group received a statistical forecast in a similar way to participants in the experiments reported in Chapter 3. Perhaps the detrimental effect obtained in the previous study was due to the inadequacy of the type of statistical forecast. A third group received a forecast based on series cleansed of promotional effects. While this does not appear to be common practice (e.g., Fildes, et al., 2009; Trapero, et al., 2013), it is the most common method used in past experimental research (e.g., Goodwin & Fildes, 1999; Goodwin, et al., 2011). Previous studies comparing the effect of the initial quality of the forecast are scarce (except for Lim & O'Connor, 1995). In this set-up, the formal method proved beneficial for forecasting accuracy, regardless of the basis of the forecast: both cleansed and non-cleansed forecasts led to an improvement in forecasting. Decomposing the error taught us that the benefit of providing a statistical forecast lies in the reduction of the Variable Error (scatter). While there was an improvement in accuracy with the presence of a statistical forecast, a large amount of error still persisted, both in forecasting normal periods and promotions. We tested an additional type of statistical forecast in a second experiment: one based on cleansed series and with an additional uplift for promotional periods. This optimal forecast led to a significantly better performance, but was still not optimal. It seems that even with an optimal forecast, the tendency to adjust persists.

In this doctoral thesis, damaging adjustments were found across the different samples: both by novices in a controlled, experimental setting, and by the professional forecasters of the publishing company (in the restricted judgment model). Chapter 3 and 4 indicate that the introduction of a formal forecast is associated with potential problems. In Chapter 3 the presence of a statistical forecast led to a worse performance. In Chapter 4 it led to a better performance but still far from optimal – even with an optimal statistical forecast (see below for an in-depth analysis of the cause of the difference in findings). Lack of use and limited use of statistical forecasts is a persistent problem in judgmental forecasting (Lawrence, et al., 2006; Önkal, et al., 2009). The integrative approach of Chapter 2 can provide a future avenue for de-biasing the combination of judgment and statistics. However, it must be noted that, while it eliminated the harmful type of adjustments, it also reduced the advantages of restrictive judgment.

2. Methodological contributions

The difference between the findings of Chapters 3 and 4 makes for an interesting methodological finding. In Chapter 3, we found that the provision of a statistical forecast damaged accuracy. In Chapter 4, it improved forecasting accuracy. This finding is surprising, as there are many similarities in both studies. In all experiments, the same stimulus material was used and the sample was similar. The only difference was in the experimental design: Experiment 1, 2 and 3 of Chapter 3 were within-subject designs with respect to the focal variable (presence or absence of the statistical forecast). Each participant received 40 trials, of which (at random) half contained a statistical forecast and half did not. In Chapter 4, each participant received 40 trials as well. However, they received either all trials with a statistical forecast, or all trials without a forecast (unaided judgment). In Chapter 4, the within-subjects manipulation was the number of promotions (40% versus 10%). This was based on the findings of Experiment 1 (40% promotions) and Experiment 3 (10% promotions) of Chapter 3. So what caused the difference between the findings in Chapter 3 and Chapter 4? It seems that the context in which the forecasting task is performed has a significant effect on the difficulty of the task. Looking at mean absolute errors, we can compare several condition types across experiments (see Table 1).

		Chapter 3	Chapter 4
Unaided judgment	40% promotions	Exp 1: 20 unaided judgment trials	Exp 1, Condition 1: 20 trials with 40% promotions
	10% promotions	Exp 3: 20 unaided judgment trials	Exp 1, Condition 1: 20 trials with 10% promotions
Aided judgment with a non-cleansed forecast	40% promotions	Exp 1: 20 trials aided with a statistical forecast	Exp 1, Condition 3: 20 trials with 40% promotions
	10% promotions	Exp 3: 20 trials aided with a statistical forecast	Exp 1, Condition 3: 20 trials with 10% promotions

Table 1. Comparisons across experiments

Table 2 provides a comparison as described above. Looking at the experiment results in the first column, where the presence of the statistical forecast was manipulated within subjects, the error for unaided judgment is lower than the error for aided judgment. This is true both for the 40% promotion trials as the 10% promotion trials. Looking at the results in column 2 of the experiment where the statistical forecast was manipulated between subjects, we see that the opposite is true: the error for unaided judgment is higher than the error for aided judgment. This is again regardless of the amount of promotions.

Independent Variables		Experimental design			
		Within (Ch.3)	Between (Ch.4)	t	p
Unaided	40% pr	31.13	31.22	-.06	.954
Judgment	10% pr	23.51	28.93	-3.93	.000
Non-cleansed	40% pr	33.70	29.12	3.66	.000
statistical forecast	10% pr	26.02	24.81	.881	.381

Table 2. Comparison of MAE across Experiments of Chapter 3 and Chapter 4

Now, comparing across experiments and across Chapters, we can investigate where this difference occurs. If we look at the 40% promotion conditions across experiments, we find a first clue. The MAE for 40% promotions is not significantly different for unaided judgment in the within-subject trials (*MAE* = 31.13, *SD* = 8.30) compared to unaided judgment in the between-subject trials (*MAE* = 31.13, *SD* = 8.30; *t* (88) = -.06, *p* = .954). Thus, for unaided judgment with a high frequency of promotions, performance is consistent, regardless of the methodological set-up.

It is only in the comparison with a statistical forecast, that a difference emerges: participants forecasted worse when a statistical forecast appeared in the within-subjects experiment (*MAE* = 33.70, *SD* = 5.94). However, in the between-subjects design, those participants that always received a statistical forecast, performed better than those who always relied solely on their judgment in the 40% promotions trials (*MAE* = 29.12, *SD* = 5.93). One possible explanation is that the appearance of a statistical forecast confused participants in the within-subjects variation. They were thrown off their game and started making more mistakes, due to an increased complexity of the task. Looking at cognitive psychology literature, it appears that the frequent task switching introduced error in the within-subjects experiment. Task-switching has two effects on performance: a transient cost is introduced by the switching, resulting in increased response time and often a higher error rate and a long-term cost, such that responses remain slower throughout the trials compared to a block with only one type of trials (Monsell, 2003). These costs are due to reconfiguration required of the task, such as a shift of attention (in this case, the introduction of an additional line on the graph) or a change in action rules (in this case, taking into account the ‘advice’ of the statistical forecast).

Looking at the low proportion of promotions trials, the pattern is different. Participants who forecasted all trials with only 10% promotions (*MAE* = 23.51, *SD* = 5.59), performed better when forecasting with judgment alone (within-subjects), than those in the unaided judgment condition (between-subjects) who switched between 10% and 40% promotions (*MAE* = 28.93, *SD* = 7.02; *t* (87) = -3.93, *p* < .001). When provided with a statistical forecast, the performance of the within-subjects group deteriorated (*MAE* = 26.02, *SD* = 6.20). In the between-subjects design, the participants that were provided with a statistical forecast in every trial (between-subjects, *MAE* = 24.81, *SD* = 6.60), had a better performance than those who forecast unaided in every trial. Further investigation here is necessary. A possible way forward here is to work with a block design. Ideally, to eliminate all effects of task-switching, both frequency of promotions (10% vs. 40%) and statistical forecast (present vs. absent) should be manipulated within-subjects. This would result in four between-subject conditions: unaided judgment with a low frequency of promotions, unaided judgment with a high frequency of promotions, aided judgment with a low frequency and aided judgment with a high frequency.

Another interesting avenue would be to look into the effect of switching after a long block of identical trial types. If participants have 20 occurrences of unaided judgment with a low frequency, will their performance drop after the introduction of a statistical forecast? A potential outcome could be that performance increases in the first twenty trials due to learning, and then drops significantly after switching to the other type. Thus, the research questions would be: do participants learn across trials? Is this learning disturbed by switching from unaided judgment to aided judgment or vice versa? We can already provide an indication. Let us take, for example, the results of the third within-subjects experiment. Error was lowest in the third experiment (10% promotions); we can therefore assume that the effect of switching tasks was least disturbed by other variables. Do participants learn across trials? Looking at the average mean error (Figure 1), it seems that they do not.

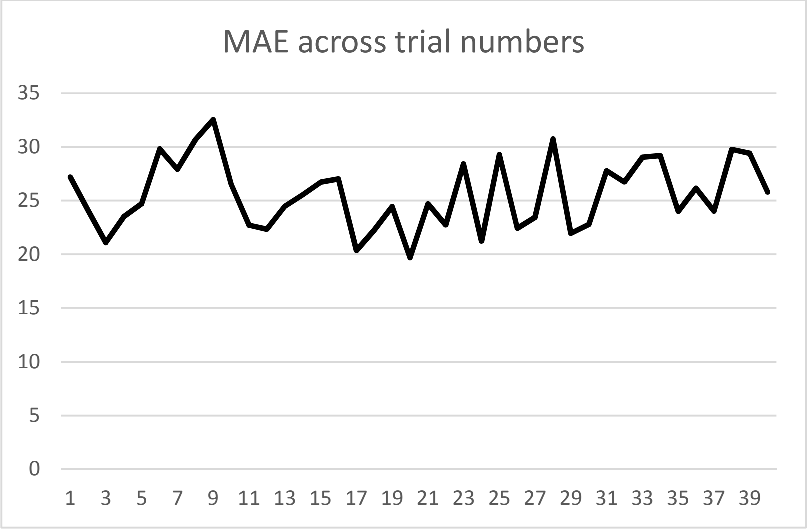


Figure 1. MAE across trial types

To investigate this statistically, we regressed the MAE on the trial number. This confirmed that the trial number does not significantly predict the MAE (*F* (1, 38) = .39, *p* = .537; R² = .01)[5]. To test whether participants might have taken a few trials to learn, the regression analysis was run again but with blocked trials: the 40 trials were divided into 5 learning blocks. The regression analysis again indicated that there was not a significant learning effect across blocks (*F* (1, 38) = .23, *p* = .635; R² = .01). A potential explanation is that the learning process was disturbed by the random switching from trial types with a statistical forecast to those without a statistical forecast. To investigate this, we looked at the trial order for every participant. For every subsequent trial number, we coded whether this was the same or a different trial type with regard to the presence of the statistical forecast (see Figure 2).

[5] The regression analyses for Chapter 4 were similarly non-significant for learning effects for unaided judgment (*F* (1, 38) = 1.38, *p* = .248; R² = .04), judgment aided with a forecast based on cleansed series (*F* (1, 38) = .24, *p* = .628; R² = .01), aided with a forecast based on non-cleansed series (*F* (1, 38) = 1.87, *p* = .179; R² = .05) and judgment aided with an optimal forecast (*F* (1, 38) = .19, *p* = .666; R² = .01).

Part.	trial 1	trial 2	trial 3	trial 4	trial 5
1	NO	SF	SF	SF	NO
2	NO	NO	NO	NO	SF
3	NO	NO	NO	NO	NO
4	SF	NO	NO	NO	NO
5	SF	NO	SF	SF	SF

⇨

Part.	trial 1	trial 2	trial 3	trial 4	trial 5
1		Diff	Same	Same	Diff
2		Same	Same	Same	Diff
3		Same	Same	Same	Same
4		Diff	Same	Same	Same
5		Diff	Diff	Same	Same

Figure 2. Categorization of trial types

We can ask whether participants who had more switches performed worse. It appears that there is no correlation between number of switches and the Mean Absolute Error (*r* (38) = -.188, p = .259). To look more closely at these results, we will need a blocked design (as described above) without the task switching from the current dataset.

3. Theoretical implications

Empirical and theoretical studies regarding judgmental adjustment (i.e., 'restrictive judgment' in Chapter 2, 'aided judgment' in Chapter 3 and 4) and judgmental forecasting are relatively scarce. The 2009 study of Fildes et al. (2009) re-started the stream of judgmental adjustment research since the landmark studies of Mathews and Diamantopolous (1986, 1989, 1990, 1992). The studies that have followed since (e.g., Franses & Legerstee, 2009; Trapero, et al., 2013) have focussed on contributing to empirical work. Recently, Syntetos, Babai, Boylan, Kolassa, and Nikolopoulos (2016, p. 15) remarked that ".. there is still considerable space for empirical investigations and more importantly theoretical and methodological contributions towards identifying the specific conditions under which judgmental adjustments do lead to improvements in forecast accuracy and forecast utility". They further state that, with the exception of Syntetos, Georgantzas, Boylan, and Dangerfield (2011), no papers currently exist within the operational research framework (in which the study in Chapter 2 is situated) that use formal models. Theoretical frameworks for judgmental adjustment as such do not exist, to the best of our knowledge. However, we can look at the broader field in which judgmental forecasting is situated: therefore, I first situate these studies at the cross-roads between Economics and Psychology.

3.1 Economics

A theory that goes back to the late sixties, but is still relevant today, is the Theory of Combining Forecasts (Bates & Granger, 1969). It recognizes that different models may be able to capture different, independent information and that by combining we can capture reality more accurately. Bates and Granger (1969) state one condition that poses a problem for applying the theory to judgmental adjustment research: namely, that the nature of the individual forecasts must be unbiased. A first step is therefore to check for unbiasedness and should bias be present, to correct for it. Bates and Granger (1969) suggest that this should be done via a correction for the average percentage (or absolute) bias. The study of Chapter 2 can be placed under this Theory for Combining Forecasts. The basic statistical model is outperformed on a number of occasions by the restrictive judgment model (combination via judgmental adjustment). However, in some cases restrictive judgment was harmful. Small adjustments for instance, were harmful and explained by the illusion of control effect, violating the unbiasedness assumption of the Theory of Combining Forecasts. Thus, we corrected them via the integrative judgment approach. After this correction, the combination of forecasts outperformed the single method of the statistical model, as we would expect within the framework of the Theory of Combining Forecasts.

Later theories have focussed on singular models for forecasting. Hendry and Clements (2003) discuss the failure of Basic Economic Theory as a framework for economic forecasting. The two main assumptions of this theory state that the model is a good representation of the economy, and the structure of the economy remains unchanged. Unsurprisingly, both assumptions are violated in reality. A less stringent theory is the Forecasting Theory proposed by Clements and Hendry (1999), based on the assumptions that models are simplified representations of reality and economies can change. The first assumption allows for error in the model, while the second recognizes the instable nature of time series. Both assumptions can be applied to forecasting beyond economic series: organizational forecasting is equally subject to error in their models and change in the environment. Overall, it appears that the oldest theory of all, is still most applicable – finding predictive accuracy in the combination of models.

3.2 Psychology

Across experiments in Chapter 3 and 4, we find evidence of under- and over-forecasting. Looking at the other overarching field of judgmental forecasting, Psychology, we find evidence in Cognitive Bias theory. Cognitive biases are deviations from rational, prescribed, or ideal judgments (Kerr, MacCoun, & Kramer, 1996). The deviations are different from random error, in that they are predictable and systematic (Arnott, 2006; Kerr, et al., 1996). Many different taxonomies of biases exist (e.g., Haselton et al., 2009; Oreg & Bayazit, 2009). The most relevant to the studies in this doctoral thesis is the very first taxonomy of cognitive biases, by Tversky and Kahneman (1974). They defined three main biases: representativeness, availability, and anchor-and-adjust. The latter bias is especially relevant for judgmental adjustment. The studies of Chapter 3 and 4 illustrate how forecasters take an initial value, being the statistical forecast, the mean of the series or the last data point, and adjust insufficiently upwards or downwards.

Other relevant theories from the field of psychology relate back to the task switching effect we mentioned above. Several theories relating to the study of judgment highlight that judgment cannot be understood or studied without taking the environment or the task into account (e.g., Cognitive Continuum Theory, Social Judgment Theory; Hammond, 1981; Hammond, Stewart, Brehmer, & Steinman, 1975). Cognitive Continuum Theory (Hammond, 1981) for instance, states that the task has three characteristics: complexity of the structure, ambiguity of the content, and form of the representation. The randomization of the trials in our experiments might have increased the complexity of the structure, such that it interfered with the performance of the participants.

3.3 Specific contributions

The paragraphs above situate the studies of this doctoral thesis within the broader fields of economics and psychology. Moving away from the large theoretical frameworks of the overarching fields of forecasting, we can specify a number of contributions to sub-fields of forecasting. Looking at *forecasting with judgment aided by Forecast Support Systems*, Chapter 2 contributes by testing the theory that the incorporation of judgment into the forecasting model itself would be beneficial for forecasting accuracy. It appears that it can compensate for damaging adjustments. Yet, restrictive judgment sometimes outperformed integrative judgment. Forecast support systems will need to be able to recognize when judgment is better than its automated counterpart. Chapter 4, Experiment 1, showed us why statistical forecasts are able to improve judgment when they are provided as aid: they reduce variable error and make the forecasts more consistent. Experiment 2 with an optimal forecast showed that a forecast able to incorporate promotional effects can result in a significant drop in overall error (MAE). For the literature on *forecasting promotions*, we refer back to the contributions mentioned above on anchoring bias. Lastly, in Chapter 4 we theorized about the process involved in *forecasting from disturbed series*. It appears that a two-stage process is involved, whereby people first create an internal baseline of the series based on the overall perceived mean. Second, they adjust for the estimated effect of the perturbations.

4. Managerial implications

An important task for researchers in the field of judgmental forecasting, is to explore avenues that may contribute to the improvement of forecasting accuracy in practice (Sanders & Manrodt, 2003). Increased understanding of the forecasting process is of critical importance for today's organizations, as it offers the potential for a competitive advantage in forecasting. It is therefore vital for managers to gain a more complete understanding of the forecasting process and the possible influencing factors (Fildes & Goodwin, 2007; Giullian, et al., 2000; Makridakis & Gaba, 1998; Shim, 2000). Forecasting can be labour intensive, and therefore a costly and time consuming process, especially if time is invested in adjusting statistical results (Adya, Collopy, Armstrong, & Kennedy, 2001; Mathews & Diamantopolous, 1989). Studying possible ways to unify judgment and formal models in an optimal manner can provide direction for efficient cost and time investments. Research may provide businesses with a measurable foundation for strategic decisions such as staff planning and ICT processes, optimizing output and quality, cost efficiency, and risk management. More concretely, this will lead to increased profits: even small improvements in predictive performance can have a substantial impact on a firm's profitability (Steenburgh, et al., 2003), as was confirmed in the results of Chapter 2. That Chapter is highly relevant to management practice, as it offers a solution for using both statistical models and judgmental adjustment, without the known damaging effects of the latter. Research has long since proven that judgmental adjustments, while holding much potential, suffer from persistent biases (Eroglu & Croxton, 2010; Fildes, et al., 2009; Webby & O'Connor, 1996). The provision of a decision support system can alleviate these biases (Fildes, et al., 2006); however, the introduction of these support systems is often met with resistance (Goodwin, et al., 2007). In a similar way to the company which provided us the dataset, other companies could adopt the integrative judgment method to improve performance, while allowing for the input of the forecaster. Importantly, this study confirmed the findings of Fildes, et al. (2009), in that there are certain circumstances under which we should trust judgment: medium to large adjustments, high volatility, downward adjustments. It should be noted that following the study documented in Chapter 2, the company involved is cultivating a forecasting culture where the forecasters are motivated to use the integrative approach.

The tentative message of Chapter 3 is: be cautious in using statistical forecasts when dealing with time series with promotions. However, Chapter 4 suggests otherwise. Given that the situation of

Chapter 3 is less likely to appear in practice (forecasters rarely forecast the time series with and without statistical forecasts), the focus will lie on the managerial relevance of Chapter 4. This does not however, take away from the important methodological finding that a difference in experimental design can lead to opposite results. This is more of an academic take-away point, with greater relevance to psychology and experimental research, than it is for business practice.

In Chapter 4, we bridged a gap between experimental design and research practice: the nature of the statistical forecast. An important message for management practice is that the provision of a statistical forecast helps to improve performance, even if it is a crude forecast that is provided. Additionally, it appears that the time, cost and effort invested in cleaning the data is not warranted. Looking specifically at promotions, the experiments suggest that forecasts are heavily subject to anchoring bias. Promotional values were consistently under-forecasted. A suggestion for business practice could therefore lie in the format in which they ask for such forecasts. The anchoring could be avoided by not working with an elevation on a graph, but by asking the forecasters to forecast a percentage increase or an absolute number. This can be added to the graph in a next step, if the company prefers a visual representation. It should be noted that when taking this approach, the forecaster is vulnerable to optimism bias (over-forecasting desirable quantities) leading to a reversal of the initial under-forecasting.

5. Limitations, strengths and further research

By employing both a company sample with forecasts made by experts (Chapter 2), and experiments with student participants (Chapter 3 and 4), this doctoral thesis attempts to balance out the advantages and disadvantages associated with both methods. A company sample has the advantage of having genuine forecasters as participants, using real-life data, subject to context effects. Given the considerable impact of the organizational context on forecasting performance (Adelman, 1981; Franses, 2013; Sniezek, 1986; Wright & Goodwin, 1998), studies like this should be an important part of judgmental forecasting research. A complex context will complicate the analysability of the forecasting task in practice. Yet, what makes the organizational context so interesting for judgmental forecasting research, is exactly what makes it so difficult to investigate and, certainly, to generalize from. As Reimers and Harvey (2011) and Harvey and Reimers (2013) conclude from their experimental, non-laboratory studies: specific organizations use specific data series. Because of their previous experience (ecological knowledge) people assumed that time series were positively auto-correlated and were unable to regress to the mean. Harvey and Reimers (2013) suggest that participants engaged in damping (or anti-damping) in a single series, depending on their expectation of the curve compared to what they experience in their environment. If the curve in the single series was steeper than they expected, they engaged in damping. If it was more shallow, they engaged in anti-damping. This suggests that people base their expectations for forecasting on experience with the broader environment, and not just the time series. This means that in an organization, experts have experience with a specific type of data series and a specific type of context. Field studies have to be careful in drawing generalizable conclusions. An experimental design on the other hand, using 'neutral' participants, enables us to investigate the 'pure' effects of judgmental forecasting, without the interference of a variable context. Using artificial series, for instance, provide certainty about the statistical underlying model and the optimal forecast.

An often heard criticism pertains to the use of student participants. Participants are essentially dealing with tasks that are unfamiliar to them. However, novices and experts are subject to the same heuristics and biases (Kahneman, 1991). Northcraft and Neale (1987), for instance, found that both novices and experts were prone to the anchor-and-adjust-heuristic in a study with real estate property. An identical susceptibility to cognitive biases regardless of expertise has been found in a variety of domains (e.g., Leventhal, Teasley, & Rohlman, 1994; McNeil, Pauker, Sox, & Tversky, 1982; Waylen, Horswill, Alexander, & McKenna, 2004). Even an inverse expertise effect has been found (Önkal & Muradoğlu, 1994; Yates, et al., 1991). Bolger & Wright (1994) document twenty studies that compare expert with novice decision making. Only 6 out of 20 studies showed better performance in experts than in novices, 9 showed worse performance (inverse expertise effect) and the remaining 5 showed equal performance. Their main conclusion is that, when there is a weak or absent feedback loop, expert judgment is not likely to outperform judgment of non-experts.

With regard to Chapter 2, the main guideline for further research would be the replication of this study in other types of environments. It is necessary that we know where our approach works and where it doesn’t. There is room for further optimization of the system: it tended to dampen the positive effects of judgmental adjustment. Interviews about the acceptability of this type of decision support system could also be relevant. With regard to Chapter 3 and 4, the blocked design as recommended above seems crucial in further understanding the effects of providing a statistical forecast to a judgmental forecaster, in time series with perturbances, without task switching. Additionally, it would be interesting to see what happens when we provide feedback. Benson and Önkal (1992) distinguish between four types of forecasting feedback: (1) outcome feedback (providing the latest outcome), (2) performance feedback (giving the forecaster information on his or her accuracy), (3) cognitive process feedback (informing the forecaster about his or her forecasting strategy) and (4) task properties feedback (giving the forecaster statistical information about the task). According to Fischer and Harvey (1999), people only learn from outcome feedback when the task is fairly simple, i.e. two or three cues maximum and a linear relation to the criterion. When the task is more complex, cognitive feedback is more useful (summarizing the relation between responses and criterion values/values of the various cues). A similar study to ours could potentially benefit from outcome feedback. This method should be easily applicable in practice to enhance forecasting accuracy.

Ideally, further judgmental forecasting research will focus on both approaches used in this doctoral thesis : fundamental experimental research to test the basic relations between judgment and models, how people make decisions, how they react; and research in business practice, as we need to test our conclusions on behaviour in companies. Ideally, a direct line of communication should exist between research and practice. Practice should employ the latest findings of basic, fundamental research, in order to improve forecasting performance. Simultaneously, researchers should acknowledge the ways of practice, including limitations and habits, and take these into account when designing their studies.

6. References

Adelman, L. (1981). The influence of formal, substantive, and contextual task properties on the relative effectiveness of different forms of feedback in multiple cue probability learning tasks. *Organizational Behavior & Human Decision Processes, 27*, 423 - 442.

Adya, M., Collopy, F., Armstrong, J. S., & Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting, 17*(2), 143 - 157.

Armstrong, J. S., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. In G. Wright & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 269 - 293). New York: John Wiley & Sons.

Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. *Information Systems Journal, 16*(1), 55 - 78.

Bates, J. M., & Granger, C. J. (1969). The combination of forecasts. *Journal of the Operational Research Society, 20*(4), 451 - 468.

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

Bolger, F., & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems, 11*, 1 - 24.

Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. Cambridge, Mass., USA: MIT Press.

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*(6), 570-576.

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 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*(1), 3 - 23.

Franses, P. H. (2013). Improving judgmental adjustment of model-based forecasts. *Mathematics and computers in simulation, 93*, 1 - 8.

Franses, P. H., & Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting, 25*(1), 35 - 47.

Giullian, M. A., Odom, M. D., & Totaro, M. W. (2000). Developing essential skills for success in the business world: a look at forecasting. *The Journal of Applied Business Research, 16*(3), 51 - 61.

Goodwin, P. (2002). Integrating management judgment and statistical methods to improve short-term forecasts. *Omega 30, 30*(2), 127 - 135.

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*(1), 37 - 23.

Goodwin, P., Fildes, R., Lawrence, M., & Nikolopoulos, K. (2007). The process of using a forecasting support system. *International Journal of Forecasting, 23*(3), 391 - 404.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega : The International Journal of Management Science, 39*(3), 242 - 253.

Hammond, K. R. (1981). Principles of Organization in Intuitive and Analytical Cognition. Boulder, CO: Center for Research on Judgement and Policy, University of Colorado.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinman, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271 - 312). New York: Academic Press.

Harvey, N., & Reimers, S. (2013). Trend damping: under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology, 39*(2), 589 - 607.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*(5), 733 - 763.

Hendry, D. F., & Clements, M. P. (2003). Economic forecasting: some lessons from recent research. *Economic Modelling, 20*, 301 - 329.

Hoch, S. J., & Schkade, D. A. (1996). A Psychological Approach to Decision Support Systems. *Management Science, 42*(1), 51 - 64.

Hughes, M. C. (2001). Forecasting practice: organisational issues. *Journal of the Operational Research Society, 52*, 143 - 149.

Kahneman, D. (1991). Judgment and decision making: A personal view. *Psychological Science, 2*, 142 - 145.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in Judgment: Comparing Individuals and Groups. *Psychological Review, 103*(4), 687 - 719.

Kremer, M., Siemsen, E., & Thomas, D. J. (2015). The Sum and Its Parts: Judgmental Hierarchical Forecasting. *Management Science, Published Online: December 18, 2015*, 2745 - 2764.

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

Leventhal, L., Teasley, B., & Rohlman, D. (1994). Analyses of factors related to positive test bias in software testing. *International Journal of Human-Computer Studies, 41*, 717 - 749.

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

Makridakis, S., & Gaba, A. (1998). Judgment: its role and value for strategy. In G. Wright & P. Goodwin (Eds.), *Forecasting with judgment* (pp. 1 - 38). Chichester: John Wiley & Sons.

Mathews, B., & Diamantopolous, A. (1986). Managerial intervention in forecasting: an empirical investigation of forecast manipulation. *International Journal of Research in Marketing, 3*, 3 - 10.

Mathews, B., & Diamantopolous, A. (1989). Judgemental revision of sales forecasts: a longitudinal extension. *Journal of Forecasting, 8*, 129 - 140.

Mathews, B., & Diamantopolous, A. (1990). Judgemental revision of sales forecasts: effectiveness of forecast selection. *Journal of Forecasting, 9*, 407 - 415.

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

McNeil, B. J., Pauker, S. J., Sox, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine, 306*, 1259 - 1262.

Monsell, S. (2003). Task switching. *Trends in Cognitive Science, 7*(3), 134-140.

Northcraft, G. B., & Neale, M. A. (1987). Expert, amateurs, and real estate: An anchoring and-adjustment perspective on property pricing decisions. *Organizational Behavior & Human Decision Processes, 39*, 84 - 97.

Önkal, D., Goodwin, P., Thomson, M., Gönul, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*, 390 - 409.

Önkal, D., & Muradoğlu. (1994). Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research, 74*(2), 350 - 358.

Oreg, S., & Bayazit, M. (2009). Prone to Bias: Development of a Bias Taxonomy From an Individual Differences Perspective. *Review of General Psychology, 13*(3), 175 - 193.

Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting, 27*, 1196 - 1214.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantiative forecasting methods in practice. *Omega, 31*, 511 - 522.

Shim, J. K. (2000). *Strategic business forecasting: the complete guide to forecasting real world company performance*. New York, Washington DC: St. Lucie Press.

Silver, M. S. (1991). Decisional Guidance for Computer-Based Decision Support. *MIS Quarterly, 15*(1), 105 - 122.

Sniezek, J. (1986). The role of variable labels in cue probability learning tasks. *Organizational Behavior & Human Decision Processes, 38*, 141 - 161.

Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively Categorical Variables: Revealing the Information in Zip Codes. *Marketing Science, 22*(1), 40 - 57.

Syntetos, A., Babai, Z., Boylan, J., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*, 1-26.

Syntetos, A., Georgantzas, N. C., Boylan, J. E., & Dangerfield, B. C. (2011). Judgement and supply chain dynamics. *Journal of the Operational Research Society, 62*, 1138 - 1158.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

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*(2), 234 - 243.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124 - 1131.

Waylen, A. E., Horswill, M. S., Alexander, J. L., & McKenna, F. P. (2004). Do expert drivers have a reduced illusion of superiority? *Transportation Research Part F – Traffic Psychology and Behavior, 7*, 323 - 331.

Webby, R., & O’Connor, M. (1996). Judgmental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting, 12*, 91 - 118.

Wright, G., & Goodwin, P. (1998). *Forecasting with judgment*. Chichester: John Wiley & Sons.

Yates, J. F., McDaniel, L. S., & Brown, E. S. (1991). Probabilistic forecasts of stock prices and earnings: The hazards of nascent expertise. *Organizational Behavior & Human Decision Processes, 40*, 60 - 79.