Judgmental Forecasting from Graphs and from Experience

Zoi Theochari

Cognitive, Perceptual and Brain Sciences Research Department

University College London

Thesis submitted for the degree of

Doctor of Philosophy (PhD)

September 2013

Abstract

Research in the field of forecasting suggests that judgmental forecasts are typically subject to a number of biases. These biases may be related to the statistical characteristics of the data series, or to the characteristics of the forecasting task. Here, a number of understudied forecasting paradigms have been investigated and these revealed interesting ways of improving forecasting performance. In a series of experiments, by controlling parameters such as the horizon and direction of the forecasts or the length, scale and presentation format of the series, I demonstrate that forecasting can be enhanced in several ways.

In Chapter 3, I examine forecasting direction as well as the use of an end-anchor to the forecasting task (Experimental Studies 1-2). In Chapter 4, I examine the way the length of the series affects forecasting performance of various types of time series (Experimental Studies 3-4). Dimensional issues related to the forecasting task are further investigated in Chapter 5, where graphs' scale is now manipulated in series with different types of noise (Experimental Studies 5-6). Task characteristics are further explored in dynamic settings in Chapter 6, in a number of experiments (Experimental Studies 7-12), where a new experimental paradigm for judgmental forecasting is introduced. Here, I test already identified robust forecasting biases in this dynamic setting and compare their magnitude and direction with those found in static environments.

I conclude that forecasting performance is affected by data series' and task characteristics in the following ways i) end-anchoring and backwards direction in forecasting tasks enhance accuracy ii) longer lengths are preferable for a number of series' types iii) dynamic settings may offer specific enhancements to the forecasting task.

The implications of these findings are discussed with respect to judgmental forecasting and corresponding cognitive mechanisms, while, directions for future research, towards the development of a unified framework for judgmental forecasting, are suggested.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

Signature:

September, 2013

(Zoi Theochari)

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Nigel Harvey for his invaluable guidance, his immense knowledge, inspiration and support of my PhD. He showed me how to draw the big picture of the Psychology discipline in applied settings like forecasting and guided me in the right direction. Besides my supervisor, I would also like to thank Nick Chater, Robin Hogarth, Stian Reimers, Aris Syntetos, Konstantinos Tsetsos, Matt Twynman and Maarten Speekenbrink for their interest and insightful discussions. During the course of this work, I was supported in part by Decision Technology, UCL, the Psychonomic Society (EPS) and the International Institute of Forecasters (IIF). I would like to acknowledge this support, which allowed me to continue my studies. Last but not least, I would like to thank my parents, Themis and Anthi. This work would not have been completed without their support. This thesis is dedicated to them.

Publications arising directly from this submission

Theocharis Z. & Harvey N. (2013). Order effects in judgmental forecasting: Higher accuracy when the last outcome is forecast first. International Journal of Forecasting. *Submitted manuscript (2nd round review)*

Theocharis Z. & Harvey N. (2013). Judgmental forecasting: How does accuracy depend on the amount of data available? *Journal of Experimental Psychology (Applied). Submitted manuscript (under review)*

Theocharis Z. & Harvey N. (2013). Judgmental forecasting from real-time experience. *Manuscript in preparation, to be submitted to the International Journal of Forecasting.*

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

Background

Forecasting with the use of human judgment or human experience is pervasive across both the operational and cognitive science domains. Typically, an estimation of future values is produced by an observer based on a sequence of past data values, known as time series. Time series can be presented to the observer either in the form of graphs and tables (static presentation) or in the form of a stream of values appearing over time (dynamic presentation).

These types of forecasting tasks have been the subject of independent research streams, within different paradigms and disciplines. Forecasting tasks with statically presented data have been traditionally studied within the decision sciences under operational and business paradigms, by assessing the forecasting accuracy in simple or group forecasting tasks (Harvey and Reimers 2013; Goodwin, Önkal and Thomson, 2010; Önkal, Sayım, and Lawrence, 2012; Pollock, Macaulay, Önkal-Atay and Wilkie-Thompson, 1999; Reimers and Harvey, 2011; for relevant reviews see Lawrence, Goodwin, O'Connor and Önkal, 2006; Goodwin and Wright, 1993; Webby and O'Connor, 1996; Armstrong and Collopy, 1998; Syntetos, Boylan and Disney, 2009; Leitner and Leopold-Wildburger, 2011 and for relevant competitions see Makridakis et al., 1993; Makridakis and Hibon, 2000). On the other hand, experiential tasks, where a stream of stimuli is presented dynamically to the observer, have been mainly studied within the cognitive sciences and, more specifically, within low-

level cognitive science studies (e.g. Tsetsos, Usher and McClelland, 2011; Wong, Huk, Shadlen, and Wang, 2007; Ratcliff, 2006; Usher and McClelland, 2001; Busemeyer and Townsend, 1993). In contrast to studies using static presentation, few of those using dynamic presentation have included sequential dependencies between successive outcomes (see, for example, Gureckis and Love, 2010; Boyer, Destrebecqz, and Cleeremans, 2005). This makes comparing the findings difficult.

Research in these different areas has produced certain suggestions regarding the performance and competence of the forecaster. With static presentation, the forecaster's responses are compared against the actual or the optimal future values of the time series. In these types of tasks, performance is also often assessed against the naïve forecasting benchmark (Lawrence, O'Connor and Edmundson, 2000; Makridakis et al., 1993; Lim and O'Connor, 1995; Sanders, 1992): this represents the accuracy achieved when the forecaster uses the last data point of the time series as a forecast. The usefulness of the naïve benchmark is not only relevant to the accuracy of the forecaster, but also to the underlying process of forecasting. This is because decisions in sequential settings can often be seen as being governed by an anchoring and adjustment heuristic (Harvey 2007; Lawrence and O'Connor, 1995; Andreassen and Kraus, 1990; Lawrence and O'Connor, 1992; Bolger and Harvey, 1993; Hogarth and Makridakis, 1981; Harvey, in press). The anchor is in most cases the last data point and adjustment is based on the patterns perceived in the data. Similarly, in the aforementioned dynamic paradigms, there is evidence that forecasts are

primarily based on the last set of observations, thereby producing recency effects; in such tasks, long-term patterns are difficult to detect and assimilate into the observer's judgment (Tsetsos, Chater and Usher, 2012).

The fact that optimal forecasts can be produced by using the series' patterns and signals to make adjustments away from the naïve benchmark, provides a useful way of understanding the underlying cognitive forecasting processes (Harvey, 2007; Goodwin and Wright, 1994; Hogarth, 1981; Speekenbrink, Twyman and Harvey, 2012; Bromiley, 1987). These forecasting processes appear to show biases, similar to those found in many judgment tasks (e.g., Lichtenstein and Slovic, 1971; Tversky and Kahneman, 1974; Gilovich, Griffin, and Kahneman, 2002). In forecasting settings, several biases have been found to be robust, with the trend damping, the autocorrelation illusion and noise introduction being the most dominant ones (Harvey and Reimers, 2013; Reimers and Harvey, 2011; Harvey, 1995). Other biases are dependant on the forecasting task characteristics (e.g feedback and advice assimilation, sensitivity to asymmetric loss, causal information incorporation).

In addition to effects arising from use of anchoring and adjustment heuristics, a number of other behavioural regularities have been found to operate when forecasts are generated by individuals. This has given rise to the suggestion that, as in other judgment tasks, forecasters use a set of diverse heuristics, each describing a special characteristic of the forecasting behaviour (Todd and Gigerenzer, 2000; Gigerenzer, 2006; Harvey, 2007). Some procedures have been suggested to reduce forecasting biases (for a review see Lawrence et al.

2006). The most dominant ones are related to structured approaches to judgment such as decomposition, role-playing, group-forecasting, feedback exploitation and the implementation of the Delphi technique (MacGregor, 2001; Armstrong, 2001; Sniezek, 1989; Mackinnon and Wearing, 1991; Rowe and Wright, 2001). In this thesis, some additional methods by which judgmental forecasting can be improved are explored.

In forecasting tasks, difficulties in generalising findings are encountered because the heuristics at play are found to be sensitive to the characteristics of the time series presented (Bolger and Harvey, 1993; Goodwin and Wright 1993). For example, in the case of the anchoring and adjustment heuristic, the forecaster employs different versions depending on the series at hand (Lawrence and O'Connor, 1992). These findings are mainly associated with research using static presentation. On the other hand, experiential forecasting processes are understudied and scarce (Wagenaar and Timmers, 1979; Hogarth, 1981; Remus and Kottemann, 1987, 1995).

Contrary to judgmental forecasting from graphs whose fundamental components have been analysed for at least 30 years (Lawrence et al, 2006), little is known about the mechanisms underlying information assimilation and use in forecasting tasks where the participant is experiencing a time-series in real-time instead of observing static graphs. In this thesis, real-time, high-frequency experiential tasks will be explored. However, these tasks should not be confused with forecasting tasks where domain experts use their professional experience to extrapolate from time series. Although forecasting tasks from

graphs and from real-time high frequency experience (i.e. experiential tasks) share a common conceptual framework, it should not be assumed *a priori* that these tasks share common cognitive processes. Revealing the processes that control judgmental forecasting from graphs and from experience and understanding whether and how they differ is central to our broader understanding of judgmental forecasting and, thus, its improvement.

The aim of this thesis is to refine the available knowledge on forecasting biases in various data series with different statistical characteristics and presentation formats. Providing grounds of understanding judgmental forecasting will not only enhance our understanding of how people perform this task but may also help us to arrive at a more thorough understanding of how anticipation of the future works in humans, and thereby affect a number of other related areas of research, such as affective forecasting, intertemporal choice and optimism (Wilson and Gilbert, 2003; Loewenstein, Read, and Baumeister, 2003; Weinstein, 1980). Findings from the research reported here contribute in both applied research areas (e.g. improving judgmental forecasting in the financial and business world) as well as more theoretical academic disciplines (e.g. cognitive science).

Characteristics of the forecasting process

First, I will define some of the basic characteristics and the general framework of the judgmental forecasting process. Later, I will refine this analysis by

describing potential cognitive processes involved in judgmental forecasting. Judgmental forecasting is characterized by a person's immediate (partly intuitive) response to a given time series of events or data points. This is akin to an unstructured judgmental impression. In particular, judgmental extrapolation, which is the main process of interest here, is defined as the subjective extension of time-series data, according to Armstrong's forecasting dictionary (Armstrong, 2001). Judgmental forecasts can be produced either by domain experts, who use both their domain knowledge as well as the historical data to come up with a final estimation (see for example Glaser, Langer and Weber, 2007), or by lay people who make predictions solely on the basis of the given series of historical data. Like the majority of studies in judgmental forecasting, the experiments reported in this thesis were conducted with participants who had no domain knowledge. Surprisingly, judgmental forecasts with or without domain knowledge have been found to be as accurate as statistical models in some studies (e.g. Lawrence, Edmundson and O'Connor 1985, 1986). This is attributed to the fact that participants might be able to pick-up patterns that are missed by the formal statistical techniques. This is not always the case though (e.g. Bunn and Wright, 1991).

Judgment can also be used to make adjustments to formal forecasts (Sanders and Ritzman, 2001; Önkal and Gönül, 2005). In these cases, which are widespread in professional environments, the forecaster makes a subjective change to a statistical forecast. For example, supply chain managers use this process frequently to try to improve predictions of future demands (Fildes,

Goodwin, Lawrence and Nikolopoulos, 2009) by incorporating their knowledge of the environment, of the product and their past experience. Revealing the structure of errors produced by judgmental adjustments of statistical forecasts has been a topic of interest in recent years (e.g. Syntetos et al. 2009). This research has involved both laboratory (Goodwin and Fildes, 1999; Lim and O'Connor, 1995) and field studies (Fildes et al., 2009; Mathews and Diamantopoulos, 1986, 1989, 1990, 1992). The evidence suggests that in some cases statistical forecasts can be improved via judgmental adjustments, especially when important domain knowledge, which is not available to the model, is incorporated in the forecasts (Goodwin, Fildes, Lawrence and Nikolopoulos, 2007; Sanders and Ritzman, 2001; Turner, 1990). In other cases, particularly when this kind of important piece of information is not available, adjustments are found to damage accuracy (Fildes et al., 2009). The investigation of errors produced by judgmental adjustments involves mainly the magnitude and the direction of those errors. Fildes et al. (2009) suggest that large and negative adjustments are likely to lead to greater accuracy, whereas smaller and positive adjustments are likely to impair accuracy. Positive adjustments are often attributed to an optimism bias (Weinstein, 1980) and are responsible for severe and consistent damaging of forecasts (Flides et al, 2009; Mathews and Diamantopoulos, 1989). Some interventions, aimed at the removal of consistent biases like this one, have been proposed for practitioners. Goodwin (2000), for example, suggested that prompting the forecaster to indicate a reason for making an adjustment reduced the frequency of unnecessary adjustments. Also, Fildes et al. (2009) discussed several

approaches to improve adjustments but highlighted the fact that methods like automatic correction of forecasts would face obstacles in practice. In this thesis, there are no experiments where judgmental adjustments are requested by the forecaster. Nevertheless, findings from Chapters 3, 4 and 5 may prove quite useful when forecasters are requested to produce an adjustment to a formal statistical forecast.

Judgmental forecasting in practice

Judgmental forecasting plays an essential role in business planning and in many other areas of life. Judgment is now considered important in a variety of forecasting tasks ranging from company sales forecasting to macro-economic forecasting (Batchelor and Dua, 1990; Clements, 1995; Fildes and Stekler, 2002; McNees, 1990; Turner, 1990; Goodwin, Önkal and Lawrence, 2011), so much so that corresponding research communities have effectively used it to develop practice guidelines as well as scientific consensus statements (e.g. Armstrong, Green and Graefe, 2013; Sanders and Manrodt, 1994; Armstrong and Collopy, 1998; Armstrong, 2001). Clearly, judgmental input is thought to have an important bearing on several important real-world forecasting issues. In this section of Chapter 1, I will review some of the most important areas where applications of forecasting with the use of judgment are being practiced.

The most important area, which first called for researchers' input has been business forecasting. Although development of formal methods of forecasting

continues apace up until now, many surveys have shown that most forecasting within businesses is associated with judgmental input (e.g., Mentzer and Cox, 1984; Mentzer and Kahn, 1997; Sanders and Manrodt, 1994, 2003; Sparkes and McHugh, 1984). Moreover, adoption of formal techniques has been shown to have reached an asymptote (Lawrence, 2000) and this has been the catalyst for early judgmental forecasting studies. Particularly, it was Lawrence, O'Connor and Edmundson's (1985, 1986) large-scale comparative studies, which first provided evidence that judgmental forecasting could be proven more accurate than quantitative models forecasting. Their studies were published after the first forecasting competition, the M1-competition (Makridakis et al., 1982); M1 compared the accuracy of most of the widely available forecasting models from a variety of domains including stock market, sales, demographic and finance. While large-scale surveys and experimental studies continued being conducted in the field of judgmental forecasting, additional large-scale competitions were launched; M2 and M3 competitions followed M1 (Makridakis et al, 1993; Makridakis and Hibon, 2000) and judgmental forecasting was now acknowledged as an official method of producing or improving forecasts.

The essential role of judgmental forecasting can be understood when one thinks of its implications in supply chain management. A persistent theme in the literature is the extent to which judgment can make a difference to sales forecasts where collaborative planning is of paramount importance; there, forecasts are produced by Forecasting Support Systems, which nowadays can integrate statistical output with judgmental input from experts in the

organization. In the specific case of sales forecasting, judgmental input is represented by the information added by experts. This is information, which cannot be provided by statistical models; moreover, it can change the final forecasting outcome drastically; experts are capable of adding information related to future promotions, aggressive marketing circumstances, increasingly competitive markets, shortening of product life cycles and other important variables to adjust the statistical forecast. These important factors are not available to the statistical model and have been proved capable of improving substantially formal forecasts (e.g. Goodwin and Fildes, 1999; Mathews and Diamantopoulos, 1992).

In addition to sales forecasting, there have been considerable developments to develop judgmental approaches in finance; portfolio managers, investors and traders task is to forecast future values of stock prices, bonds and predict market movements to take their investment decisions. Judgmental forecasting is extremely relevant to these settings and has also been studied extensively (e.g. Muradoglu and Önkal, 1994; Önkal, 1998; Goodwin, Önkal-Atay, Thompson, Pollock and Macauley, 2004; Önkal and Muradoglu, 1994, 1996).

It has now been acknowledged that judgmental forecasting can play an important role with business settings and that research into judgmental forecasting has real potential for increasing business effectiveness (Syntetos, Nikolopoulos, Boylan, Fildes and Goodwin, 2009). Approximately thirty years after the first official attempts to reconcile the mechanisms underlying judgmental input to forecasts, there is now a large corpus of such research

(Lawrence et al, 2006) and findings have been used to develop principles of good practice (Armstrong, 2001). Since the early steps in the field, many approaches to forecasting with judgment have been, and are being, developed. This is achieved mostly by using experiments that involve presenting participants with series of stimuli to be judged and interpreting the underlying processes accordingly. Apart from accuracy measures, emphasis is given to the way people think when anticipating the future. In other words, a large set of studies operates in the interface of business and cognitive science research (see for example Harvey and Reimers, 2013; Reimers and Harvey, 2011; Harvey, 2007). In other words, judgmental forecasting has been proven useful for identifying the underlying cognitive mechanisms that govern human anticipation of the future.

Interestingly, there are several approaches where researchers conducted forecasting research in the interface of cognitive science and environmental science. Those were mainly related to the field of climate forecasting. The main goal of the research was to reveal how people react when presented with graphical information related to future and past values of climate variables. So, for example, Lewandowsky (2011) demonstrated that lay people, who use their judgment, put emphasis on long-term climate trends and ignore local information when extrapolating data. This finding has significant implications because it suggests that presentation of climate data would counteract evidence that global warming has stopped. However, the notion that cognitive biases are playing an important role, when climate data is presented to people, is not new.

A review paper of the Bulletin of the American Meteorological Society (see Nicholls, 1999), discusses the role of cognitive mechanisms in climate predictions. It shows that users have difficulties in understanding and translating probabilistic information in operational settings. Also Lemos, Finan, Fox, Nelson and Tucker (2002) presented research from Brazil, where forecasters tested various information formats of geoclimatic maps in order to discover the best approach for users.

Short-term weather forecasts are also of interest in a variety of domains. For example, daily temperatures were studied in conjunction with weekly and daily sales in a Brewing company (Nikolopoulos and Fildes, 2013), in order to estimate the impact of temperature fluctuations in the company sales. The forecasters were found to take advantage of weather information by adjusting their sales forecasts according to short-term weather predictions.

The optimal use of weather and climate data and predictions is also of interest for several practitioners as well as policy makers. A relevant example involves a forest management application in British Columbia lands, where flexible policies had to be produced on the basis of climate information (McDaniels, Mills, Gregory and Ohlson, 2012). There, forecast scenarios from a panel of 14 experts were combined to produce stochastically flexible policies. It is not only environmental policies that can take advantage of judgmental input. Policy makers have to forecast the impact of future legislation by using several sources of information before applying them. In this context, an interesting forecasting system with judgmental input was recently proposed by Savio and

Nikolopoulos (2009) to facilitate policy-making at country level; the aim was to provide forecasts of policies' success before implementation.

A large variety of contemporary domains of judgmental forecasting practice have evolved over the past decades; one can find applications in virtually all domains of economic activity, where insights and knowledge of human experts provide essential aids to the forecasting process. Thus, practical judgmental tools are used in econometric forecasting (Allen and Fildes, 2001), in macroeconomic forecasting (McNees, 1990), in real estate market forecasting (Ong and Chew, 1996), web-tourism demand (Song, Witt and Zhang, 2008), livestock production (Vere and Griffiths, 1995), as well as sports forecasting (Andersson, Edman and Ekman, 2005). The number of these different areas and the rate at which they are developing provide considerable scope for forecasting researchers and cognitive scientists.

There are, however, limitations to the use of judgmental forecasting that still require a broad-based exploratory research in order to be overcome. If judgmental forecasting tools are to improve forecasting performance, research needs to be systematic and provide practical guidance (e.g. Armstrong et al., 2013). A critical point is to clearly define judgmental biases in all areas of application and to develop corresponding tools that provide improvements (e.g. Goodwin, 2000; Fildes et al., 2009; Syntetos et al., 2009; Bunn and Wright, 1991). In this thesis, understudied areas of judgmental forecasting will be examined so that practitioners can be provided with specific strategies to aid their performance.

1.1 Judgmental forecasting basic mechanisms

Research in the field of judgmental forecasting has shown that, when people use their judgment to forecast future values of various financial or environmental variables, they exhibit several biases that impair their forecasts (Harvey, 2007; Eroglu and Croxton, 2010). Research suggests that many of these biases arise from the simple heuristic mechanisms that people use in their attempts to take into account the pattern, autocorrelation and noise information in the series.

Among these simple heuristic mechanisms, the anchor and adjustment heuristic (Tversky and Kahneman, 1974) has been proposed as a way of explaining some of the biases in judgmental forecasting tasks (Harvey, 2007). The anchor point is usually defined as the last data point provided to the judge or the long-term average of the series, and adjustment is based on other elements of the time series. For example, Andreassen and Kraus (1990), as well as Lawrence and O'Connor (1992), used this heuristic to model people's judgment performance by specifying those anchors and arguing that adjustment comprises a proportion of the difference between the two most recent data points multiplied by a parameter which was dependent on the series' characteristics. In another experiment, Bolger and Harvey (1993) found that people employed different versions of this heuristic for trended and untrended series. For trended series the anchor was the last point but the adjustment was towards the trend of the series. For untrended series, the anchor was the last data point and adjustment was towards the mean of the series. For cyclical series, people anchor again on the

last data point and adjust by taking into account a proportion of the last difference in the data (Harvey, Bolger and McClelland, 1994).

Adjustment away from the anchor is typically insufficient when this heuristic is used, thereby producing biases in forecasts (Epley and Gilovich, 2001). A well-documented bias related to insufficient adjustment is trend damping: people's forecasts lie below upward trends and above downward ones. (Wagenaar and Sagaria, 1975; Eggleton, 1982; Bolger and Harvey, 1993; Harvey et al., 1994; Lawrence and Makridakis, 1989; Sanders, 1992). Another bias is the positive autocorrelation illusion: people's forecasts imply serial dependence even when the series are independent (Reimers and Harvey, 2011; Bolger and Harvey, 1993; Eggleton, 1982). This also means that forecasts often lie closer to the last data point than they should.

In most cases the anchoring and adjustment heuristic is modelled using exponential smoothing algorithms with the amount of the adjustment depending on the latest error value and the value of the smoothing constant (Andreassen and Kraus, 1990; Lawrence and O'Connor 1992, 1995). A time series is a sequence of events related one to each other in various ways. Thus, the effects of the use of the anchoring and adjustment heuristic by participants will be examined in all the experiments.

Another set of cognitive biases associated with the anchor and adjustment heuristic, are recency and primacy biases. Recency effects (Anderson, 1981) occur when people evaluating a sequence of items are unduly influenced by those received later in the sequence. In other words, later data dominate a

decision makers' judgment. Recency has important implications in judgmental forecasting from graphs. Especially in the case of graphical rather than tabular representations, it is easier for the participant to focus on and overweight the most recent data points and their patterns. Thus, when Lawrence and O'Connor (1992) investigated the influence of the slope of the last segment of an ARMA time series, they found that "the judgemental forecaster, on average, utilises the segment slope information correctly in judging the direction of adjustment but incorrectly estimates the amount of adjustment". Primacy effects refer to the influence of items early in a sequence: they are most likely to be at play in judgmental forecasting from experience (e.g. Tsetsos et al., 2011). In the experiments outlined in this thesis, I use a variety of time series. In cases of highly autocorrelated or persistent data, the forecaster is justified in applying high weight to recent events and, thus, in exhibiting a conservative anchor and adjustment heuristic strategy. In other cases, where larger shifts are required to provide enhanced accuracy, the forecaster would have to alter this strategy.

Apart from the family of anchoring heuristics, there are other heuristic mechanisms involved when people make forecasts by using their judgment. One such is the representativeness heuristic (Tversky and Kahneman, 1974), which assumes a high degree of correspondence between a sample and a population. As Bagnoli, Guazzini, and Lio (2008, p.2) mention, "this heuristic can be thought of as the reflexive tendency to assess the similarity of characteristics on relatively salient and even superficial features, and then to use these assessments of similarity as a basis of judgment". An example of the

involvement of this heuristic in forecasting was documented by Harvey, Ewart and West (1997) in an investigation on the influence of noise levels in people's predictive accuracy. According to their account, people add noise to their forecasts to make each forecast typical of the data points in the presented series. If past data show more scatter, more noise is included in forecasts to represent that scatter. In the experiments outlined in this thesis, I examine whether there is evidence for such effects in a variety of series types.

Another heuristic involved in judgmental forecasting is the availability heuristic. The use of this heuristic is related to probability or frequency judgments which rely upon the available knowledge. In the case of judgmental forecasting from time series, this heuristic is likely to be involved when people predict data points that belong to large classes of events (for example, values that are close to the average of the time series).

Other sources of error generation that affect forecasting stem from behaviours that reflect optimism. The optimism bias (Weinstein, 1980) potentially explains elevations in people's forecasts. Optimism bias (Weinstein, 1980), desirability bias (Crandall, Solomon, and Kelleway, 1955), overforecasting bias (Eggleton, 1982) and outcome bias (Cohen and Wallsten, 1992) all relate to a judgmental phenomenon in which people overestimate the probability of desirable future events, while also underestimating the probability of undesirable ones (Weinstein, 1980). Reimers and Harvey (2011) argued that this may occur in the forecasting domain because the forecasting scenarios often involve profit or sales scenarios. While optimism seems to affect forecasts by elevating the final

estimation, overconfidence causes the forecaster to think that the probability that a forecast is correct is greater than the actual probability. This is likely to explain another robust finding in judgmental forecasting: prediction intervals are estimated to be too narrow (O'Connor and Lawrence, 1989).

1.2 Types of judgmental forecasting phenomena

The aforementioned mechanisms are considered central to judgmental forecasting. Nevertheless, the diversity of time-series, as well as the richness in presentation formats and task characteristics, renders generalisations difficult (Goodwin and Wright, 1993). Broadly speaking the mechanisms outlined in the previous section produce biases, which can be grouped into a) those related to the statistical characteristics of the data series, b) those related to the way in which series are presented to forecasters, and c) those related to characteristics of the forecasting task. In the next sections, I discuss what is known about problems with judgment input to the forecasting process and outline how they relate to the main issues to be investigated in the research reported in this thesis.

1.2.1 Statistical characteristics of the series

Various forecasting anomalies or 'biases' are related to the way that forecasters perceive the statistical characteristics (patterns and noise) in the series. Research has revealed that the forecasters typically produce forecasts that are too close to the last data point. As a result, they appear to underestimate the steepness of trends in series (Harvey and Reimers, 2013) and to overestimate

first-order sequential dependence (Reimers and Harvey, 2011). Second, they may add noise to a sequence of forecasts that reflects the level of noise in the data series (Harvey, 1995). This may be because they use the representativeness heuristic or because they see patterns in the noise where none exist (O'Connor, Remus and Griggs, 1993). Third, forecasts may be influenced by what forecasters consider to be desirable and by whether they think that the series can be controlled to counteract any undesirable features that may be revealed as the future unfolds (Lawrence and O'Connor, 1992).

The time series mean, noise, autocorrelation, persistency levels and trends form the basic series characteristics I will be concerned with in this section. Webby and O'Connor (1996) list a subset of those factors as important ones in their extensive review of judgmental and statistical time series forecasting. Specifically, they scrutinize the role of trend, seasonality, noise and discontinuities and they conclude that trend and discontinuities impair judgmental forecasts. Here, I will discuss time-series characteristics in light of their work and new evidence from more recent research.

The time series mean or average long-term value is a quantity perceived and taken into account in judgmental forecasts (Andreassen, 1990; Lawrence and O'Connor, 1992; Armstrong and Collopy, 1993; Harvey et al. 1994). Lawrence and O'Connor (1992), who modelled people's statistical judgment when presented with ARMA models, suggested that their behaviour could be simulated as if the long-term mean of the time series was taken as a mental anchor from which people adjusted away to take into account other elements of

the time series. Nevertheless, when there is seasonality in the time series, which is perceived as part of the pattern, errors tend to depart from the mean; Harvey et al., (1994), for example, in their cognitive algebra analysis present different prediction equations for trended, cyclical or untrended time series. Hence, participants take into account the long-term mean in various ways depending on the time series under examination. However, the long term mean seems to be a quantity that is important when judgmental forecasts are produced.

Equally important are the last observations of the time series, which are treated in a special manner by participants when they produce their forecasts. This is realised by use of anchoring and adjustment heuristics. The last observations' strong influence on the decision-maker's judgment is also highlighted in several cognitive studies under the term 'recency' effects (Tsetsos et al., 2012). The last data points provide an anchor from which people adjust to allow for other important elements of the time series to be taken into account (Bolger and Harvey, 1993; Lawrence and O'Connor, 1995). Thus, the influence of the last observations, though present in all judgmental forecasting, depends on the series type. The position of the last data point, of course, depends on the noise in the series. This makes the anchoring and adjustment strategy prone to errors when an unrepresentative noise pattern occurs on these last points (Harvey et al., 1997). In graphical presentations of a time series, the most recent data points and the slope of their line segments are likely to be excessively weighted in the judgemental forecast leading to a bias in the forecast value. To avoid such

effects impacting on the conclusions drawn from this thesis, I use different exemplars for each participant and condition under examination.

Apart from the long-term mean of the series and the last observations, there are a number of elements that influence judgmental forecasts, which are associated with the complexity in the series. Goodwin and Wright (1993) suggest that the complexity of a series includes three components:

- the underlying signal, comprising its seasonality, cycles and trends and response to shocks;
- the level of noise around the signal and
- the stability of the underlying signal

Here, I use this complexity categorisation to structure the next sections but I add more recent evidence to enrich it.

When people are presented with different series' types, signal detection or pattern extraction is considered to play an important role. Research has investigated the question of how well people can identify patterns of various time series types and how they use this information (Andreassen and Kraus 1990; Lawrence and O'Connor, 1992; Bolger and Harvey, 1993; Lawrence and O'Connor, 1995; Lawrence and Makridakis, 1989; Mosteller, Siegel, Trapido, and Youtz, 1981; Edmundson 1990). Sanders (1992) showed that forecasters can incorporate recognition of a signal in their adjustments to extrapolation forecasts. Participants in these experiments made adjustments that led to the improvement of judgmental accuracy when the series had recognizable patterns.

People, therefore, do extract and use some information about the patterns in series. However, as Goodwin and Wright (1993) argue, the variety of different types of time series used in judgmental forecasting tasks leads to difficulties in generalizing findings implicating particular cognitive mechanisms in the prediction process.

However, fairly robust findings regarding pattern recognition are associated with trended and seasonal patterned series. Trend 'eyeballing' skills were first studied by Lawrence and Makridakis (1989) and Mosteller et al. (1981). These studies showed that people are relatively good at perceiving a trend. However, Andreassen and Kraus (1990) found that noise had an impact to the participants' ability to detect the trend. Subsequent research showed that trend detection skill is insufficient to allow people produce unbiased forecasts regardless of the noise levels: people underestimated or damped both upward and downward trends, with the latter being damped more than the former (Harvey and Reimers, 2013; Bolger and Harvey, 1993; Harvey and Bolger, 1996; Harvey et al., 1994; Lawrence and Makridakis, 1989; Sanders, 1992; Eggleton, 1982; Wagenaar and Sagaria, 1975). Interestingly, Harvey et al. (1997) showed that although positive linear trends were recognized more easily than untrended series, forecasting was worse from them. Significant damping has also been identified in forecasts from non-linear trends (Timmers and Wagenaar, 1977).

While unlimited trends are rarely found in the environment, unlimited periodicity is often a property of the real-world time series. According to

Edmundson (1990), people are efficient at perceiving and utilising seasonal patterns. Nevertheless, there is a limit in this cognitive ability. This limit was stressed by Harvey, Bolger and McClelland (1991), as well as by Lawrence and O'Connor (1992), who argued that the presence of high and complex seasonality or a strongly cyclical component impairs people's judgement. Moreover, this ability to perceive the cyclical nature of time series has been found to depend on the series' noise levels. Harvey et al. (1997), for example, report a set of experiments where a sinusoidal signal was overlaid with different trends and noise levels. The lower the noise levels, the easier the participants recognized the signal. Harvey (1988) suggested, that people do acquire some information about the pattern in the series but do not use it as a basis for their forecasts. Instead they appear to use heuristics based on a few salient elements of the data (Bolger and Harvey, 1993; Lawrence and O'Connor, 1992).

Lack of pattern in a time series is rightly seen as more consistent with random generation of the data (Wagenaar 1972). Lopes and Oden (1987), though, pointed out that even random processes occasionally produce highly patterned sequences. Also, Armstrong et al. (2013), for example, argue that participants tend to see patterns where none exist and that they tend to suffer from illusions of control even when the underlying process is purely random. Here a variety of time series with and without patterns are used in the experiments and unique exemplars are shown to each participant to avoid effects related to the false perception of patterns in the noise in the series.

Following Goodwin and Wright's (1993) categorisation, the next important component is noise level. This heavily influences judgment accuracy. Harvey (1995) showed that series noise causes people to add noise to their judgments in their attempt to represent the time series better and that high noise renders the mental anchors discussed above less effective. Payne (1993) also suggested that an increase in data noise may affect the strategy people use to produce their judgments. He anticipated that an increase in noise levels may cause people to switch from a pattern-extraction based statistical cognitive strategy to one based on heuristics, thereby adapting their decision making strategy to the task that they have been given. However, no matter which strategy is selected, noise makes patterns in data series harder to discern and people add noise to their forecast sequences that tend to mask the patterns that would otherwise appear in those sequences in an effort to make their forecast representative of the time series under examination (Harvey, 1995). Noise introduction effects are examined in the experiments presented here, especially in Chapter 3, Experiment 2 and in Chapter 6, where pure noise introduction is studied in an experiential setting. Also, in Chapter 5, noise type effects are studied. Uniform and Gaussian noise terms are tested to determine whether this manipulation has an effect on forecasting performance.

Series autocorrelation should also be included in Goodwin and Wright's (1993) first category. This is a property that expresses the relation of the last data point to the previous one. In a random process, there is no relation between successive data points. On the other hand, there are many other processes,

where a data point is related to earlier value(s) in some way. Reimers and Harvey (2011) showed that people's forecasts are sensitive to autocorrelation in series. In other words, they take into account the relation between successive data points and forecasts are closer to the last data point when series autocorrelation is higher. A positive autocorrelation illusion was also revealed: people's forecasts imply that they overestimate serial dependence for low (including zero) levels of autocorrelation but underestimate it for very high ones.

Apart from noise and first-order autocorrelation elements in the series, long memory components in the series or higher order autocorrelations should also be considered within this category. People appear to be sensitive to these features of time series though their level of sensitivity remains in dispute (Gilden, Schmuckler and Clayton, 1993, Westheimer, 1991). Degree of sensitivity to them is important because they are present in real series. For example, financial and environmental time series contain important long memory components (Koutsoyiannis, 2002; Cont, 2001; Cajueiro, 2008).

Hurst (1951) was the first to have discovered this special behaviour in hydrological and other geophysical time series; this behaviour is known as the "Hurst phenomenon". The generalised Hurst exponent, which governs the generation of such series, is directly associated with the fractal dimension of a time series. Long memory series are characterized by a tendency to contain clusters of neighbouring values. Mandelbrot (1977, p.248) used the term "Joseph effect" for the same behaviour. Since then, the Hurst phenomenon or

Joseph effect has been verified in several environmental variables, such as global mean temperatures (Bloomfield, 1992), indices of the North Atlantic Oscillation (Stephenson, Pavan and Bojariu, 2000), climate change (e.g. Evans, 1996), River Nile flows (Eltahir, 1996), annual streamflow records across the continental United States (Vogel, Tsai and Limbrunner, 1998), and many others. There is also an extensive literature suggesting evidence of long memory in economics fundamentals (Diebold and Rudebusch, 1989) and, therefore, stock returns and volatility (Cajueiro and Tabak, 2008) and a variety of financial assets (Barkoulas and Baum, 1998). Moreover, a number of psychological variables have recently been revealed to possess such fractal properties. They include self-esteem (Delignières, Fortes and Ninot, 2004), mood (Gottschalk, Bauer and Whybrow, 1995), serial reaction time (Gilden, 2009; Van Orden, Holden and Turvey, 2003) and many others (see for example, Madison, 2004).

From a mathematical point of view, several types of models have been proposed to reproduce the Hurst phenomenon when generating synthetic time series. These include Fractional Gaussian noise (FGN) models (Mandelbrot and Wallis, 1969), Fast Fractional Gaussian noise models (Mandelbrot, 1971), Fractional autoregressive integrated moving-average models (Hosking, 1981), and symmetric moving average models based on a generalized autocovariance structure (Koutsoyiannis, 2000). For the scope of this thesis, and specifically in Chapter 4, I use an approximation to fractional Gaussian noise, the multiple time-scale fluctuation approach (Koutsoyiannis, 2002). This approach was

selected because it provides a very good approximation, which can be tuned to be as accurate as demanded. Additionally, it is not a black box method; this means that the experimenter is able to control the characteristics of the elements of the fractal series and its internal structure. There have been no studies of judgmental forecasting studies from long memory series, which hold their autocorrelation structure for many time steps: this is one reason why I included them in Chapter 4 of this thesis.

Research has shown that forecasters tend to take into account the series statistical characteristics in various ways, depending on the time series under examination, by using context sensitive strategies (Bolger and Harvey 1993). This means, that in order to obtain generalizable results concerning factors that serve to enhance judgmental forecasting performance (Goodwin and Wright 1993), more than one type of series should be used in experiments. This was done here: experiments were conducted with a variety of series' types in order to produce generalizable results regarding accuracy and underlying cognitive processes.

1.2.2 Presentation format

Presentation format (static versus dynamic) and graph format (e.g. points, lines, bars) influence the forecasts that people provide as well. For example, people are generally better at extrapolating from trends when data are presented in graphs (Harvey and Bolger, 1996) and forecasts from graphs are better when

data are represented as points (whether or not they are joined by lines) than when they are represented as bars (Harvey and Reimers, 2012).

Scale of graphs used to represent the data series may also influence quality of forecasts (Lawrence and O'Connor, 1992). Number of forecasts that are made from a given data series and the order in which they are made also appears to have an effect (Harvey, et al., 1997). Length of data series has also been found to affect the quality of judgmental forecasts (Andersson, Gärling, Hedesström and Biel, 2012; Lawrence and O'Connor, 1992).

Presentation format decisions are taken in many business settings such as the stock market and supply chain management as well as other managerial activities. However, this group of important factors is critically understudied. Findings that do exist suggest that it would be useful to carry out more work on how presentation format affects judgmental forecasts. In this thesis, presentation format elements are studied in depth: series' scale, series' length and horizon length variables are scrutinized in Chapters 3, 4 and 5, while, dynamic presentation of stimuli is analysed in Chapter 6.

1.2.3 Task characteristics

Characteristics of the forecasting task beyond the statistical features of the data series and the way it is presented can also influence the quality of forecasts made from it. First, feedback to forecasters about the outcomes they have previously forecast and about the quality of their performance provides a means of training forecasters (Goodwin et al., 2004; Mackinnon and Wearing, 1991;

Remus, O'Connor and Griggs, 1996; Sanders, 1997) but its effectiveness is likely to depend on the delay in providing it, frequency of provision, and various other factors (Harvey, 2011). Second, forecasters' sensitivity to asymmetric loss functions has also been shown to be influencing the forecasting process (Goodwin, 2005; Lawrence and O'Connor, 2005). Third, forecasters have difficulty in incorporating into their forecasts information about causal factors that are likely to perturb the pattern in a time series (Goodwin and Fildes, 1999; Lim and O'Connor, 1996). Fourth, it is well known that errors in aggregated forecasts from a number of independent individuals are lower than average errors of the individuals because of cancellation of random error. However, errors in forecasts produced by interacting groups of forecasters can, under certain circumstances, be even lower than those in the aggregated forecasts (Rowe and Wright, 1999; Sniezek, 1990). Fifth, use of advisors also reduces forecasters' error but forecasters tend to place insufficient weight on advice they receive (Harvey and Fischer, 1997; Yaniv and Kleinberger, 2000). For some applications, forecasters can receive advice in the form of formal forecasts produced by models of the underlying processes. Again, forecasters often place insufficient weight on the advice. Thus, research has shown that they are inclined to make unwise adjustments to model-based forecasts, thereby causing their final forecasts to be worse than those originally produced by the model (Fildes et al., 2009). As this brief review demonstrates, these elements affecting judgmental forecasting have been subject to a considerable amount of research. Further research into them is, therefore, perhaps not as urgent as it is

for the effects of presentation format: they are not the focus of experiments reported here.

1.3 Understudied areas in judgmental forecasting

The above analysis provided an overview and a categorisation of the main research findings in judgmental forecasting. This detailed analysis allows for identification of research areas that are currently understudied and that need further investigation; these are effects of horizon, length and scale of series as well as effects of dynamic rather than static data presentation. These are the areas I chose to study within the present thesis. In subsequent sections, I will provide a brief overview of the literature in these four areas of interest. Further and more detailed analysis will be offered in each Chapter devoted in the corresponding research theme.

1.3.1 Order effects in judgmental forecasting

Forecasting horizon appears to influence judgmental forecasting accuracy. Shorter horizon lengths are associated with smaller judgmental errors and longer horizons are associated with larger ones in line with expectations. Evidence for this finding can be traced in several judgmental forecasting experiments using different types of time series. Bolger and Harvey (1993), for example, used trended and untrended series with various degrees of aurocorrelation and found that forecasters' errors increased with forecasting horizon. Many researchers who have studied the trend damping phenomenon

and found that judgmental forecasts deviated from the trend line as a function of the forecast horizon, thereby increasing their error (e.g., Andreassen and Krauss, 1990; Bolger and Harvey, 1993, 1995; Eggleton, 1982; Harvey and Bolger, 1996; Harvey et al., 1994; Harvey and Reimers, 2013; Keren, 1983; Lawrence and Makridakis, 1989; Mackinnon and Wearing, 1991; O'Connor, Remus and Griggs, 1997; Sanders, 1992; Timmers and Wagenaar, 1977; Wagenaar and Sagaria, 1975; Wagenaar and Timmers, 1978, 1979). Finally, Harvey (1995) who studied the effects of noise levels in forecasting accuracy, by presenting seasonal series to the participants, confirmed this result. He reported increasing errors with an increase in forecast horizon. He also revealed that the magnitude of this effect depended on the series' noise levels as well as the series' frequency. Steeper gradients of the seasonal series and greater noise levels were associated with larger errors.

Although several studies have tentatively identified the cognitive mechanisms involved in one-step-ahead forecasts, little has been done to explain the deterioration of forecasting performance for longer horizon forecasts. For short horizons, research suggests that people use simple heuristic mechanisms to take into account pattern, autocorrelation and noise information in the series. As described in previous sections, the anchor and adjustment heuristic has been proposed as one way of explaining performance in judgmental forecasting tasks (Harvey, 2007). But what happens with longer horizons? Is this deterioration in performance only an effect of errors' superposition for various time steps? Or is it also an effect of the cognitive strategy chosen by the participants?

Bolger and Harvey (1993) used stepwise regression to reveal whether longer horizon forecasts were based on a set of preceding forecasts. Their results suggested that while one-step ahead forecasts exploit pattern information, this is not the case for longer horizon ones; it is the immediately preceding forecast that mainly influences longer horizon forecasts. This means that beyond the one-step-ahead horizon, people use a simple heuristic strategy, which resembles the naïve forecasting approach. This finding was confirmed by Lawrence and O'Connor's (1992) research; they found that people adopt different smoothing constant values for different forecast horizons when employing the averaging heuristic for untrended series. Was this an effect of suboptimal parameterization? For shorter horizons, the use of heuristic mechanisms often produces acceptably low levels of error and participants take into account the patterns and the autocorrelation of the series. For longer horizons, however, pattern elements, though essential for optimal forecasting, seem to be ignored. Instead, longer horizon forecasts seem to be mere repetitions of the previous data point. But why is this so and are there any task characteristics that would help forecasters improve their performance?

Two papers have dealt with this presentation format issue and its impact on the cognitive strategies adopted by the forecaster. The first one by Welch, Bretschneider and Rohrbaugh (1998) concluded that, by making the long-term elements of the series more salient to the forecaster, MAPE decreases. Participants assigned to an experimental condition, in which the only the basic series information was presented to them, were less accurate than those

assigned in a condition in which the long term trends and long-term levels of the series were highlighted.

Harvey et al. (1997) deal with the same research question; in their second experiment, they tested the idea that a single forecast for a distant horizon would be better than a forecast for the same horizon embedded within a set of forecasts for multiple horizons. Their hypothesis was based on the argument that if people introduce noise in successive forecasts in an attempt to represent the series, thereby impairing overall accuracy, they should not do so for single forecasts. Hence, forecasting performance should be enhanced for single forecasts. To test this hypothesis they assigned half of the participants in a six-horizon successive forecasting task and the rest of the participants in a single forecasting task either for the first or for the six forecast horizon. They used seasonal series and forecasts started at a 0.375 phase of the sinusoid. Their results, though, did not show any significant differences between successive or single forecasting conditions. These findings undermined the pattern masking account, which posited that participants are aware of the pattern they should produce, but they mask it by adding noise. Instead, the representativeness account was supported: participants added noise even when they produce single forecasts.

Judgment processes in forecasting can include intuitive or analytical modes of thinking (Kahneman, 2011). Intuitive modes of thinking, such as heuristic processing, are quick and automatic, producing approximate judgments to a problem. On the other hand, analytic thinking requires more time and can

produce more accurate judgments. Welch at al.'s (1998) paper emphasized the need for the forecaster to use more analytic modes of thinking when forecasting for longer horizons in order to take into account the long term characteristics of the series, something that is essential for distant forecasts. Is there a way to prime the forecaster to think more analytically? And would that be beneficial for the forecaster performance? It might be the case that forecasters who are faced with a more difficult forecasting task (for example, requiring high deliberative effort) might need to think more analytically and, thus, produce more accurate forecasts.

The limited research related to longer horizon judgmental forecasts and the fact that pattern components might not be exploited by heuristic mechanisms as in the case of one-step ahead forecasts create an interesting area for research. In Chapter 3, I report an investigation into horizon length errors and order effects in judgmental forecasting from various types of time series, and I identify which horizons and presentation formats are optimal for forecaster accuracy. The aim of this research is to enhance the accuracy of judgmental forecasting and, at the same time, to describe the cognitive mechanisms involved in each case.

1.3.2 Length Effects in Judgmental forecasting

Judgmental forecasts are widely used in practice either alone or in combination with statistical forecasting tools. People using their judgment, though, tend to make forecasts that are not in agreement with statistical techniques (Lawrence

et al., 2006). To date, most research on how people use their judgment to make forecasts from time series suggests that the process involves extraction and use of pattern information such as trends, seasonality and noise. Pattern extraction, though, seems to be dependent on graphical characteristics of the series considered. Andreassen and Krauss (1990), for example, suggest that people need to have a series sufficiently long for them to have confirmation of any patterns thought to exist. Andersson et al. (2012) find similar evidence in a stock investment paradigm; in their second experiment price predictions improve with price-series length. Nevertheless, contrary to expectations, Lawrence and O'Connor (1992) found that the length of time series affects forecast accuracy in the opposite way: forecast accuracy decreases when participants are presented with longer time series. The same conclusion was reached by Waagenar and Timmers (1978) who used an exponential task: extrapolation from exponential functions was improved when fewer data points were shown to the participants.

Time series length is expected to influence judgmental forecasting accuracy especially in experiments with static (e.g., graphical) presentation. By varying the length of the series presented to a subject, the amount of information available for processing changes. Shorter time series provide evidence for elements such as the last data point and the local trend. Longer time series, on the other hand, contain more information. These series carry evidence related to the series' signal, overall trends, introduced randomness, autocorrelations and so on: these elements can be combined to produce a forecast. Thus, the

mechanisms of data processing might change when the amount of information presented to people varies (Einhorn, 1971).

Limited research related to people's sensitivity to time series length in judgmental forecasting and the fact that pattern components are needed by heuristics create an interesting area for research. In Chapter 4, I report an investigation into length effects in judgmental forecasting from various types of time series, and identify which lengths are optimal for the forecaster accuracy with each series' type. An additional aim of the research was to investigate whether different cognitive mechanisms are involved for series of different lengths. For example, different versions of the anchor and adjust heuristic may be used to make forecasts from short and long data series.

1.3.3 Scale effects in judgmental forecasting

The scale used for graphically presented time series may influence judgmental forecasting accuracy. Legge, Gu and Luebker (1989) support this notion by arguing that the scale at which data is presented will influence the graphical perception of the behaviour of a time series. Nevertheless, contrary to expectations, Lawrence and O'Connor (1992) failed to find any effect of the scale of data presentation on forecast accuracy. This result might stem from a general law of stimulus perception, the scale invariance law (Chater and Brown 1998, 2008). This law posits that the perception of stimuli is independent of their size. Lawrence and O'Connor (1992) findings might also relate to the fact

that humans seem not to be able to attribute absolute coding magnitudes to stimuli but only relative ones (Stewart, Brown and Chater 2005).

However, a restricted number of experiments concerning scale effects in judgmental forecasting accuracy have been reported. Results might be different for various types of time series. Also, within the same type of time series, differences might occur if the distribution of the underlying noise function is manipulated. This issue of noise type effects has not been studied before. Thus, this area of research is appropriate for further investigation. In Chapter 5, I report an investigation into series' scale and noise type effects in judgmental forecasting from various types of time series.

1.3.4 Forecasting from experience

The type of display used for presenting the time series may also influence judgmental forecasting accuracy. Will people react with the same way when presented with static or dynamic data? This is a seriously understudied area in forecasting with the use of judgment. Dynamic series' presentation, where the forecaster experiences individually each point of the time series sequentially is a seriously understudied area in judgmental forecasting. However, this is a widespread task in the domain of finance where professionals receive real-time information from which they have to extrapolate in order to make their investment decisions. Traders for example make instant decisions on the basis of data they receive in real-time on their computer screens. Managers also receive real-time information for developments in the market and base their

subsequent decisions on their judgmental forecasts (see, for example, Nuthall, 2001). Policy makers also receive real-time information for real GDP growth and other indicators such as inflation to base their decisions. Monetary policy decisions, for example, are taken in real-time with the use of judgment and models on the basis of assessments of current and future economic conditions (for relevant nowcasting models, see for example Giannone, Reichlin and Sala, 2004). Weather forecasters also use their judgment in real-time settings (see, for example, the experiment by Lusk and Hammond, 1991**).**

Also, in life, people receive streams of data of interrelated events and base their anticipations on real-time, sequential information, or updated information of a single cue they have experienced (for example, the weather, prices in the supermarket and so on). Adaptation accounts would suggest that such a successful interaction with cues that are interrelated is essential for human beings. These types of experiential tasks should not be confused with forecasting tasks where domain experts use their professional experience to produce their forecasts. Here, the target area of research is high-frequency information assimilation through experience. These tasks do not refer to experience gained over the years. Forecasting from real-time experience involves the exposure to streams of data, the assessment of whether patterns are present in these data, and finally the assimilation of all the information for the final forecast to be produced. Given the potential usefulness and practical application of judgmental forecasting from experience, there is good reason to study it. I do this in Chapter 6, where I present a set of exploratory

experiments, which deal with the basic characteristics of the forecasting process.

1.4 Summary and Overview of the Thesis

The scope of this thesis is to examine understudied areas in judgmental forecasting from graphs and from experience and to suggest improvement strategies. I will specifically examine presentation format phenomena that concern horizon, order, length, scale and dynamic display effects. In order to obtain generalisable results, the aforementioned phenomena will be investigated using various types of time series in the experiments.

In the second chapter, I will present the basic methodological approaches to studying judgmental forecasting phenomena. In the literature, experimental research involves mainly the tasks with static graphs of series. In the present thesis, judgmental forecasts from static graphs will be explored in Chapters 3, 4 and 5, while judgmental forecasts from real-time experience of series will be explored in Chapter 6. Thus, in Chapter 2, I will first present the main methodological issues for judgmental forecasts from static graphs along with the statistical methods used to address these phenomena. In subsequent sections of Chapter 2, I will present a novel experimental paradigm designed to study forecasting from experience. In this paradigm, participants are presented with a time series in an experiential way with the use of bar charts. Error measures and

other statistical techniques used in the present thesis will also be described in Chapter 2.

In Chapter 3, horizon effects will be investigated in judgmental forecasting from graphs (Experimental Studies 1-2). A novel way of requiring participants to make their forecasts will be presented; the forecaster will first produce his forecast for distant horizons and then for the remaining horizons (e.g. to the most distant horizon will be used as an end-anchor). Forecasting performance with the aid of end-anchors will be compared with traditional forecasting where the forecaster begins forecasting from the closest horizon. Order effects will also be examined in this Chapter and, more specifically the effect of direction on forecasting performance will be thoroughly investigated.

In Chapter 4, I will examine the way the length of the series affects forecasting performance from various types of time series (Experimental Studies 3-4). A set of lengths will be selected on the basis of previous findings in the literature. Forecast performance for various lengths will then be assessed; also, the anchoring and adjustment mechanisms will be examined in conjunction with a naïve benchmark. In this chapter, length effects for later horizons will also be examined in an effort to reconcile findings from previous research on the issue.

Dimensional factors related to the forecasting task will be further investigated in Chapter 5, where graphs' scale will be manipulated (Experimental Studies 5-6). The types of time series entered in these experiments are selected in order to uncover the effects related to the series' noise distribution; two different noise types (uniform and Gaussian noise) will be introduced to the series of interest.

Apart from forecasting performance, anchoring and adjustment mechanisms will be also scrutinised in this section to determine whether participants are sensitive to different scales and noise functions.

Presentation format issues will be further explored in dynamic settings in Chapter 6, in a number of experiments (Experimental Studies 7-12), where the new experimental paradigm for judgmental forecasting will be tested. Here, I will test already identified robust phenomena in judgmental forecasting within this novel dynamic setting and compare their magnitude and direction with those found in static environments.

Finally, findings will be summarised in Chapter 7 and their implications will be discussed.

Chapter 2 Methodology

Overview

In this chapter, I present a general description of the experimental and statistical methods and techniques used throughout this thesis. I start by describing the time series stimuli and methods employed to construct them in both judgmental forecasting tasks from graphs and from experience. Next, I outline the measures used to assess forecasting performance and evaluate their appropriateness for each of the tasks I used. Finally, I outline the basic techniques for measuring the robust biases found in judgmental forecasting: trend damping, autocorrelation illusion and noise introduction.

2.1 Experimental Methods

Judgmental forecasting from graphs

For the study of judgmental forecasting from graphs, I employed tasks commonly used in the field of judgmental forecasting (e.g. Sanders, 1992; Önkal, Gönül and Lawrence, 2008; Goodwin and Fildes, 1999; Reimers and Harvey, 2011; Harvey and Reimers, 2013). In these tasks, series are presented to participants as line graphs. In each trial of the experiments found in this thesis, participants observe a graph and are requested to extrapolate from that. After the end of each series, a number of vertical lines are presented in the next

time periods to indicate where forecasts have to be made. The number of these lines depends on the number of forecasts requested by the participant in each experiment. When a forecast is made, by clicking on one of the vertical lines, a coloured dot appears and this point is connected with a line with the previous point.

As discussed in Chapter 1, there might be cases where presentation format can play an important role in the way forecasts are produced from graphs (Harvey and Reimers, 2012). The choice between point, line or bar formats seems to be important, with preliminary evidence showing higher errors to be associated with bar graphs. Throughout the present thesis, such presentation format biases are not investigated in depth and, thus, a homogenous experimental paradigm is used across all the experiments that involve graphs; series are always presented in line graphs, where successive points are interconnected with a line. Forecasts provided by participants are also connected with a line with the previous data point. Figure 2.1 shows a basic display for this experimental paradigm.

In each of the experiments requiring forecasts from graphs, time series are generated uniquely for each participant and the types of series used in each experiment are randomly ordered separately for each of the participants. This methodology ensures that results are not artifacts of the specific series used or of the order in which those were presented (e.g. context effects).

In the majority of tasks where forecasting from graphs is requested, tasks are not performed within particular scenario, such as one associated with sales forecasting, to avoid introduction of frame-specific biases (e.g., elevation biases arising from optimism or perceived control effects). Hence, the vertical axes of the graphs used to present the series are unlabelled.

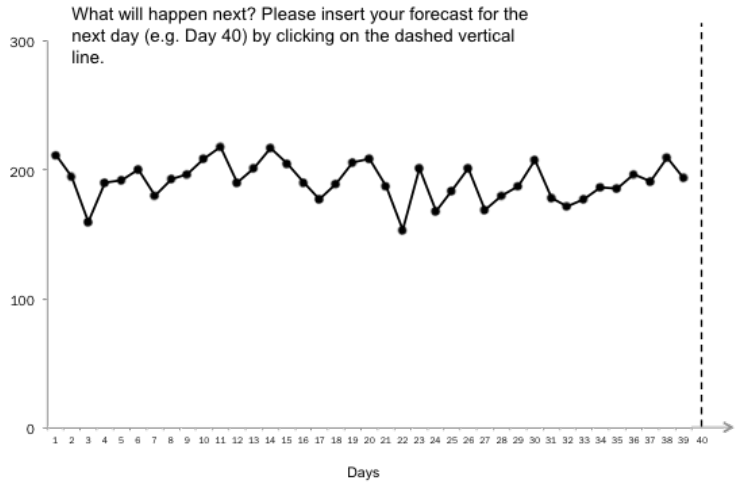

Figure 2.1 Standard experimental paradigm showing 39 data points (seen by participants) followed by a vertical line, where participants are requested to mark their forecast.

In these tasks, each participant performs the task individually. They read a short introduction to the study and then enter their demographic details (age, sex). They then are instructed to view each series and click on each of the vertical lines to show where they expected future points in the series to appear. In the majority of the experiments contained in this thesis, forecasts are made from the nearest horizon to the most distant one with the exception of two experiments in Chapter 3, where the order of forecasts is reversed. In this case, all vertical lines

are again presented at the same time as the series. However, an explicit screen message prompts participants to make their forecasts in a reverse order. In order to ensure that forecasts are actually made in this order, this task is constrained; thus, the programme accepts a forecast (and show a blue dot to indicate its position) only if it is made in the required order.

There are also two conditions in Chapter 3 (Experimental Studies 1 and 2) and Chapter 4 (Experimental Study 4) where participants first make their forecast for the most distant horizon for each of the series presented to them. In this case, a single vertical line is presented at furthest horizon with each series to signal that only the forecast for that horizon is required. In Chapter 3, once this forecast is made for all series, participants return to each series (presented in the same order as before) to make the remaining required forecasts. To enable them to do this, the remaining vertical lines appear on the screen at this point to indicate the positions of these required forecasts. As forecasts are made, a blue line links each new forecast with the last data point, or with both the immediately preceding forecast and the forecast for the most distant horizon. In Chapter 4, participants are not required to forecast the remaining points.

Judgmental forecasting from experience

People often need to deal with streams of information that they receive over time. In this thesis, I propose a new way to directly investigate forecasting performance, by introducing a simple forecasting task, where the forecaster experiences the time series in real-time instead of observing it via graphs. In this task, values that the forecasters experience were presented as a sequence of bar graphs and participants were asked to forecast the next values (Figure 2.2). Importantly, the structure of the time series within this paradigm could be modified to match that used in judgmental forecasting tasks from graphs. The only difference was the presentation format. Based on the ability of the brain to make predictions when processing sequential stimuli (see for example Fiedler and Juslin, 2006) in a variety of domains (for a review, see Bubic, Cramon and Schubotz, 2010), I expected that the forecaster would be able to process values across time and forecast accordingly. This dynamic paradigm, which I label "the experiential forecasting paradigm", lies at the intersection of low and higher level cognition. Prediction judgments deriving from this paradigm can be used as a proxy to understand more about judgmental forecasting from graphs but may also cast light on more complex forecasting decisions such as those that take place when real-time data, such as news, influence processes such as group forecasting (Önkal et al., 2012).

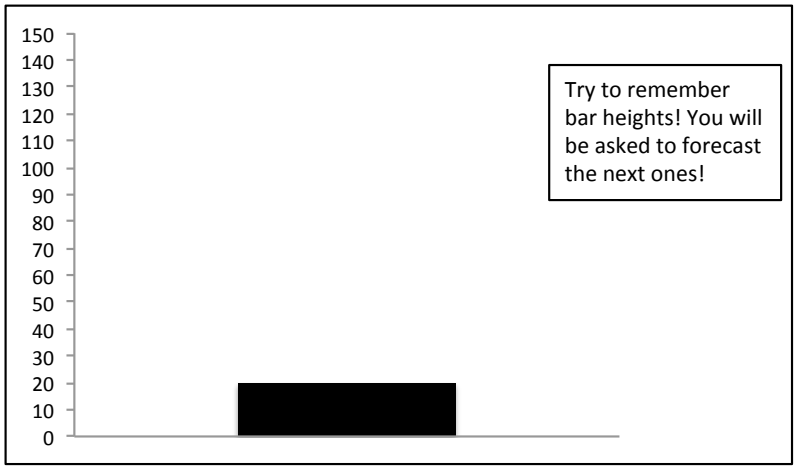

Figure 2.2 Screenshot of the experiment. The bar chart represents a specific data point in a time series.

In these experiments, participants were presented with a sequence of bars that formed a specific time series. In the beginning of each experiential experiment, they were asked to enter their age and gender. Participants are randomly assigned to one of the experimental conditions. They then read instructions, where a particular scenario was described in order to render the task more representative of a typical experiential forecasting situation:

"Imagine you are a trader… You are now at Wall Street premises and you are observing a specific stock price in this screen. The stock price values are not presented in numbers. Instead, they are presented with the use of bar charts. The greater the height of the bar is, the larger the price of the stock. A first bar appears in your screen with the initial price. When the stock price changes (it does within seconds in the stock market), the next bar appears in your screen. The previous one disappears. At the end of the task and after observing approximately 40 consecutive stock price changes, you will have to predict the height of the next two bars (i.e. stock prices) by mouse-clicking the height of the bar. Will the price of the stock increase or decrease? Will it remain the same? Your prediction will show whether you are appropriate to become a trader!"

Once they had indicated that they had understood the instructions, they experienced the successive data points. They then had to forecast the next data points by using the mouse and clicking at the heights at which they thought that the next points would appear. Their understanding of the task was checked twice: first, during the experiment where they had to answer to the experimenter whether the task was clear to them and then at the end of the experiment, where they had to describe what type of time series they have just experienced by selecting among different graphs. The number of forecasts requested from participants was determined by the need to test the experimental hypothesis. For example, in tasks where the goal was to understand whether the forecasters introduce noise to their forecasts, they were asked to provide five forecasts. In other cases, where trend damping was investigated, two forecasts were enough to determine whether forecasts were below the trend line. In the next section, I provide more detailed information about the stimuli used in these tasks.

Participants and subject pools

To conduct the experiments both in graphical and experiential settings, I used participants from two sources: either UCL's subject pool or Amazon Mechanical Turk, a crowdsourcing web service commonly used for data collection by psychologists. UCL's subject pool comprises mainly undergraduate and postgraduate students but also a small minority of individuals outside UCL who are interested in acting as human subjects. These individuals were excluded from the experiments reported here in order to maintain the homogenous characteristics of the University sample. On the other

hand, Amazon Mechanical Turk pool of subjects comprises individuals around the world, who register in this service in order to accumulate a monthly allowance by participating in various experiments published via Amazon.

Manipulations within the experimental conditions of this thesis were designed in such a way so that all comparisons referred to the same pool. Moreover, cross-experimental comparisons were only conducted in cases where subjects came from the same pool of subjects. Although research has demonstrated equivalence between results obtained from online and laboratory studies (Mason & Suri, 2012), there were several differences between these two pools of participants, which might have produced experimental artefacts if one directly compared results from these two sources. For example, the average age of subjects drawn from the UCL pool was 26 years old while, Amazon participants' average age was 31 in the experiments of the current thesis. Additionally, the majority of UCL participants have attended at least one course of inferential statistics, which was not the case for subjects drawn from Amazon pool. Also, in the laboratory, the experimenter could assess the subjects' understanding of the task before trials begun; the same was not possible for online tasks. Nevertheless, to avoid possible discrepancies between results from these two different pools of subjects I significantly increased the number of participants in each of the web experiments conducted in the current thesis.

2.2 Stimuli and Tasks

In judgmental forecasting tasks, the stimuli used are time series of different types. Some of the research findings and corresponding phenomena are often verified for specific series' types and not for others (see for example the studies of Andersson et al., 2012; Wagenaar and Timmers, 1978; Lawrence and O'Connor, 1992), a fact that causes difficulties in generalizing results (Goodwin and Wright, 1993). This observation suggests that carrying out experiments with a variety of series' types would help to clarify whether specific phenomena are replicated for a variety of series' types and, consequently, generalizable.

For the purposes of this thesis, I used mainly five types of series: untrended series of independent data points with various levels of noise depending on the hypothesis of the experiment; untrended series of autocorrelated data points with various autocorrelation coefficients, again depending on the experimental design; series of independent data points with a linear trend imposed upon them; series of independent data points with a seasonal trend and, finally, untrended non-linear long memory (fractal) series.

More specifically, in Chapter 3, where I study horizon and order effects I presented participants with untrended series of independent data points, untrended series of highly autocorrelated data points, series of independent data points with a linear trend imposed upon them and series of independent data points with a seasonal trend imposed upon them. In Chapter 4 where I studied length effects, I used untrended non-linear long memory series, untrended series

of highly autocorrelated data points, a series of independent data points with a linear trend imposed upon them and a series of independent data points with a seasonal trend imposed upon them. In Chapter 5, where scale effects were investigated, I used only untrended series of independent data points and untrended series of highly autocorrelated data points. Nevertheless, two types of noise distributions were used in this paradigm. Finally, in the experiential forecasting experiments, I used series of independent data points with a linear trend imposed upon them (both upward and downward ones) to study whether trend damping occurs. I also employed untrended series with various autocorrelation coefficients to study the autocorrelation illusion and, finally, I presented untrended series of independent data points with different noise levels to study whether noise introduction occurs.

These types of time series were chosen for the following reasons. Non-linear long memory series of high persistency appear to be interesting because their optimal forecast lies close to the last data point. Non-linear long memory series have not been studied in the past in judgmental forecasting settings. Theoretically, the degree of persistence in this type of series and, thus, the optimal forecast, could be extracted by the forecaster by observing the smoothness of the series; higher persistency series are represented by smoother graphs. Seasonal and trended series on the other hand contain a signal, which should be detected by the forecaster. Random series represent deviations around a mean value and are always useful as a control. Finally, autoregressive time

series of various autocorrelation levels were used to detect whether participants' implied autocorrelation matches the autocorrelation of the series.

Examples of the types of series used can be seen in individual figures in each of the experimental chapters of this thesis.

To construct untrended series, I followed Box and Jenkins (1976) methodology by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon$, where X_{t-1} was the previous observation, μ was the mean of the series, α was the degree of autocorrelation, and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2. The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen. The autocorrelation coefficient was selected according to the goals of the experiment and varied between $\alpha = 0$ for random series and $\alpha = 0.9$ for highly-autocorrelated series. The variance σ^2 depended again on the experimental hypothesis (high, medium or low noise components).

Patterned series, such as trended and seasonal series, were produced by imposing the appropriate pattern on an independent series. More specifically, linear trended series were produced by using the equation: $X_t = \alpha t + \varepsilon_t$, where α represented the gradient of the series (shallow, medium or steep gradient) and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2. The gradient and variance of the series were chosen according to the experimental design. Seasonal series were constructed by using the equation: $X_t = \alpha \cos(\beta t) + \varepsilon_t$. The starting point of

these series was chosen so that the last data point was close to the vertical mid-point of the screen.

Finally, to construct the non-linear long memory series, I used the multiple time-scale fluctuation approach (Koutsoyiannis, 2002). The autocorrelation and variance restrictions were calculated from the equations after the Hurst exponent value was selected to be equal to 0.9. Time series, such as those chosen for the purposes of this thesis, with high Hurst values (H = 0.9) exhibit a long memory autocorrelation function of many time steps. This means that optimal forecasts lie very close to the last data point, rendering anchoring and "conservative" adjustment a very efficient way of producing forecasts. Macroscopically, this property can be traced by the smoothness of the series. Koutsoyiannis (2002) has shown that by superimposing three or more Markovian functions, one can obtain a good approximation of fractal Gaussian noise by applying specific restrictions on the relations of their variances, autocorrelations and fluctuation time scales. The algorithms used are based on the same principles as the fast fractional Gaussian noise (FFGN) algorithm (Mandelbrot, 1971) with the difference that this approach uses only three AR(1) components (many fewer than the FFGN) and that the parameters of the algorithm are determined by much simpler equations. The multiple time-scale fluctuation approach thus makes use of three Markovian processes to construct the fractal Gaussian noise approximation. These Markovian processes AR_1, AR_2 and AR_3 have the following properties:

- Means μ

- Variances σ_1^2, σ_2^2, σ_3^2

$\sigma_1^2 = (1 - c_1 - c_2)\ \gamma_0$

$\sigma_2^2 = c_1\ \gamma_0$

$\sigma_3^2 = c_2\ \gamma_0$

where c_1 and c_2 are positive constants (with $c_1+c_2 < 1$)

- Autocorrelation functions ρ, φ, ξ:

$\rho = 1.52\ (H-0.5)\ 1.32$

$\varphi = 0.953 – 7.69\ (1-H)\ 3.85$

$\xi = 0.932 + 0.087\ H$, for $H < 0.76$

$\xi = 0{,}993 + 0{,}007\ H$, for $H > 0.76$

and

$\varphi = e^{(-\delta/\lambda)}$

$\xi = e^{(-\delta/\nu)}$

where δ, λ, ν are the time scales of interest and H represents the Hurst exponent value.

2.3 Error measures

Forecast error in this thesis is defined as the difference between the forecasted value (F) and the actual value (A).

In the experiments outlined in the present thesis, forecast errors serve two important functions: 1) to measure overall accuracy of the participant who is

requested to make judgmental forecasts and 2) to measure specific cognitive biases or deviations from the true value that tend to be in one direction. There are several types of biases identified in judgmental forecasting, as discussed in Chapter 1. Among those, trend damping and elevation effects are those revealed via error measures; other cognitive biases such as noise introduction and autocorrelation illusion are assessed with alternative methodologies, presented in the next section.

In the experiments reported in this thesis, I use a variety of time-series types. Nevertheless, I only compare errors produced from the same types of series. In other words, within-series comparisons are assessed. Error comparisons between different types of series (cross-comparisons) were not considered here; the focus was mainly to understand whether specific manipulations related to the task characteristics improved performance individually for each type of series and whether these manipulations affected the underlying cognitive mechanisms. This renders measures such as the mean absolute error (MAE = |Absolute value – Forecast| appropriate for within-series comparisons (see also, Armstrong and Collopy, 1992). Relative and percentage measures would have been useful for cross-comparisons between different series' types.

In order to evaluate forecasting performance, I used the mean absolute error measure (MAE) and the root mean square error (RMSE) – for multiple forecasts - whereas to evaluate specific underlying phenomena such as trend damping and positive elevation biases, I used the simple and symmetric mean signed error as well as cumulative error, which put equal weights on both positive and

negative errors. RMSE is a common measure, but it is well known to suffer from a number of problems; these mainly concern cases where forecasting performance is compared across series with different scale (Armstrong and Fildes, 1995; Armstrong and Collopy, 1992). For the purposes of the present research, RMSE can be still be useful for comparing accuracy from the same type of series under different conditions.

MAE is the absolute forecast error made by participants and corresponds to the difference between the judgmental forecast and the actual outcome of the series whereas mean signed error is calculated by subtracting the optimal forecast instead of the actual value. Optimal forecasts were calculated by dropping the noise component from the generating algorithm of the series. As mentioned before, while MAE was used as a measure of forecasting performance, MSE was especially useful in order to nicely reveal phenomena such as trend damping and elevation biases. Although randomization of series and trials was employed to avoid experimental artifacts, actual and optimal values were employed both in MAE and MSE error equations respectively to ensure that the whole process of collection and analysis of the data was conducted in an unbiased manner.

Cognitive biases were also studied by extracting the mean absolute distance of the forecast from the last data point. This measures the degree of adjustment from the previous data point.

Extreme values were mainly treated statistically; participants whose forecasts were at least three inter-quartile ranges from the median of each group were removed and replaced (6% of all participants).

2.4 Robust biases in judgmental forecasting

Although difficulties emerge when trying to generalise research findings from judgmental forecasting studies due to the great variety of series that can be used in forecasting tasks (Goodwin and Wright, 1993), the most robust phenomena associated with forecasting biases in a variety of studies and a diversity of series' types are trend damping (Harvey and Reimers, 2013), noise introduction (Harvey, 1995) and the autocorrelation illusion (Reimers and Harvey, 2011). In the present thesis, these biases are investigated by using methodologies suggested in the literature.

It is useful to explain here the distinction between elevation effects and trend damping; elevation effects occur when forecasts are consistently above or below the trend line but with no difference in the slope between the data series and the sequence of forecasts. To measure whether trend damping occurs, two methodologies are used in the literature. One is associated with the exploitation of the signed error measure, which is calculated for each time-step as the difference between the forecast and the corresponding trend value. A repeated-measures ANOVA is then run with the dependent variable of the signed error and the independent variable of time horizon. The number of levels of this

GLM model depends on the forecasting task characteristics. Trend damping occurs when significantly higher errors are associated with the most distant horizons. Another methodology proposed by Harvey and Reimers (2013) suggests the use of regression lines to fit each participant's data individually, with time horizon being the x axis variable. Regression fitting results in obtaining a slope value for each individual. These values are then compared with the actual gradient of the series; significantly shallower slopes indicate that trend damping occurred.

Noise introduction effects (Harvey, 1995) are treated with a similar technique except that now, after fitting linear regression lines to the forecasts, residuals in each noise condition are compared via a one-way ANOVA. If those are significantly different from each other (i.e. significantly greater for higher noise), then the researcher can conclude that subjects introduced more noise in the higher noise condition. If differences are not significantly different, then there is no evidence that noise is introduced into the forecast sequence in proportion to the noise level in the data series.

Finally, the autocorrelation illusion is assessed via the calculation of implied autocorrelations, a methodology introduced recently by Reimers and Harvey (2011). Implied autocorrelation can be calculated by dividing the following quantities: the distance between the forecast and the series mean and the distance between the forecast and the previous data point. This estimation is directly derived by the equation for the autocorrelation. In Chapter 6,

Experimental Study 12, I also describe an alternative methodology for estimating participants' sensitivity to autocorrelation.

With these methodologies, one can tackle issues associated with robust biases in judgmental forecasting. Throughout this thesis, these methodologies are used numerous times. New directions towards enriching these methods are presented in the last chapter of this thesis.

Chapter 3 Order Effects in Judgmental Forecasting

Overview

As discussed in Chapter 1, the order in which judgmental forecasts of proximal or distal periods are made is an understudied area in judgmental forecasting. Uncertainty increases as we move into the future. Unsurprisingly, therefore, both statistically based forecasts and judgmental forecasts are worse for more distant forecast horizons (Lawrence et al., 1985). Rate of deterioration, measured by increase in mean absolute percentage error (MAPE), is broadly similar for the two types of forecasts (Lawrence et al., 1986) but reasons for it differ. As discussed in Chapter 1, judgmental forecasts, unlike most statistical forecasts, show trend damping. This causes their signed error to increase over the forecast horizon (Harvey and Reimers, 2013; Lawrence and Makridakis, 1989). What cognitive processes produce this phenomenon? To make forecasts for the first horizon, people appear to use the last data point as a mental anchor and then make some adjustment away from that point to take account of the pattern in the series. Typically, these adjustments are insufficient. As a result, trend damping is observed with trended series and forecasts from non-trended series appear to exaggerate the sequential dependence in the data. Furthermore, people add random noise to the result of the anchoring and adjustment process to produce their forecasts (Harvey, 1995; Harvey et al, 1997). They may do this to make their sequence of forecasts look similar to the data series. Forecasts for

later horizons are made in a similar way except that the previous forecast rather than the last data point is used as a mental anchor (Bolger and Harvey, 1993). As a result, the random noise added to previous forecasts accumulates as people make forecasts for increasingly distant horizons. If this accumulation of random noise could be eliminated, forecasts for these more distant horizons would improve in accuracy and variability across forecasters in the trajectory of the forecast sequence would be reduced.

End-Anchoring

This analysis suggests that forecasting performance would be improved by preventing forecasts for horizons beyond the first one being made in sequence and, thereby, accumulating the random errors associated with each one. One obvious way of doing this is to ask forecasters to make their forecast for the most distant horizon first. One might assume that forecasters do this by using the anchoring and adjustment heuristic that is normally used to make an initial forecast. For example, for trended series, instead of making a forecast for the first horizon by anchoring on the last data point and adjusting away from that value by a proportion (P) of the difference between the last two data points (Bolger and Harvey, 1993), they could make a forecast for, say, the fifth horizon by anchoring on the last data point and adjusting away from that value by 5P (i.e. five times the size of the adjustment used when forecasting for the first rather than the fifth horizon). Forecasters may find making an initial forecast for the most distant horizon more difficult than making an initial

forecast for the first horizon and it may take them a little longer. However, once that forecast has been made, their task is transformed from one of extrapolation to one of interpolation. This manipulation is expected to produce its greatest improvement on the forecast for the most distant horizon. This is the horizon that would be most affected by accumulation of noise components in previous forecasts. However, because interpolation is a more constrained task than extrapolation, the end-anchoring produced by making an initial forecast for the most distant horizon may also improve forecasts for less distant horizons. To make the intervening forecasts, people may simply use linear interpolation between the last data point and their forecast for the most distant horizon. They are still expected to add a noise component to the results of each forecast in this interpolation (Harvey, 1995) but this would not determine the trajectory of the forecast sequence.

Based on the above rationale, I will test the following hypotheses.

H_1: Requiring forecasters to make their initial forecast for the most distant horizon will produce more accurate forecasts for that horizon than when they make their forecast for it last.

H_2: Requiring forecasters to make their forecast for the most distant horizon first rather than last will also increase the accuracy of forecasts for less distant horizons.

Reversing the direction of the forecasting

Once forecasters have made their initial forecast for the most distant horizon, they could proceed in one of two ways. They could forecast forwards in time from the end of the data series towards their existing forecast for the most distant horizon. So, for example, forecasts for five horizons would be made in the order: 5, 1, 2, 3, 4, where lower numbers represent horizons closer to the end of the data series. I shall refer to this as forward forecasting. Alternatively, they could make forecasts in in the reverse direction, working from their initial forecast for the most distant horizon back towards the end of the data series. Thus, when forecasts for five horizons were required, they would make them in the order: 5, 4, 3, 2, 1, where lower numbers again represent horizons closer to the end of the data series. I shall refer to this as backward forecasting. There are reasons to suppose that direction of forecasting will influence accuracy but that the effect of this variable will depend on the characteristics of the time series.

First, consider forecasting from series containing linear trends. Trend damping effects tend to be greater with downward than with upward trends (Harvey and Bolger, 1996; Lawrence and Makridakis, 1989; O'Connor et al., 1997). An upward trend when forecasting forwards, is transformed into a downward trend when forecasting backwards. Therefore, errors in forecasting upward trends are likely to be larger when people forecast backwards than when they forecast forwards. Second, suppose that the final point of an autocorrelated data series

has been perturbed well away from the mean or trend line of the series by noise. When forecasting forwards, forecasters could take the effects of autocorrelation into account (Reimers and Harvey, 2011): for example, if the last point of an untrended series with a first order autocorrelation of .5 was 8 points above the series mean, optimal forecasts for the next three horizons would be four, two, and one point above the mean. However, when forecasting backwards, they would be unable to make any allowance for autocorrelation. Third, with untrended independent data series, there is no obvious reason to expect any major asymmetries between forward and backward forecasting if interpolation is reasonably good. However, if it is poor (perhaps because people have some difficulty taking into account the position of the anchor they are moving towards), forecast errors for horizon 1 may be larger for backwards than for forwards forecasting whereas errors for horizon 4 may be larger for forwards than for backwards forecasting. These suggestions are merely examples of how forecasting direction may influence accuracy. There are many other factors that could differentially affect forward and backward forecasting. Therefore, the hypotheses that I test are fairly general in nature:

H_3: Accuracy of people's judgments when they forecast forwards from the end of the data series towards a forecast that they have already made for the most distant horizon will be different from their accuracy when they make forecasts in the opposite direction.

H_4: The effects of reversing the direction of the forecasting sequence will depend on the characteristics of the data series.

3.1 Order effects in judgmental forecasting (Experimental Study 1)

In this experiment, participants were presented with time series comprising 35 points and asked to make forecasts for the next five points. To test the above hypotheses, I manipulated the horizon for which the initial forecast was made (first versus last), the direction of forecasting when the forecast for the final horizon was made first (forwards versus backwards), and series' type.

3.1.1 Method

Participants

One hundred and twenty students (48 men, 72 women) from University College London acted as participants. They were recruited from UCL's subject pool. They had basic knowledge in statistics and had never attended advanced time series analysis classes. Their mean age was 26 years. They were paid £1.00 for their participation.

Design

Participants were divided into two groups. The first group (no end-anchoring) made their forecasts for the five horizons in the order in which the data points appeared (i.e. 1, 2, 3, 4, 5). The second group (end-anchoring) did not. Instead they made their forecast for the most distant horizon (i.e. 5) first. In this second group, there were two sub-groups. The forward forecasting sub-group made

their forecasts from the end of the data series towards the forecast that they had already made for the most distant horizon. Thus, their five forecasts were made in the order 5, 1, 2, 3, 4. In contrast, the backward forecasting sub-group made their forecasts in the reverse direction moving from their initial forecast for the most distant horizon back towards the final point of the data series. Thus, their forecasts were made in the order 5, 4, 3, 2, 1. All participants made predictions for four different types of time series. Hence, they each produced a total of 20 forecasts (five horizons for each of four types of series). Characteristics of the four types of series are described in the next section.

Stimulus materials

The four types of series were: an untrended series of independent data points; an untrended series of highly autocorrelated data points; a series of independent data points with a linear tend imposed upon them: a series of independent data points with a seasonal trend imposed upon them. Series were presented graphically. Examples of the four types of series can be seen in Figure 3.1. Each panel in the figure shows 35 data points (seen by participants) followed by five optimal forecasts (not seen by participants).

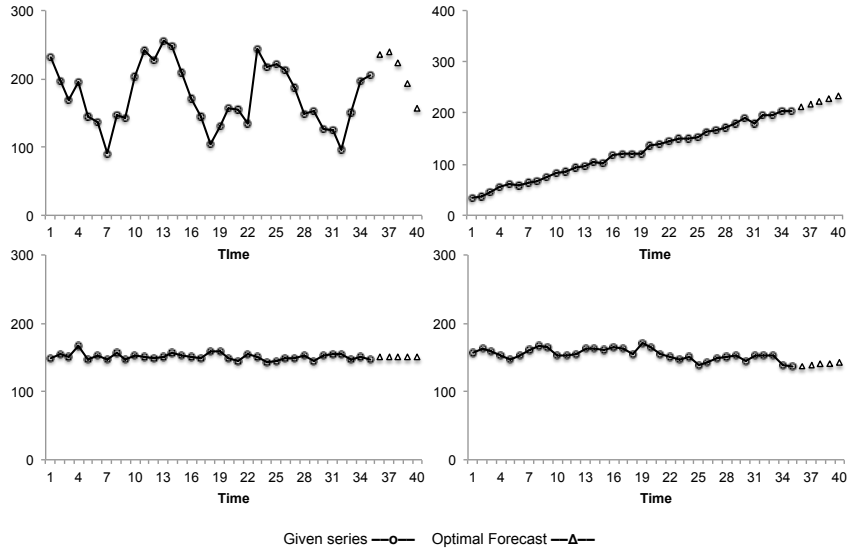

Figure 3.1 Examples of the four types of series, showing 35 data points (seen by participants) followed by five optimal forecasts (not seen by participants) for seasonally trended, linearly trended, autocorrelated, and random series (clockwise from top left).

Untrended series were constructed by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon$, where X_{t-1} was the previous observation, μ was the mean of the series, α was the degree of autocorrelation ($\alpha = 0.9$ for autocorrelated series and $\alpha = 0$ for random series), and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2 ($\sigma^2 = 30.0$ for both autocorrelated and independent series). The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen. Linear trended series were produced by using the equation: $X_t = 5t + \varepsilon_t$. Its noise term, ε, had a mean of zero and a variance of 19.0. The final data point of these

trended series was approximately 10% of the screen height above its vertical mid-point. Seasonal series were constructed by using the equation: $X_t = 70\cos(100t) + 170 + \varepsilon_t$, where the noise term had a mean of zero and a variance of 225. The starting point of these series was chosen so that the last data point was a) close to the vertical mid-point of the screen and b) one third of the way from the mid-point of the seasonal cycle towards its peak (Figure 3.1). Each wavelength phase lasted for 12 time periods. There were 3.25 wavelengths in the screen. Each wavelength's width corresponded to a 30% of the screen width. Time series were generated uniquely for each participant and the four types of series were randomly ordered separately for each of them. The task was not performed within a particular scenario, such as one associated with sales forecasting, to avoid introduction of frame-specific biases (e.g., elevation biases arising from optimism or perceived control effects). Hence, the vertical axes of the graphs used to present the series were unlabelled. Series were presented as line graphs. After the end of each series, five vertical lines were presented in the next five time periods to indicate where forecasts had to be made. When a forecast was made by clicking on one of the vertical lines a blue dot, appeared in the position of the cursor when the mouse was clicked.

Procedure

Each participant performed the task individually on a computer in a separate cubicle. They read a short introduction to the study and then entered their demographic details (age, sex). They were instructed to view each series and then click on each of the vertical lines to show where they expected future

points in the series to appear. Before starting, they were told the order in which they had to make their forecasts. However, the task was constrained to ensure that their forecasts were actually made in this order. Thus the programme would accept a forecast (and show a blue dot to indicate its position) only if it was made in the required order. For participants in the no end-anchoring group, all five vertical lines were presented at the same time as the series. Forecasts were made from the nearest horizon to the most distant one in the order 1, 2, 3, 4, 5. As forecasts were made, a blue line linked each new forecast with the last data point (forecast for horizon 1) or with the immediately preceding forecast (all other forecasts). For participants in the backwards sub-group of the end-anchoring group, all five vertical lines were again presented at the same time as the series. However, an explicit screen message prompted participants to make their forecasts backwards (in the order 5, 4, 3, 2, 1). As forecasts were made, a blue line linked each new forecast with its predecessor. Participants in the forwards sub-group of the end-anchoring group first made their forecast for the most distant horizon for each of the four series. Thus, initially, only a single vertical line at furthest horizon was presented with each series to signal that only the forecast for that horizon was required. Once that forecast had been made for all series, participants returned to each one (presented in the same order as before) to make the remaining four required forecasts working forward from the end of the data series. To enable them to do this, the remaining four vertical lines appeared on the screen at this point to indicate the positions of these required forecasts. Thus forecasts were made in the order 5, 1, 2, 3, 4. As forecasts were made, a blue line linked each new forecast with the last data

point (forecast for horizon 1), with the previous forecast (forecasts for horizons 2 and 3), or with both the immediately preceding forecast and the forecast for the most distant horizon (forecast for horizon 4).

3.1.2 Results

Participants whose forecasts were at least 3 inter-quartile ranges from the median of each group were excluded. This resulted in a total of 116 participants, 58 in each of the two conditions. To test H_1 and H_2, I compared mean absolute error (MAE) between Group 1 (no end-anchoring) and Group 2 (end-anchoring). To cast more light on the effects of end-anchoring, I also report some supplementary analyses. Then, to test H_3 and H_4, I compare MAE between Group 2a (forward forecasting after end anchoring) and Group 2b (backward forecasting after end-anchoring). Again, I also report supplementary analyses.

Effects of end-anchoring Graphs of MAE in the two conditions are shown in Figure 3.2 for each of the four series types. They show accuracy decreasing with increasing horizon and the decrease appears to be higher in the no end-anchoring group for seasonal, linear trended, and autoregressive series. To examine the significance of these effects, I carried out separate two-way analyses of variance (ANOVA) on the MAE data for each series type. Here and later, Huynh-Feldt corrections were applied to address violations of sphericity.

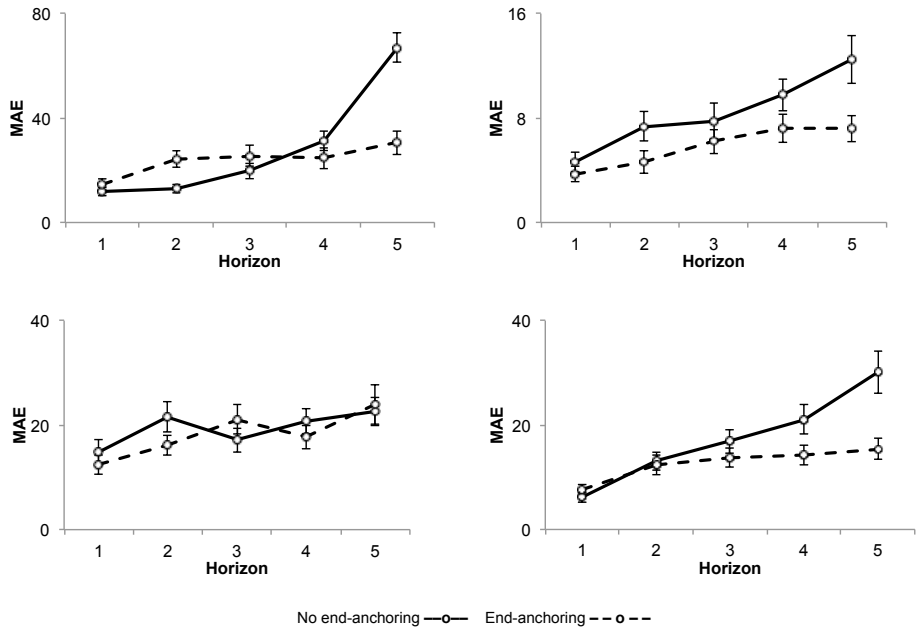

Figure 3.2 Graphs of mean values of absolute error (together with standard error bars) in the no end-anchoring group (continuous lines) and in the end-anchoring group (dashed lines) for seasonally trended, linearly trended, autocorrelated, and random series (clockwise from top left).

For seasonal series, there was an effect of horizon (F (2.40, 273.45) = 74.31; p < .001), and analysis using polynomial contrasts showed that it contained linear, quadratic and cubic components. There was also an effect of end-anchoring (F (1, 114) = 4.72; p < .05) and an interaction between that variable and horizon (F (2.4, 273.45) = 35.67; p < .001). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 2 (F (1, 114) = 19.16; p < .001) and horizon 5 (F (1, 114) = 52.52; p < .001). For the linear trended series, there was an effect of horizon (F (2.69, 307.20) = 24.66; p < .001), with only the linear component significant in an analysis using polynomial contrasts. There

was also an effect of end-anchoring (F (1, 114) = 10.16; $p < .01$) and an interaction between that variable and horizon (F (2.695, 307.20) = 3.49; $p < .05$). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 2 (F (1, 114) = 7.21; $p < .01$), horizon 4 (F (1, 114) = 5.08; $p < .05$), and horizon 5 (F (1, 114) = 12.23; $p < .001$). For the autocorrelated series, there was again an effect of horizon (F (2.13, 243.09) = 62.13; $p < .001$), with only the linear component significant in an analysis using polynomial contrasts. There was also an effect of end-anchoring (F (1, 114) = 7.25; $p < .01$) and an interaction between that variable and horizon (F (2.13, 243.09) = 18.17; $p < .001$). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 4 (F (1, 114) = 7.99; $p < .01$) and horizon 5 (F (1, 114) = 21.00; $p < .001$). For the random series, only the effect of horizon was significant (F (3.92, 447.04) = 7.53; $p < .001$). As the interaction was not significant, the effects of this variable were not analysed in each of the groups separately. For all series that contain a pattern as well as noise, these analyses are consistent with the first hypothesis (H_1) that end-anchoring improves the accuracy of the forecast for the most distant horizon: in each case, the simple effect of group was significant for horizon 5. Other aspects of the results are consistent with the second hypothesis (H_2) that end-anchoring also improves accuracy of forecasts for less distant horizons: significant interactions showed that the linear increase in MAE with horizon was faster in the no end-anchoring group and significant simple effects of group occurred for horizons 2 and 4 in the linear trended series and for horizon 4 in the autocorrelated series.

I now report two supplementary analyses designed to throw light on the reasons for these effects. The first analysis is of Mean Signed Errors (MSE) for each series type (Figure 3.3). Signed errors are calculated as actual forecast minus optimal forecast. Hence, the increasing signed error for forecasting the downward section of the seasonal series and the decreasing signed error for forecasting the upward sloping linear trended series are both evidence of trend damping. It is immediately apparent from Figure 3.3 that one effect of end-anchoring is to reduce trend damping.

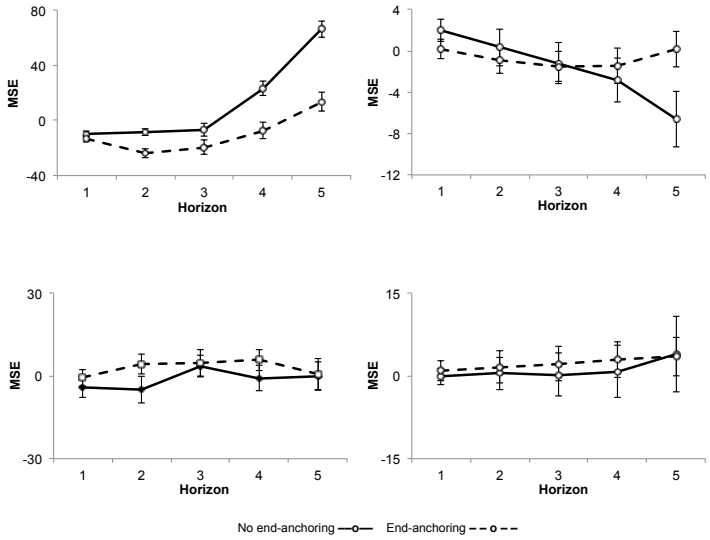

Figure 3.3 Graphs of mean values of signed error (together with standard error bars) in the no end-anchoring group (continuous lines) and in the end-anchoring group (dashed lines) for seasonally trended, linearly trended, autocorrelated, and random series (clockwise from top left).

Two-way ANOVAs on MSE confirmed that this was so. For seasonal series, there were significant effects of horizon (F (2.31, 263.58) = 160.62; p < .001),

end-anchoring (F (1, 114) = 48.21); $p < .001$), and an interaction between these variables (F (2.31, 263.58) = 27.41; $p < .001$). For linearly trended series, there were significant effects of horizon (F (2.23, 254.60) = 8.43) and of the interaction between that variable and end-anchoring (F (2.23, 254.60) = 9.20; $p < .001$). In these analyses, the effects of horizon indicate trend damping and the interaction demonstrates that end-anchoring reduces that effect. (ANOVAs on MSE in autocorrelated and random series showed no significant effects.)

The second analysis was designed to investigate the effect of end-anchoring on the slope of the sequence of five forecasts in more detail. I used an approach employed by Harvey and Reimers (2013). Linear regression models were fitted to each one of the four sequences of five forecasts produced by each participant. Thus, for each sequence, I fitted the model: forecast = a + b (horizon) + error. Then, for each series type, t-tests were used to examine whether the constants (a) and trend coefficients (b) in the two conditions differed from one another and whether each of them differed from the optimal values derived from the generating equation. I also tested whether the variance of the coefficients and the error variance in the model were greater in the no end-anchoring group than in the end-anchoring group. Values of coefficients in each condition and in the generating equation and levels of error variance in each condition are shown in Table 3.1 for each series type. This table also indicates the comparisons that reached significance. However, I will highlight the main results here.

Table 3-1 Linear regressions of forecast sequences for each series type: mean values (variances in parentheses) of constants, trend coefficients and residual error variances. Actual values in the generating equations are shown for comparison.

		Constant (*a*)	*Trend* (*b*)	*Error* (*e*)
Seasonal	Actual	270.86	-20.50	
	No end-anchoring	228.69**† (23.48)	-2.14**† (9.71)	149.86
	End-anchoring	240.03**† (19.30)	-13.57**† (9.05)	235.12
Linear	Actual	207	5	
	No end-anchoring	211.45*† (8.60)	2.97*† (3.51) †	14.31
	End-anchoring	206.45† (7.22)	4.94† (2.63) †	16.25
Auto correlated	Actual	150	0	
	No end-anchoring	148.51 (9.78)	0.86 (7.51) †	31.55
	End-anchoring	150.32 (15.02)	0.65 (5.03) †	44.88
Random	Actual	150	0	
	No end-anchoring	144.96 (25.52)	1.24 (7.74)	311.03
	End-anchoring	151.75 (17.25)	0.42 (6.36)	275.73

*Mean value different from that in the generating equation, $p < .05$

**Mean value different from that in the generating equation, $p < .01$

† Values differ between the no end-anchoring and end-anchoring groups, $p < .05$

The mean slope of the forecast sequence was significantly lower in the end-anchoring group than in the no end-anchoring group for seasonally trended (t (114) = 6.557; $p < 0.05$) and linearly trended (t (114) = -3.519; $p < 0.001$) series. This confirms that, where trends are present in the data series, end-anchoring acts to decrease trend damping. Variance of the trend coefficients was significantly lower in the end-anchoring group than in the no end-anchoring group for linearly trended (F (57, 57) = 1.78; $p < .05$) and autocorrelated series (F (57, 57) = 2.22; $p < .05$). (Data for the other two series types are in the same direction but the comparisons did not attain significance). This shows that there was a tendency for end-anchoring to reduce the degree to which the slope of the forecast sequence drifted away from its correct value.

Effects of direction of forecasting One sub-group made forecasts in a forwards sequence after end-anchoring: horizons were forecast in the order 5, 1, 2, 3, 4. The second sub-group made forecasts in a backwards sequence after end anchoring: horizon were forecast in the order 5, 4, 3, 2, 1. Here I test hypotheses H_3 and H_4 by comparing overall forecast error (MAE) in the forwards and backwards sub-groups. Graphs of MAE in the two conditions are shown in Figure 3.4 for each of the four series types. I carried out separate two-way ANOVAs on each of them using horizon as a within-participants variable and condition as a between-participants variable.

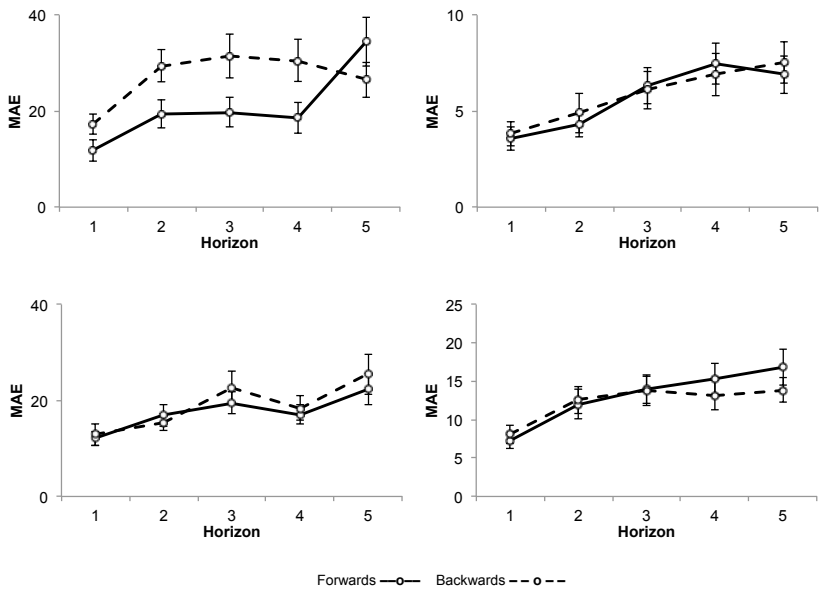

Figure 3.4 Graphs of mean values of absolute error (together with standard error bars) in the forwards forecasting sub-group (continuous lines) and the backwards forecasting sub-group (dashed lines) for seasonally trended, linearly trended, autocorrelated, and random series (clockwise from top left).

For seasonal series, there was an effect of horizon (F (2.89, 162.02) = 7.12; $p <$ .001), and analysis using polynomial contrasts showed that it contained linear and cubic components. There was also a significant interaction between forecast direction and horizon (F (2.89, 162.02) = 3.66; $p < .05$). Tests of simple effects showed that the effect of forecast direction to be significant for horizon 2 (F (1, 56) = 4.95; $p < .05$), horizon 3 (F (1, 56) = 4.45; $p < .05$), and horizon 4 (F (1, 56) = 4.95; $p < .05$). The other three series types showed effects only of horizon: linear trended (F (3.16, 177.13) = 7.83; $p < .001$); autocorrelated (F (2.98, 166.97) = 9.62; $p < .001$); random (F (3.37, 188.47) = 6.40; $p < .001$). In

all three cases, analysis using polynomial contrasts showed that only the linear components of these effects were significant. Thus, results for seasonal series are consistent with the third hypothesis: effects of direction of forecasting affected accuracy for that series type. Furthermore, the results as a whole are consistent with the fourth hypothesis: effects of direction of forecasting depended on series type.

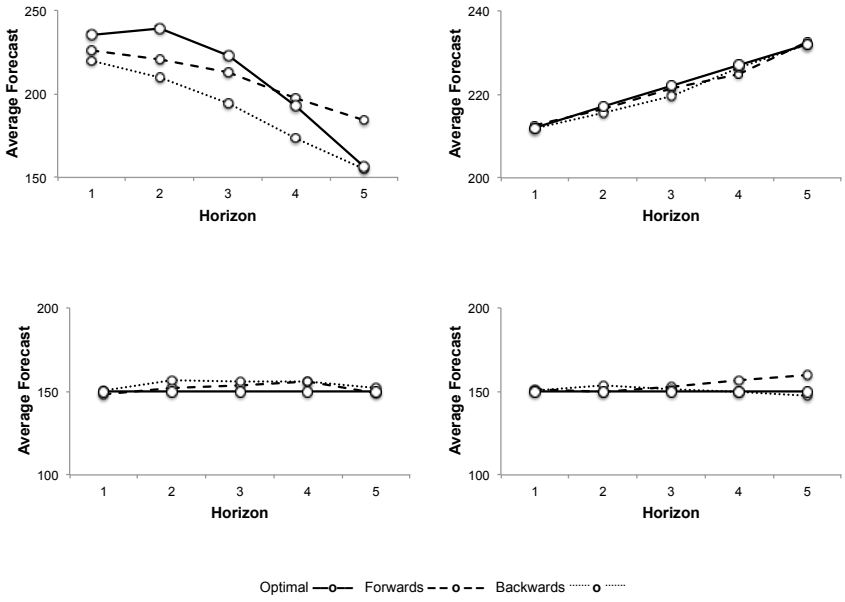

Figure 3.5 Graphs showing optimal forecasts (continuous lines) and participants' mean forecasts in the forwards forecasting sub-group (dashed lines) and the backwards forecasting sub-group (dotted lines) for seasonally trended, linearly trended, autocorrelated, and random series (clockwise from top left).

Why were only seasonal series influenced by direction of forecasting? For the other three types of series, reasonably good forecasts could be made by making linear interpolations between the last data point and the forecast already made for the most distant horizon. This was because the sequence of outcomes that required forecasting was linear. In contrast, the sequence of outcomes that had to be forecast in the seasonal series was non-linear: its values first increased markedly and then decreased. However, forecast sequences did not show this pattern: the value of the first two forecasts stayed close to that of the last data point and values of later forecasts then decreased more slowly than the values of the outcomes to be forecast. In other words, the forecast sequence did not show such a clear point of inflection as the sequence of outcomes that had to be forecast: it was more linear than it should have been. These patterns are shown in Figure 3.5. These impressions were confirmed in regression analyses. Using a step-up procedure, I found that 40 of the 58 participants' forecasts showed significant linear components. In those cases, the linear models explained an average of 86% of the variance. Adding a quadratic component significantly increased the variance explained by the model in only three of these 40 participants and, on average, it explained only an additional 10% of the variance. Also, comparing the 40 models (with quadratic components included) to a model of the outcomes to be forecast (produced by continuing the generating function) showed that the coefficient for the quadratic component was significantly lower in the participants' forecast sequences (t (39) = 6.10; p < .001) than in the sequence of outcomes. These analyses imply that the interpolations that participants made between the last data point and the forecast

that they had already produced for the most distant horizon were more linear than they should have been.

As Figure 3.5 shows, the near-linearity of participants' sequence of forecasts reflected their failure to increase the value of their first few forecasts above the value of the last data point. As a result, their forecasts were too low – and the extent to which they were too low was greater in backwards forecasting group than in the forwards forecasting group. A two-way ANOVA on MSE confirmed this. It showed a significant effect of forecast horizon (F (2.45, 136.96) = 27.21; $p < .001$) and an analysis using polynomial contrasts indicated that it had linear and quadratic components. There was also a significant effect of forecasting direction (F (1, 56) = 16.94); $p < .001$) and a marginally significant interaction between the two variables (F (2.45, 136.96) = 2.78; $p = .055$). Tests for simple effects showed that the effect of forecasting direction was significant at horizon 2 (F (1, 56) = 5.84; $p < .05$), horizon 3 (F (1, 56) = 7.90; $p < .01$), horizon 4 (F (1, 56) = 9.68; $p < .01$) and horizon 5 (F (1, 56) = 10.68; $p < .01$).

Why did this pattern of results occur? It appears that participants forecasting in a backwards direction anchored their judgments on the low value of the forecast that they had already made for the most distant horizon. As a result, although they then increased the value of their forecasts for horizons 4, 3, and 2, they did so insufficiently. Their forecast for horizon 1 was then made by linearly interpolating between their forecast for horizon 2 and the last data point. In contrast, those forecasting in a forward direction anchored on the relatively high

value of the last data point. As a result, their forecasts for horizons 1, 2, and 3 were made at the same level as that point. Then, their forecast for horizon 4 was made by linearly interpolating between their forecast for horizon 3 and the forecast that they had made earlier for horizon 5. Thus when participants had to forecast a nonlinear sequence of four outcomes between the end of the data series and a forecast that they had earlier made for the most distant horizon, they used an initial strategy based on anchoring to make the first three of those forecasts and then made the final forecast by linear interpolation. This produced different levels of accuracy for backwards and forwards forecasting.

Discussion

The experiment showed that, when the data series contain a pattern, judgmental forecasts for a sequence of outcomes can be improved by making the forecast for the most distant horizon first. It also showed that, when that is done, the order in which the remaining forecasts are made does not matter if the sequence of outcomes that require forecasting lie in a straight line. However, if they contain some other (i.e. nonlinear) pattern, accuracy of forecasts made in a forwards direction (from the last data point towards the previously produced forecast for the furthest horizon) can differ from forecasts made in a backwards direction (from the previously produced forecast for the furthest horizon towards the last data point). I shall discuss these findings in turn.

End-anchoring Making the forecast for the furthest horizon first clearly improved accuracy not only of that forecast but of other forecasts too. This result was expected. I found that it occurred for two reasons. First, as the regression analyses showed, the trajectory of the forecast sequence became less variable. This was expected this because, without end-anchoring, each forecast after that for the first horizon is based on its noisy predecessor but is not constrained by a forecast for a more distant horizon: the task is one of extrapolation. In contrast, end-anchoring constrains the forecast trajectory: the task is one of interpolation. Second, as the analyses of MSE and regressions show, end-anchoring reduced trend damping. This finding was not expected. It might have occurred because participants found forecasting in the end-anchoring condition more difficult. As a result, they devoted more cognitive resources to the task and performed it better. To check this account, I compared the mean time taken to make the first forecast in the two groups. This analysis showed that it was less in the no end-anchoring group (4.37 seconds) than in the end-anchoring group (6.96 seconds) ($t(221.79) = 12.35$; $p < .001$). I also compared the time to make all five forecasts in the no end-anchoring group (9.60 seconds) with backwards forecasting sub-group of the end-anchoring group (13.69 seconds) and found it to be greater in the latter case ($t(65.46) = 7.29$; $p < .001$). These two analyses confirm that forecasters devoted more cognitive resources to their task in the end-anchoring condition. This finding can be interpreted in terms of models that posit different modes of cognitive processing: an intuitive system that acts rapidly, heuristically, non-consciously, and with little effort and a deliberative system that acts slowly, analytically,

consciously, and with effort (Kahneman, 2011). Thus forecasting in the normal way from the data series for increasingly distant horizons may be an intuitive process that relies on anchoring heuristics and produces 'biases' such as trend damping. In contrast, making forecasts for the most distant horizon first is likely to be a slower, more cognitively demanding, deliberative process that is less susceptible to the sort of biases produced by heuristic processing.

Direction of forecasting After end-anchoring, the forecasting task was transformed from one of extrapolation to one of interpolation. When linear interpolation was appropriate (random, autocorrelated or linearly trended series), there was no difference in accuracy between interpolating forward from the end of the data series towards the anchor provided by the forecast for the most distant horizon and interpolating backwards from that anchor towards the end of the data series. However, when linear interpolation was not appropriate (seasonal series), interpolating backwards produced higher levels of error than forecasting forwards. The reason for this appears to be that people adopted different strategies for forecasting in the two cases. The section of the seasonal series that had to be forecast comprised the peak of a cycle followed by a descending segment (Figure 3.5). When forecasting backwards, the descending segment became an ascending one and was forecast in the same way as a linear trend. For the first three forecasts they made (horizons 4, 3, and 2), participants anchored on the forecast that they had made immediately before and then adjusted upwards to take the trend into account. As there adjustments were insufficient, some trend damping was observed (Figure 3.5). Then they made

their final forecast for horizon 1 by linearly interpolating between their previous forecast for horizon 2 and the last data point. When forecasting forwards, participants approximated the peak of the seasonal series as an untrended linear series and forecast it as if it were one. Thus, their forecasts for horizons 1, 2, and 3 were forecast at the same level as the last point of the data series. Then they made their final forecast for horizon 4 by linearly interpolating between the forecast that they had just made for horizon 3 and the forecast that they had made earlier for horizon 5. This strategy for forecasting was, unlike the one for backwards forecasting, not subject to trend damping: it therefore produced forecasts that were higher and closer to the target outcome series (Figure 3.5).

3.2 Order effects and noise levels in judgmental forecasting (Experimental Study 2)

In this experiment, I examine the effects of a) increasing the level of noise in the data series and b) changing the phase of the seasonal series so that the sequence of outcomes that had to be forecast was approximately linear rather than nonlinear. Increasing noise in the data series is likely to impair forecasting performance. However, there are two reasons that higher noise levels should increase (or, at least, preserve) the effects of end-anchoring. First, higher levels of noise in series produce greater trend damping effects (Eggleton, 1982; Harvey and Bolger, 1996). Hence, a manipulation that removes (or greatly reduces) trend damping should improve accuracy more when series noise is higher. Second, when data series are noisier, a sequence of forecasts made via

forward extrapolation is likely to deviate more from the correct trajectory. This is because forecasts are made by using immediately preceding forecasts as anchors and those forecasts contain more noise when series are noisier (Harvey, 1995). Hence, a manipulation that changes the task from one of extrapolation to one of interpolation should reduce the variance in participants' forecast trajectories even more (and improve their accuracy even more) when noise in the data series is higher. Hence, I test the following hypothesis.

H_5: Higher levels of noise will depress forecasting performance but preserve or even enhance effects of end-anchoring

Requiring people to forecast an approximately linear section of the seasonal series should eliminate the difference between backwards and forwards forecasting sub-groups of the end-anchoring condition. This is because linear interpolation, forecasters' default strategy after end-anchoring, would be as appropriate as it is for linearly trended or untrended autocorrelated series. I would expect it to be used irrespective of forecasting direction. Hence the higher levels of MSE that are observed for backwards forecasting from seasonal series in Experiment 1 should no longer be obtained. Consequently, a cross-experiment comparison on seasonal series should reveal a significant interaction between forecasting direction (forwards versus backwards) and experiment (Experiment 1 versus Experiment 2).

H_6: Requiring participants to forecast a linear rather than a nonlinear sequence of outcomes will eliminate the effect of forecasting direction on MSE and this

will produce a significant interaction between forecasting direction (forwards versus backwards) and experiment (Experiment 1 versus Experiment 2).

3.2.1 Method

Participants

Participants comprised 120 students (57 men, 63 women) drawn from the same pool as before. Their mean age was 28 years. They were paid £1.00 for their participation.

Design

As the end-anchoring effect did not occur when there was no pattern in the data, I excluded random series from this experiment. In all other respects, the design was identical to that outlined for Experiment 1.

Stimulus materials

For the seasonally trended series, the amplitude of the seasonal variation was doubled: the equation used to generate these series was therefore $X_t = 140\cos(100t) +170+ \varepsilon$. Also the variance of the noise component was increased by a factor of four to 900. The starting point of these series was chosen so that the last data point was a) close to the vertical mid-point of the screen and b) at the peak of the seasonal cycle (Figure 3.6). The linearly trended series and the autocorrelated series were generated in the same way as in Experiment 1, except that the variance of the noise was increased by four times to a value of 120 in the former case and to a value of 76 in the latter one.

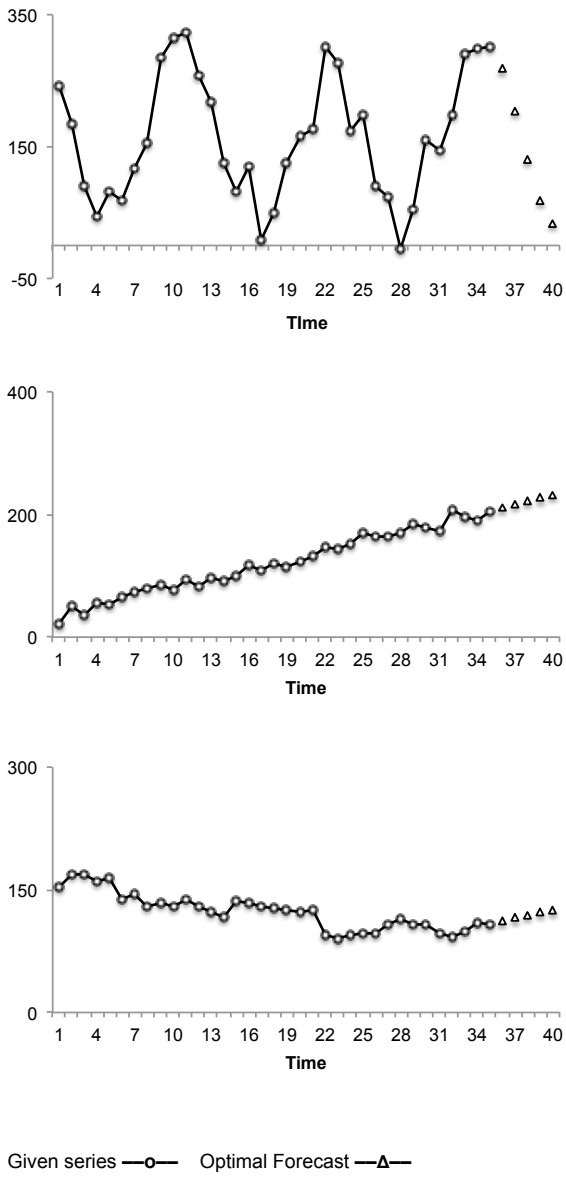

Figure 3.6 Examples of the three types of series, showing 35 data points (seen by participants) followed by five optimal forecasts (not seen by participants) for seasonally trended (top panel), linearly trended (middle panel) and autocorrelated series (lower panel).

Procedure

The procedure was identical to that used in Experiment 1.

3.2.2 Results

To test H_5, I compare MAE in the no end-anchoring and end-anchoring groups and then compare the effect of this variable in this experiment with the effect it had in Experiment 1. To test H_6, I compare MAE in the forward forecasting and backward forecasting sub-groups of the end-anchoring group and then examine whether the effects of direction of forecasting are different in this experiment from those in the previous one.

Effects of end-anchoring Graphs of MAE in the two conditions are shown in Figure 3.7 for each of the three series types. They show accuracy decreasing with increasing horizon and the decrease again appears to be higher in the no end-anchoring group for seasonal, linear trended, and autoregressive series. To examine the significance of these effects, I carried out separate two-way analyses of variance (ANOVA) on the MAE data for each series type.

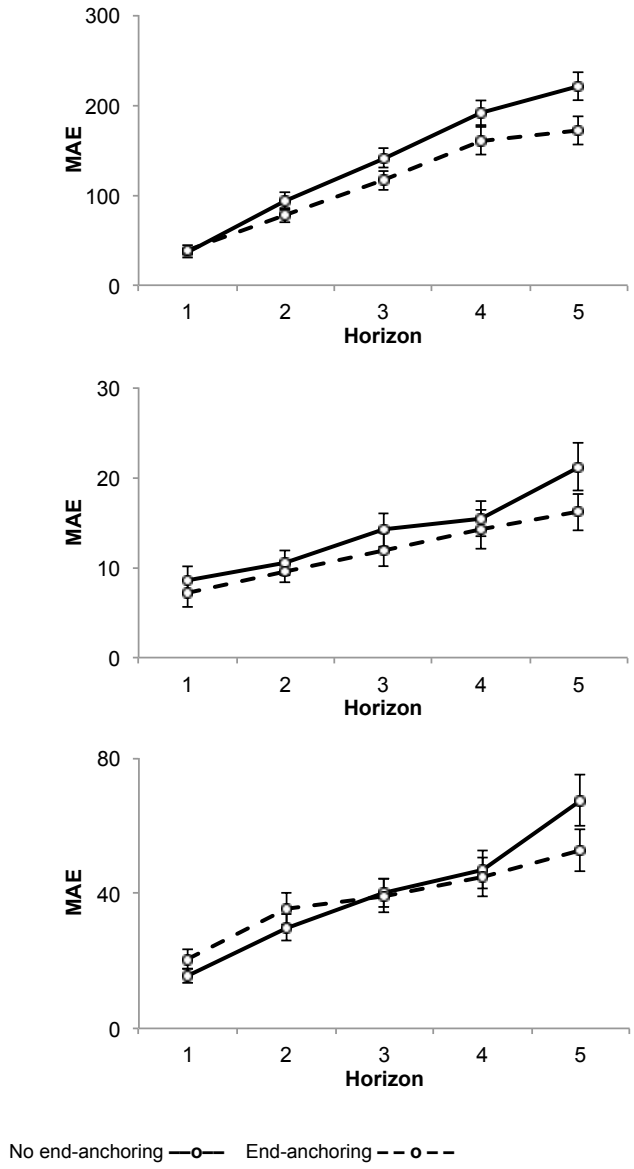

Figure 3.7 Graphs of mean values of absolute error (together with standard error bars) in the no end-anchoring group (continuous lines) and in the end-anchoring group (dashed lines) for seasonally trended (top panel), linearly trended (middle panel) and autocorrelated series (lower panel).

For seasonal series, there was an effect of horizon (F (2.69, 306.71) = 213.55; p < .001), and analysis using polynomial contrasts showed that it contained linear components. There was also an effect of end-anchoring (F (1, 114) = 7.54; p = .007) and an interaction between that variable and horizon (F (2.69, 306.71) = 4.65; p = .005). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 3 (F (1, 114) = 4.94; p < .05), for horizon 4 (F (1, 114) = 4.83; p < .05) and horizon 5 (F (1, 114) = 9.48; p < .05). For the linear trended series, there was an effect of horizon (F (3.25, 370.32) = 27.98; p < .001), with only the linear component significant in an analysis using polynomial contrasts. There was also a marginally significant effect of end-anchoring (F (1, 114) = 3.45; p = .066) and no interaction between that variable and horizon. Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 5 (F (1, 114) = 4.64; p < .05). For the autocorrelated series, there was again an effect of horizon (F (1.99, 227.25) = 75.39; p < .001), with only the linear component significant in an analysis using polynomial contrasts. There was no effect of end-anchoring but there was an interaction between that variable and horizon (F (1.99, 227.25) = 5.14; p < .05). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 5 (F (1, 114) = 4.39; p < .05). These analyses are consistent with hypothesis H_1 that end-anchoring improves the accuracy of the forecast for the most distant horizon: in each case, the simple effect of group was significant for horizon 5. For seasonal series, end-anchoring also improved accuracy of forecasts for less distant horizons (H_2).

To throw light on the reasons for these effects, I now report the same two supplementary analyses that I carried out for Experiment 1. The first analysis was carried out on MSE (again calculated as actual forecast minus optimal forecast) for each series type (Figure 3.8). The increasing signed error for forecasting the downward section of the seasonal series and the decreasing signed error for forecasting the upward sloping linear trended series are both evidence of trend damping. It is clear that end-anchoring again acted to reduce trend-damping in these series (Figure 3.8).

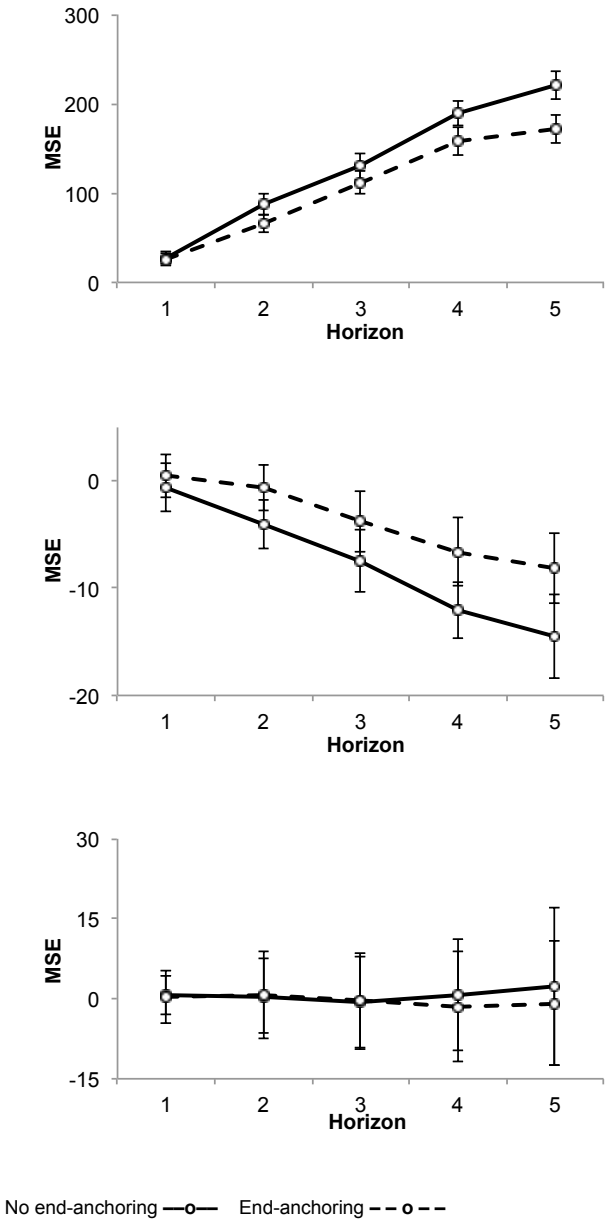

Figure 3.8 Graphs of mean values of signed error (together with standard error bars) in the no end-anchoring group (continuous lines) and in the end-anchoring group (dashed lines) for seasonally trended (top panel), linearly trended (middle panel) and autocorrelated series (lower panel).

Table 3-2 Linear regressions of forecast sequences for each series type: mean values (variances in parentheses) of constants, trend coefficients and residual error variances. Actual values in the generating equations are shown for comparison.

		Constant (a)	*Trend (b)*	*Error (e)*
Seasonal	Actual	322.14	-60.49	
	No end-anchoring	305.71**† (57.67)	-11.38**† (22.81)	1443.71
	End-anchoring	313.41**† (58.77)	-22.56**† (22.06)	1155.03
Linear	Actual	207	5	
	No end-anchoring	210.04 (12.76)	1.42** (4.49)	125.11
	End-anchoring	210.20* (11.79)	2.68** (4.16)	95.74
Autocorrelated	Actual	150	0	
	No end-anchoring	149.68 (20.83)	0.34 (15.88)	144.16
	End-anchoring	151.02 (35.75)	-0.46 (13.71)	260.35

*Mean value different from that in the generating equation, $p < .05$

**Mean value different from that in the generating equation, $p < .01$

† Values differ between the no end-anchoring and end-anchoring groups, $p < .05$

Two-way ANOVAs on MSE confirmed that this was so for two of the three series types. For seasonal series, there was an effect of horizon (F (2.97, 338.32) = 221.07; $p < .001$), and analysis using polynomial contrasts showed that it contained linear components. There was also an effect of end-anchoring (F (1, 114) = 5.83; $p = .017$) and an interaction between that variable and horizon indicated that trend damping was reduced by end-anchoring (F (2.97,

338.32) = 3.57; p = .015). Tests of simple effects showed that the effect of end-anchoring to be significant for horizon 4 (F (1, 114) = 4.29; $p < .05$) and horizon 5 (F (1, 114) = 9.475; p = .003). For the linear trended series, there was an effect of horizon (F (3.63, 414.28) = 15.86; $p < .001$), with only the linear component significant in an analysis using polynomial contrasts. There was also an effect of end-anchoring (F (1, 114) = 4.57; $p < .05$) but no interaction between that variable and horizon. Tests of simple effects showed that the effect of end-anchoring to be marginally significant for horizon 4 (F (1, 114) = 3.38; p = .068), and horizon 5 (F (1, 114) = 3.09; p = .08). For the autocorrelated series, neither the main effects of horizon and end-anchoring nor the interaction between them were significant.

To carry out the second analysis, regression models were again fitted to each one of the four sequences of five forecasts produced by each participant. As before, for each sequence, I fitted the model: forecast = a + b (horizon) + error. Mean values of constants, trend coefficients and residual variance in each condition, together with optimal values derived from the generating equations are shown in Table 3.2. Also shown is the significance of statistical comparisons between the two groups and between each of them and the values in the generating equations. For seasonal series, there was evidence that trend damping was reduced in the end-anchoring condition. In other words, the mean absolute value of the linear trend coefficient (b) was significantly lower in participants' forecast sequences in both conditions than in the generating equation and also lower in forecast sequences in the no end-anchoring condition

than in the end-anchoring condition (Table 3.2). The effect did not reach significance for the linearly trended series. Also, in this experiment, there was no statistical evidence that the variability of b coefficients was greater in the no end-anchoring condition though, for all three series types, the difference across conditions was numerically in that direction.

Effects of direction of forecasting Figure 3.9 shows MAE scores for each of the three series types. Separate two-way ANOVAs on each series type, using horizon as a within-participants variable and forecasting direction as a between-participants variable, showed an effect of horizon for seasonally trended series (F (2.57, 143.72) = 1.96; $p < .001$), linearly trended series (F (3.53, 197.54) = 11.96; $p < .001$), and autocorrelated series (F (2.13, 119.43) = 23.93; $p < .001$). Analyses using polynomial contrasts revealed that, in all cases, these effects contained only linear components. Effects of forecasting direction and the interactions between this variable and horizon did not reach significance for any series type. Trend damping contributed to the effects of horizon on MAE for seasonally trended and linearly trended series (Figure 3.10). Thus two-way ANOVAs showed effects of this variable on MSE in the seasonally trended series (F (2.78, 155.79) = 93.25; $p < .001$) and in the linearly trended series (F (3.98, 223.29) = 5.56; $p < .001$) but not in the autocorrelated series. Analyses using polynomial contrasts showed that the significant effects in the trended series contained only linear components. Effects of forecasting direction and the interactions between this variable and horizon did not reach significance for any series type.

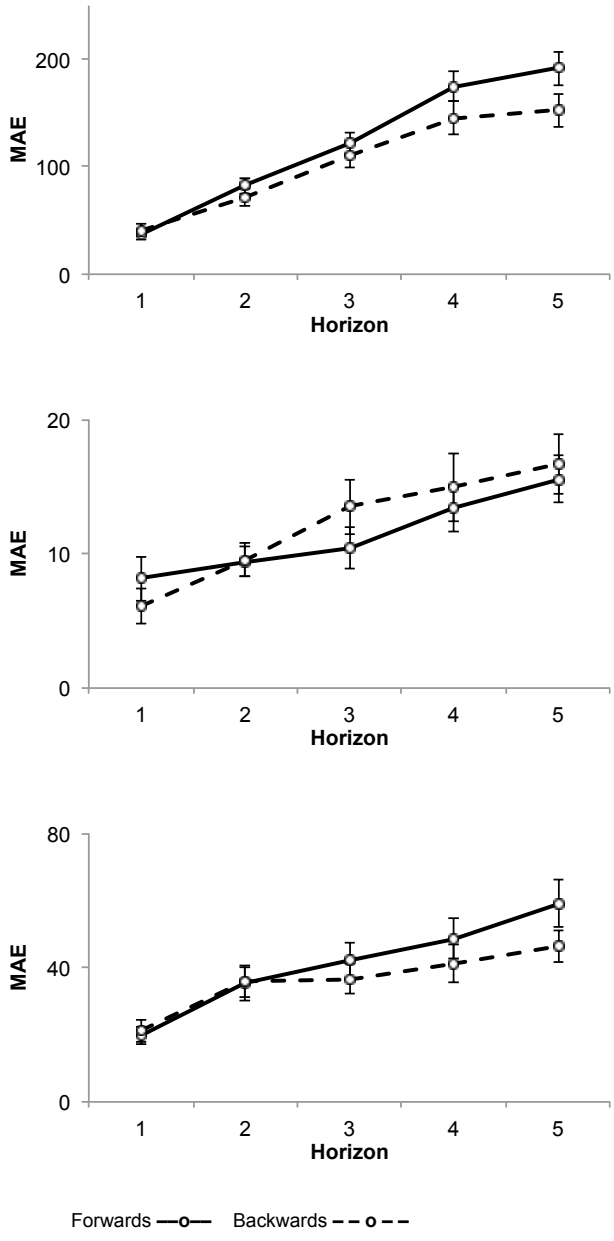

Figure 3.9 Graphs of mean values of absolute error (together with standard error bars) in the forwards forecasting sub-group (continuous lines) and the backwards forecasting sub-group (dashed lines) for seasonally trended (top panel), linearly trended (middle panel) and autocorrelated series (lower panel).

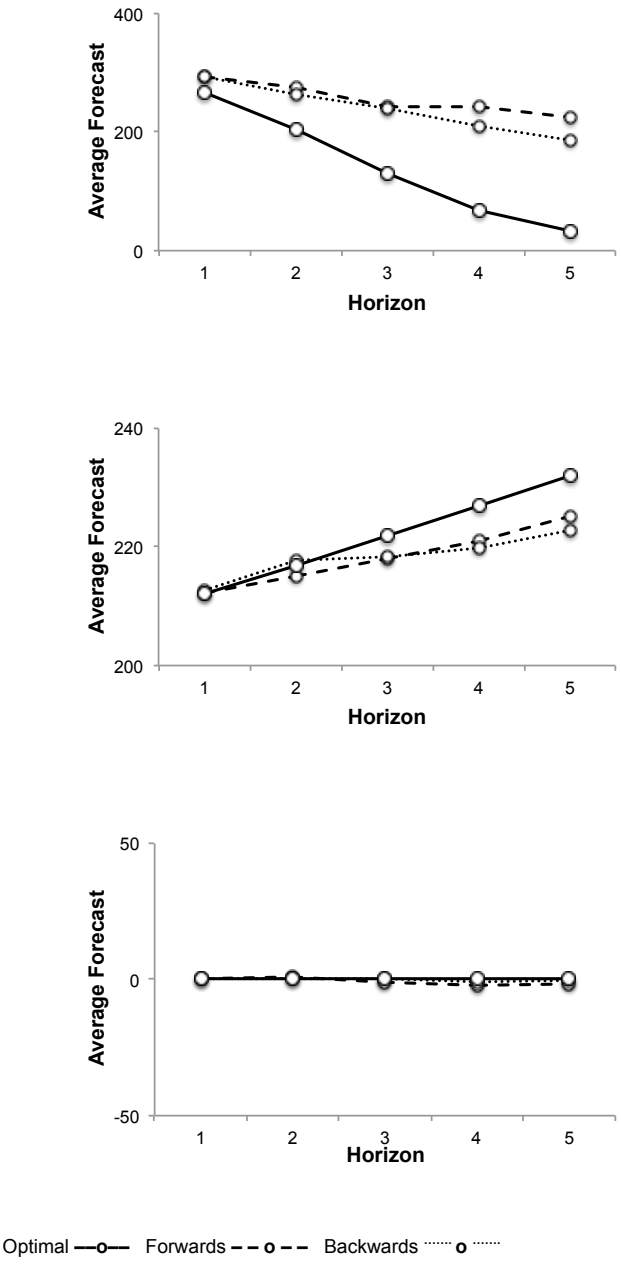

Figure 3.10 Graphs showing optimal forecasts (continuous lines) and participants' mean forecasts in the forwards forecasting sub-group (dashed lines) and the backwards forecasting sub-group (dotted lines) for seasonally trended (top panel), linearly trended (middle panel) and autocorrelated series (lower panel).

In Experiment 1, backward forecasting of seasonal series produced higher MAE scores than forward forecasting. I showed that this occurred because participants' strategy for backwards forecasting of the highly nonlinear sequence of outcomes was different from their strategy for forwards forecasting of that sequence. In particular, backwards forecasting produced lower forecasts and, hence, higher MSE scores. In contrast, this experiment showed no effect of forecasting direction on MAE for seasonal series. This was because participants' linear interpolation strategy for forecasting the near linear sequence of outcomes was appropriate and the same for both backwards and forwards conditions. As a result, MSE scores were no higher when participants were forecasting backwards than when they were forecasting forwards.

Cross-experiment comparisons

According to H_5, higher levels of noise depress forecasting performance but preserve or even enhance effects of end-anchoring. Separate three-way ANOVAs were performed on MAE for seasonal, trended and autocorrelated series using experiment (Experiment 1 versus Experiment 2) and end anchoring as between-participants variables and horizon as a within-participants variable. These confirmed that, in this second experiment, forecasting was not only worse (Seasonal series: F (1, 228) = 497.21; $p < .001$; Trended series: F (1, 228) = 64.22; $p < .001$; Autocorrelated series: F (1, 228) = 125.70; $p < .001$) but also deteriorated more with increasing horizon (Seasonal series: F (2.72, 619.76) = 128.03; $p < 0.001$; Trended series: F (3.17, 724.31) = 5.15; $p < 0.001$; Autocorrelated series: F (2.01, 458.31) = 25.02; $p < 0.001$). However, the size

of the end anchoring effect was preserved for trended and autocorrelated series (i.e. there was no interaction between this variable and experiment) and, for seasonal series, it was larger in the present experiment than in the previous one (F (1, 228) = 4.61; p = .033). All these results are consistent with H_5. I carried out a three-way ANOVA comparing MSE in forecasts from seasonal series in the two sub-conditions of the end-anchoring condition across experiments. Thus this analysis used experiment (Experiment 1 versus Experiment 2) and forecasting direction (backwards versus forwards) as between-participants variables and horizon as a within-participants variable. It revealed that the interaction between experiment and forecasting direction was significant (F (2.547, 285.309) = 3.816; $p < 0.025$). This finding is consistent with H_6.

Discussion

As expected, increasing the noise in the data series impaired forecasting and this impairment was greater for more distant horizons. This additional noise also resulted in effects of end-anchoring being only marginally significant for linear series. However, the more powerful cross-experiment comparison showed that the effect of end-anchoring was either maintained (linearly trended and untrended autocorrelated series) or magnified (seasonal series). End-anchoring had its effect by reducing trend damping effects (just as it did in Experiment 1). These effects tend to be greater with noisier series (Harvey and Reimers, 2013) and, as a comparison of Tables 3.1 and 3.2 shows, the b coefficient in forecast sequences underestimated the coefficient in the continuation of the data series by a larger amount here than in Experiment 1.

Hence, the greater effect of end-anchoring on seasonal series in this experiment may be attributed to the fact that there was more trend damping to be reduced in this experiment. In this experiment, direction of forecasting after end-anchoring did not affect forecast accuracy. This is in accord with H_6. It indicates that the original effect found in Experiment 1 arose because the section of the seasonal series that required forecasting was strongly non-linear. In this experiment where the section of the seasonal series that required forecasting was close to linear, no effect of direction of forecasting was obtained. In other words, the original effect was not caused by the type of series represented by the data (seasonally rather than linearly trended) but by the characteristics (linear or nonlinear) of the ideal forecast sequence. In seasonal series, these characteristics depend both on the phase of the seasonal cycle at which forecasting must start and on the length of the forecast sequence.

3.3 Summary and General Discussion

In the current chapter, I examined the influence of order on forecasting accuracy and corresponding anchoring and adjustment processes. A primary aim was to investigate the effects of end-anchoring. It was anticipated that it would lead to improvements in the accuracy of judgmental forecasts. It is well-known that people add noise to their forecasts (Harvey, 1995) and, when making a sequence of forecasts in order from nearest to most distant horizon, they use their previous forecast as a mental anchor (Bolger and Harvey, 1993). As a result of these two phenomena, a sequence of forecasts may be akin to a

random walk and drift away its original trajectory. By requiring the most distant horizon to be forecast first, my aim was to eliminate this drift. The first experiment did indeed show that end-anchoring reduced the variability across participants of the trajectories of forecast sequences made from the same underlying pattern.

However, the end-anchoring manipulation also had another effect. It made the mean trajectory of the forecast sequence more appropriate. This is because it reduced trend damping. I suggested that this was a response to forecasters finding their task more difficult. Making an initial forecast for five periods ahead is more challenging than making an initial forecast for one period ahead. Kahneman (1973) has argued that people cope with increased difficulty by allocating more cognitive resources to their task; for example, they may switch from using a rapid, heuristic, non-conscious, intuitive mode of processing to a slower, more analytic, conscious, deliberative mode of processing (Kahneman, 2011). The latter approach, though slower, tends to be more accurate. I suggested that end-anchoring improves accuracy because it results in more cognitive resources being devoted to the forecasting task (perhaps via a change from intuitive to deliberative processing). In support of this account, I demonstrated that initial forecasts took over fifty percent longer to produce in the end-anchoring group than in the no end-anchoring group. In the second experiment used noisier data series. Forecasts were worse, showed greater trend damping, and deteriorated more rapidly as the forecast horizon increased. However, end-anchoring still decreased trend damping and, therefore, increased

forecast accuracy for more distant horizons. In this experiment, variability of the forecast trajectories across participants was not significantly reduced by end-anchoring. Noisier data series produce noisier forecasts, which, in turn, reduce the likelihood of real effects attaining significance.

The experiments had a secondary aim. This was to investigate the effects of the direction in which forecasts were made after end-anchoring. There were plausible reasons to expect such a manipulation to have an effect on forecast accuracy (these are analysed in the introduction of the current Chapter), though it was recognised that the nature of any effect was likely to depend on series type. Thus, for different types of series, I compared the accuracy of forecasts made in the order 54321 with that of those made in the order 51234. In fact, results showed that the effect of forecast direction depended not on the type of series from which forecasts were made but on whether the ideal sequence of forecasts was linear or nonlinear. Forecast direction had an effect on accuracy only when that sequence was strongly nonlinear. In this case, forecasting backwards from the end-anchor (54321) produced higher levels of error than forecasting forwards towards the end-anchor (51234). This result could be explained by assuming that participants produced their first three forecasts after the end anchor by using (imperfect) extrapolation and then produced their final forecast by linear interpolation.

Limitations

The recommendations outlined above are only relevant when forecasters produce at least four or five forecasts from each data series. Advantages in

terms of accuracy generally increase as forecast horizon extends into the future; accuracy for close horizons is unaffected by changes in forecast order. There has to be some pattern in the data series for order of forecasting to influence accuracy. If forecasting merely requires mental extraction of the mean of an untrended random series, there is no advantage to be gained by end-anchoring (Experiment 1). Although costs of end-anchoring are low relative to other techniques for improving judgmental forecasting, the technique imposes a greater cognitive load on forecasters and increases the time that they require to make their forecasts by about fifty percent.

Chapter 4 Length Effects in Judgmental Forecasting

Overview

As discussed in Chapter 1, the length of time series graphs is another understudied area in judgmental forecasting, which might be proved extremely useful if appropriate amounts of data are found to produce more adequate judgmental forecasts. When forecasts are produced by formal statistical means, it is natural to expect those forecasts to be better with longer time series. This is because longer series enable the patterns in those series to be extracted from the noise more effectively. Of course, this expectation would not be borne out if the formal approach was merely to extract the naïve forecast (i.e. to use the last data point as the forecast for the next one). Furthermore, as Makridakis, Wheelwright and McGee (1983, p.555) point out, it is more likely that the patterns in longer series will change; when they do, any approach not taking this into account may produce worse forecasts from longer series. Generally, however, formal methods produce more accurate forecasts with more data (though the rate of improvement declines as series lengthen).

Will the same phenomenon to occur in judgmental forecasting? Andreassen and Kraus (1990) showed that the quality of forecasts implied by performance in a simulated trading task was better when trends did not change over a series of 120 data points than when they did. This finding implies that judgmental forecasts do not take sufficient account of regime change (cf. O'Connor et al.,

1997) but it does not directly address the issue of whether sample size affects the quality of judgmental forecasts when patterns in the series do not change. To the best of my knowledge, there are just three studies that do address this issue directly. In the first one, Wagenaar and Timmers (1978) required people to make forecasts from three, five or seven points of an exponential growth series presented as a sequence of numbers (i.e. in tabular form). The points in each condition were approximately equally spaced over a total time period. As a result, the interval between successive points was greater when there were fewer of them. Wagenaar and Timmers (1978) found that, while the length of the total time period had no effect on forecasting performance, accuracy of predictions was higher when there were *fewer* data points. This is just the opposite of what it is expected from formal approaches to forecasting. In the second study, Lawrence and O'Connor (1992) presented people with graphs of either 20 or 40 successive data points in Autoregressive Moving Average (ARMA) series. In both conditions, data points represented quarterly data and the last of them was one quarter before the first of the four quarterly points that had to be forecast. Lawrence and O'Connor (1992) found that absolute error in the forecasts averaged over the four horizons was approximately twice as large when series comprised 40 data points than when they comprised 20 data points. Not unreasonably, they found this finding 'both surprising and counter-intuitive'. Again, it is just the opposite of what would be expected if people were using some cognitive analogue of a formal technique to make their forecasts. These two studies produced similar findings despite differences in series type (exponential versus ARMA), range of data points examined (3, 5,

and 7 versus 20 and 40), data spacing (different inter-point intervals over the same total time period versus the same inter-point intervals over different total time periods), and data format (tabular versus graphical).

What could have produced such a generalizable finding? Lawrence and O'Connor (1992) and reviewers of these results (Goodwin and Wright, 1994; Webby and O'Connor, 1996) have suggested two possibilities. First, people may suffer from cognitive overload when they are presented with more data. For this to account for performance becoming worse (rather than merely not becoming any better), it has to be assumed that adding data causes people to become so overwhelmed by their task that they put less effort into it (Lawrence and O'Connor, 1992). A second alternative is that the longer the total time period over which the series extends, the more likely people are to think that the patterns in it will change. Hence, for series extending over a longer period of time, they are more likely to forecast away from points produced by simple extrapolation of the existing patterns in the series. Lawrence and O'Connor (1992) liken this to the 'gamblers' fallacy', where runs or trends are expected to reverse. However, without elaboration, it is not clear how this explanation accounts for Wagenaar and Timmers (1978) findings. This is because they found the effect for series with different numbers of data points that extended over the *same* total period of time and because they found that varying the total period of time had *no effect* on accuracy.

The third study was carried out by Andersson et al. (2012). They required people to make forecasts from either five, 10 or 15 daily 'share prices' in series with positive linear, negative linear, or no trend. With graphical but not tabular presentation, they found a highly significant effect of series length: mean absolute error (MAE) in forecasts from series with five points (MAE = 70.5) was much higher than it was from series with 10 points (MAE = 55.5) or 15 points (MAE = 49.7). Clearly, results of this study contradict those of the other two. Unlike them, they are consistent with what it would be expected if people use some cognitive analogue of a formal process to make their forecasts. Why do the results of this third study differ from those of the other two? Andersson et al.'s (2012) study used series of independent data points with or without a linear trend. In Wagenaar and Timmers (1978) study, series had non-linear trends and, in Lawrence and O'Connor's (1992) study, points were not independent: in other words, series were more complex than in Andersson et al.' s (2012) study. There is also another difference that may help to explain the difference in results. The range of data points examined was low in Wagenaar and Timmers (1978) study (3, 5 and 7), high in Lawrence and O'Connor's (1992) study (20 and 40) but between these two extremes in Andersson et al.'s (2012) experiments (5, 10, and 15). These observations suggest that it would be worthwhile carrying out experiments with a variety of series types and with a much broader range of series lengths. It appears that the counter-intuitive findings occur when series contain more complex patterns and/or that there may be a non-linear relationship between series length and forecast accuracy. Hence,

I test the hypothesis (H_1) that the relation between forecast accuracy and series length is non-linear.

For series with high levels of autocorrelation, naïve forecasts produce fairly accurate predictions. For such series, suppose that people use the naïve forecast as a default when series are too short for them to perceive the autoregression in the series. Suppose also that they appropriately use a forecast close to this naïve one when series are long enough for them to perceive the autoregression in the series. It then follows that the distance of forecast from the last data point should vary little with the length of highly autoregressive series. For series with long-term linear or seasonal trends, naïve forecasts fail to produce accurate predictions. However, suppose that, as before, people use the naïve forecast as a default when series are too short for them to perceive the trends in the series. However, when series are long enough for them to perceive the trends in the series, they should make forecasts that are appropriately distant from the naïve forecast. Thus, distance of forecasts from the last data point should increase with the length of series that contain trends. Hence, I also test the hypothesis (H_2) that the absolute distance between forecasts and the last data point increases with the length of trended series but does not do so with series that have high levels of autoregression.

4.1 Length effects in judgmental forecasting (Experimental Study 3)

In this experiment, participants were presented with graphical representations of time series and asked to make forecasts for the next point (one-step ahead forecast). To test the above hypotheses, I manipulated the length of the time series and the complexity of the pattern in the data series.

4.1.1 Method

Participants

One hundred and fifty students (52 men, 98 women) from UCL's subject pool acted as participants. Their mean age was 26 years. They were told (truthfully) that the five participants with the lowest Mean Absolute Error scores would each be rewarded with a payment of £5.00. Although Remus, O'Connor and Griggs (1998) found no significant incentive effect on the accuracy of time series forecasting, the £5.00 award for top performance rendered the experiment popular among students and, thus, data collection was conducted at a quicker rate.

Design

Participants were divided into five groups, each one corresponding to one length condition. The experiment used a mixed design in which participants made forecasts from four time series of different types, each of which contained

40 or 20, five, two, or one data points depending on the condition to which they assigned. Thus each participant was tested in a specific length condition but experienced all four types of series. Time series were generated uniquely for each participant and the order in which the four different series occurred was randomly ordered for each of them.

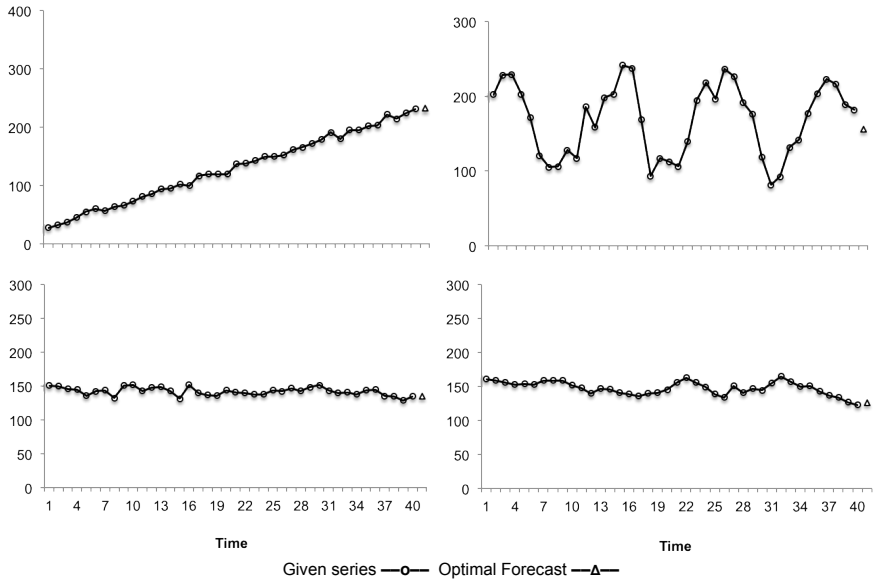

Figure 4.1 Examples of the four types of series, showing 40 data points (seen by participants) followed by the optimal forecast (not seen by participants), shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Stimulus materials

Four types of series were selected to ensure that they varied in complexity. The simplest were series of independent data points with a linear trend imposed upon them. More complex were series of independent data points with a seasonal trend imposed upon them and untrended series of highly autocorrelated

data points. More complex still were untrended non-linear series with a fractal structure. These also had high levels of autocorrelation but the autocorrelation function decayed more slowly: they showed a longer memory than the linear autoregressive series. All series were presented graphically. Examples are shown in Figure 4.1 with optimal forecasts.

Linear trended series were generated from the equation: $X_t = 5t + \varepsilon$. The noise term, ε, had a mean of zero and a variance of 19.0. The final data point of these trended series was approximately 10% of the screen height above its vertical mid-point. Thus, the trend imposed on the series was a mild one. Seasonal series were constructed by using the equation: $X_t = 70\cos(100t + 20) + 170 + \varepsilon$, where the noise term had a mean of zero and a variance of 225. The starting point of these series was chosen so that the last data point was a) close to the vertical mid-point of the screen and b) one third of the way from the mid-point of the seasonal cycle towards its peak. Each wavelength phase lasted for 12 time periods. There were 3.33 wavelengths in the screen. Each wavelength's width corresponded to a 30% of the screen width. The autocorrelated series were produced by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon$, where X_{t-1} was the previous observation, μ was the mean of the series, α was the degree of autocorrelation ($\alpha = 0.9$), and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2 ($\sigma^2 = 36.0$). The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen. To construct the untrended non-linear long

memory (fractal) series I used the multiple time-scale fluctuation approach (Koutsoyiannis, 2002). The autocorrelation and variance restrictions were calculated from the corresponding equations after the Hurst exponent value was selected to be equal to 0.9. Fractal time series with high Hurst values (H = 0.9) exhibit a long-range memory autocorrelation function: it decays as a power function rather than as an exponential function typical of non-fractal autocorrelated series (Gilden, 2009).

The task was not performed within a particular scenario, such as one associated with sales forecasting. This was to avoid introduction of frame-specific biases, such as elevation effects arising from optimism or perceived control (Eggleton, 1982; Lawrence and Makridakis, 1989). Hence, the vertical axes of the graphs used to present the series were unlabelled.

Procedure

Each participant performed the task individually on a computer. They read a short introduction to the study and then entered their demographic details (age, sex). Then the trials began. Series were presented as line graphs. After the end of each series, a vertical line was presented in the next time period to indicate where forecast had to be made. When a forecast was made, a blue dot appeared in the position of the cursor when the mouse was clicked. This dot was linked with a blue line with the last data point of the graph. Once a forecast had been made in this way, the next data series appeared. Participants were not given immediate feedback regarding the quality of their forecasts. When projected data points were fewer than 40 (i.e. L = 20, L = 5, L = 2 and L = 1), a label was

presented on the screen informing participants that earlier data were not available. An example of the task screen with a seasonal series of 20 data points is shown in Figure 4.2. In this figure, I have also depicted the vertical bar on which participants made their forecasts.

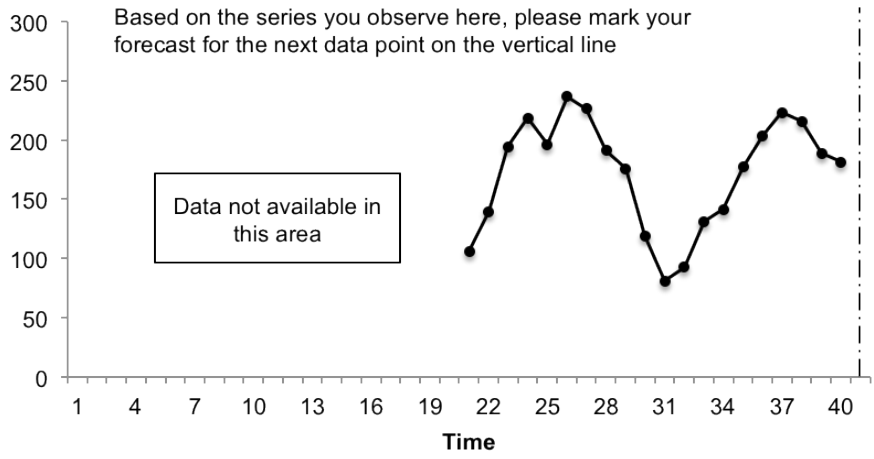

Figure 4.2 Example of the task with 20 data points of a seasonally trended series and showing the vertical bar on which participants made their forecast for the immediate (one step ahead) forecast horizon

4.1.2 Results

Six participants whose forecasts were at least 3 inter-quartile ranges from the median of each group were removed and replaced. This resulted in a total of 150 participants, thirty in each length condition. To assess H_1, absolute errors were calculated and compared across the five length conditions. Then, to test H_2, I use independent t-tests.

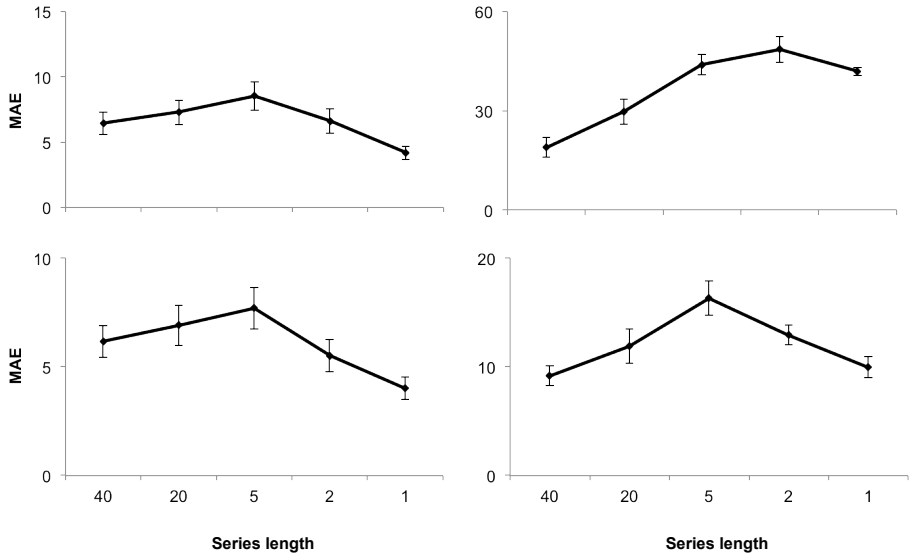

Figure 4.3 Graphs of mean values of absolute error (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Effects of series length on accuracy Graphs of MAE against series length (Figure 4.3) show an inverse U-shape function for all series' types. To examine the significance of these effects, I carried out separate one-way analyses of variance (ANOVA) with polynomial contrasts on the MAE data for each series type. Here and later, Welch tests were performed to examine whether the homogeneity of variance assumption had been violated: if it had been, the F-test was adjusted accordingly. Independent t-tests were used to follow up results of these analyses of variance. When variance across groups in these tests was heterogeneous, Games–Howell post hoc tests were used. For the rest of the cases, Bonferroni corrections were applied.

For the linearly trended series, there was a main effect of length across groups (F (4, 70.75) = 4.78; $p < 0.05$). Absolute error described an inverted U-shape function. Polynomial contrasts showed the quadratic component to be significant ($p < 0.05$). The error was lower for long lengths (L = 40) and increased as length decreased (L = 20) until it reached its maximum value for L = 5. Then, it decreased again for shorter lengths (L = 2 and L = 1). Independent two-sample t-tests, with Games-Howell corrections, were again used to compare participants' predictions among the ten different pairs of lengths. Two-tailed tests ($p < .05$) showed that a very short length (L = 1) produced higher accuracy than the medium length (L = 5) but no other differences between specific length conditions were significant.

For the seasonal series, there was a main effect of length across groups (F (4, 66.57) = 15.88; $p < 0.001$). Absolute error described an inverted U-shape function. Polynomial contrasts analysis showed the linear and quadratic component to be significant ($p < 0.001$). The error was lower for long lengths (L = 40) and increased as length decreased (L = 20) until it reached its maximum value for L = 2. Then, it decreased again for length L = 1. Independent two-sample t-tests, with Games-Howell corrections, were used to compare participants' predictions among the ten different pairs of lengths. Two-tailed tests showed significant differences in MAE between the predictions for 40-5, 40-2, 40-1, 20-5, 20-2, 20-1 ($p < 0.05$); in all other cases, differences were not significant.

For the autoregressive series, there was a main effect of length across groups (F (4, 71.67) = 5.05; $p < 0.001$). Absolute error again described an inverted U-shape function. Polynomial contrasts showed the quadratic component to be significant ($p < 0.001$). The error was lower for long lengths (L = 40) and increased as length decreased (L = 20) until it reached its maximum value for L = 5. Then, it decreased again for shorter lengths (L = 2 and L = 1). Independent two-sample t-tests, with Games-Howell corrections, were used to compare participants' predictions among the ten different pairs of lengths. Two-tailed t-tests ($p < .05$) showed significant differences in MAE between the predictions for 40-5 and 5-1 but no other differences between specific length conditions attained significance.

For the fractal series, the ANOVA revealed a main effect of length across groups (F (4, 71.39) = 4.14; $p = 0.015$). Absolute error described an inverted U-shape function. Polynomial contrasts analysis showed the linear and quadratic components to be significant ($p < 0.05$). The error was lower for long lengths (L = 40) and increased as length decreased (L = 20) until it reached its maximum value for L = 5. Then, it decreased again for shorter lengths (L = 2 and L = 1). Independent two-sample t-tests, showed significant two-tailed differences for errors between the predictions for L = 5 and L = 1 ($p = 0.011$); in all other cases, no significant differences occurred. For all series that contain a pattern as well as noise, these analyses are consistent with the first hypothesis (H_1) that length increase does not impair accuracy: in each time series type, the contrasts analysis showed the quadratic component to be significant. The analyses also

show that the very short length (L = 1) produced higher forecast accuracy than the medium length (L = 5).

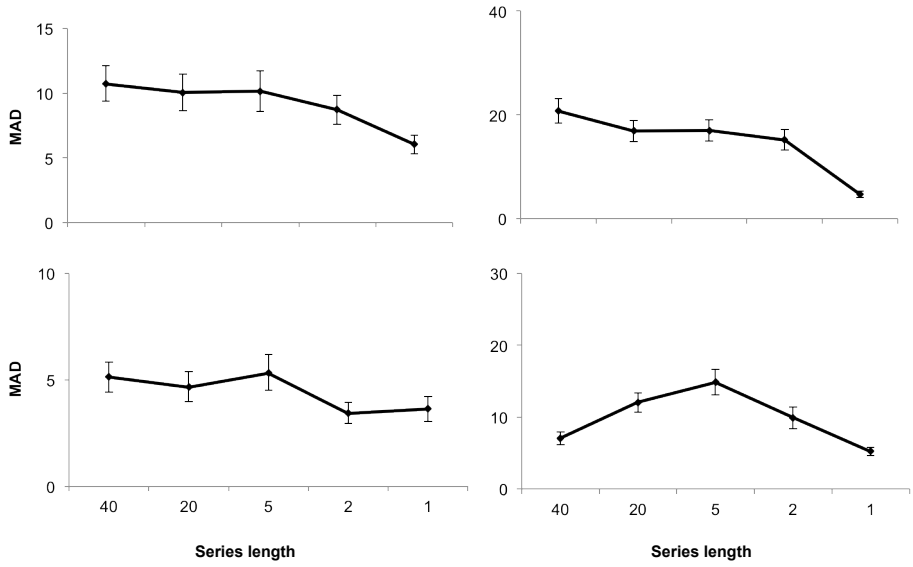

Figure 4.4 Graphs of mean values of absolute differences between forecasts and the last data point (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Distance of forecasts from the last data point. Figure 4.4 shows, for each series type, the Mean Absolute Differences (MAD) between participants' forecasts and the last data point. These differences are calculated as absolute value of the forecast minus last data point of the series. These analyses should demonstrate whether the amount that forecasters adjust away from the naïve forecast is appropriate for the data series. For the linearly trended series, there was a main effect of length across groups (F (4, 70.39) = 3.83; p = 0.05). Polynomial

contrasts showed the linear component to be significant ($p < 0.05$). The absolute distance of forecasts from the last data point increased with series' length. As series length increased, participants produced forecasts that were further away from the last data point. This is what I would expect if they were increasingly able to identify the trend signal as series length increased. Differences between MAD values for longer series (L = 40, L = 20, L = 5) and those for shorter ones (L = 2, L = 1) were significant ($p < .05$).

For the seasonally trended series, there was a main effect of series length across groups (F (4, 65.35) = 25.30; $p < 0.001$). Polynomial contrasts again showed the linear component to be significant ($p < 0.001$). Absolute distance of the forecasts from the last data point increased with an increase in the series' length. Again, this is what I would expect if participants were increasingly able to perceive the trend in the series as they increased in length. Differences between MAD values for longer series (L = 40, L = 20, L = 5, L=2) and those for the shortest one (L = 1) were significant.

For the linear autoregressive series, there was a main effect of series length across groups (F (4, 71.67) = 5.05; $p < 0.05$). Polynomial contrasts showed the quadratic component to be significant ($p < 0.001$). The distance of forecasts from the last data point was higher for medium lengths (L = 5) than for short or long lengths and significant pairwise differences are only found between pairs which contained the L = 5 condition. In the next section, I discuss possible reasons for this unexpected pattern in the data.

For the fractal series, there was no main effect of series length across groups. Values of the MAD scores were very close to zero for all length conditions. With short series, participants anchored their judgments strongly on the last data point, thereby producing predictions very close to the naïve forecast. With longer series, they continued to do so.

Analyses of signed errors The effect of series' length on mean signed error (MSE) was also examined. I calculated as the value of the forecast minus the value of the noise-free signal for the point at which the forecast was made. For the fractal and linear autoregressive series, no differences were found. For the linearly trended series, there was a main effect of length across groups (F (4, 69.76) = 3.71; $p < 0.05$). Polynomial contrasts analysis showed the linear component to be significant ($p < 0.001$). Signed error described a linear function, signifying greater trend damping for shorter series. For the seasonally trended series, there was a main effect of length across groups (F (4, 66.57) = 14.87; $p < 0.001$). Polynomial contrasts analysis showed the linear and quadratic component to be significant ($p < 0.001$), which again signifies greater trend damping for shorter series. The results are shown in Figure 4.5.

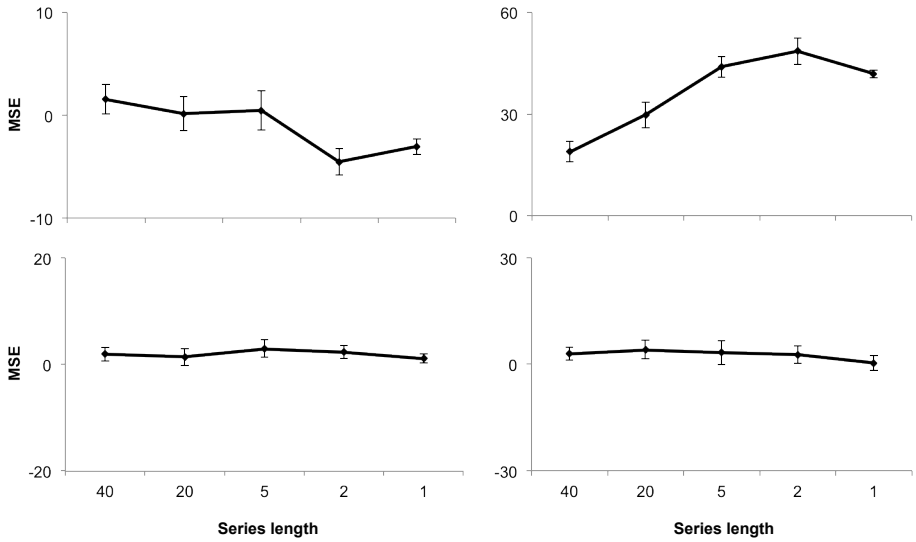

Figure 4.5 Graphs of mean values of signed error (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Discussion

For all series types, forecast error described an inverted U-shaped function: MAE was low for long series (L = 40), increased as series length decreased (L = 20), took a maximum value for L = 5 (L = 2 for seasonally trended series), and then decreased again for L = 1 and L = 2 (L = 1 for seasonally trended series). These results are consistent with those of Andersson et al. (2012). They found that MAE was higher when series had five points than when they had 10 or 15 points. They are also consistent with results reported by Wagenaar and Timmers (1978): they found that, with very short series (three, five, or seven points), forecasts were more accurate with shorter series. Thus, apparently conflicting

findings showing that accuracy decreases with longer series (Wagenaar and Timmers, 1978) and that it increases with longer series (Andersson et al., 2012) can be reconciled taking the values over which series length was varied into account and recognizing that there is an inverted U-shaped function relating forecast accuracy to series length. Results are not consistent only with those of Lawrence and O'Connor (1992). However, their experiment differed from that of Wagenaar and Timmers (1978) and from the present findings in a number of ways. For example, they calculated the accuracy of forecasts by averaging over four horizons whereas I examined MAE only for the forecast for the most immediate horizon. It is possible that MAE of the forecast for the immediate horizon and MAE of forecasts for more distant horizons are differentially affected by the length of the data series.

The data shown in Figures 4.4 and 4.5 permit some tentative inferences about the cognitive processes underlying forecasting performance. Judgmental forecasts may be produced by heuristics that are independent of the long-term pattern in the data series. The naïve forecast is one such a heuristic: it can be used when data series comprise a single data point or when they contain many data points. Alternatively, forecasts may be produced by heuristics that are dependent on the forecaster's ability to extract the long-term pattern from the series. Thus, for example, forecasters may produce damped extrapolations of the long-term trend in the series (Harvey and Reimers, 2012) or may use their assessment of the level of autocorrelation in the series to decide how much to regress from the last data point towards the series mean (Reimers and Harvey,

2011). The analysis of the effect of series' length on the absolute distance between the last data point and the forecast (Figure 4.4) indicates that forecasts tend to be close to the naïve forecast when series are short. Because the last data point is the only or is the most salient piece of information available to forecasters, they have to rely on a heuristic that does not require extraction of a pattern from the series.

With linearly or seasonally trended series, use of the naïve forecast for shorter series lengths produced high levels of trend damping (Figure 4.5). However, the distance between forecasts and the last data point increased as the data series lengthened and, as a result, the degree of trend damping declined. I assume that this occurred because forecasts for longer series relied on a heuristic that required extraction of the pattern in the series. When series were longer, forecasters were better able to extract this pattern and were more confident on relying on it: as a result, the mean forecast moved further from the last data point.

With linear series containing high levels of autocorrelation, forecasters are likely to have used the naïve forecast for short data series (L = 1, L = 2) and to have switched to using a heuristic based on pattern extraction for longer ones. People are sensitive to levels of autocorrelation in linear series having many data points (Reimers and Harvey, 2011). Thus, for the longest series (L = 20, L = 40), I can assume that they were able to determine that autocorrelation was high and, therefore, they produced forecasts appropriately close to the last data point (i.e. similar to the naïve forecast). But why were MAD scores higher for

series with five data points than they were for longer or shorter series? It is known that simple statistical estimators radically underestimate high levels of autocorrelation when series are short (e.g., Huitema and McKean, 1991). If people used some approach to assessing autocorrelation that approximated to these formal methods, they would also have underestimated the autocorrelation in the series. As a result, they would have regressed away from the last data point towards the mean too much when making their forecasts.

With fractal series, MAD scores when L = 5 appear somewhat elevated but this effect was not significant. I assume that autocorrelation was not extracted from fractal series in the way that it was from linear ones. In fact, there is currently no evidence that people are sensitive to levels of autocorrelation in fractal series. Most forecasters may simply have treated the fractal series as if they comprised pure noise around a mean. As a result, they would have maintained their default strategy of using the naïve forecast for all series lengths.

4.2 Length and horizon effects in judgmental forecasting (Experimental Study 4)

Experiment 3 was able to reconcile the apparently conflicting results of Andersson et al. (2012) and Wagenaar and Timmers (1978): the former compared longer series drawn from that part of the inverted U-shaped curve where error increased with decreasing length whereas the latter compared shorter series drawn from that part of the curve where error decreased with

decreasing length. However, Lawrence and O'Connor's (1992) results remain anomalous: they used longer series but found that error decreased with decreasing length. There are a number of underlying factors in the characteristics of their study, which might be able to provide explanations for this discrepancy; for example, if one investigates more carefully the stimuli used in this study, it will be immediately obvious that Lawrence and O'Connor did not use conventional ARMA series. Stimuli generated according to Model 1 had a parameter outside the bounds of invertibility, rendering it not directly equivalent with traditional AR models, such as the one used in the current thesis; this model produced declining weights on the observations with time, implying that some older observations may have been more correlated with the current observation than more recent ones. Moreover, the equations used to generated stimuli for Model 2 were actually equivalent to white noise. In the current thesis, such types of series were not investigated at all so it is difficult to speculate what would have happened under the current circumstances if these series were to be used. Here, I used two types of highly correlated series: AR and High Hurst long memory series. Results seem to have coincided for those two types of series with high degrees of autocorrelation. If random or anti-persistent (low or negative autocorrelation (see Koutsoyiannis, 2000)) series were to be used, for example, the optimal strategy to achieve accuracy would no longer be achieved by taking into account the patterns in the series but rather by forecasting the average of the series; optimal forecasts would have derived from an averaging heuristic strategy in this case because it is impossible to predict randomness in these types of series. Therefore, it might have been more

beneficial to provide subjects with less data points to avoid the use of heuristics, which are closely tied with the use of patterns. Another factor that might have rendered Lawrence and O'Connor's (1992) study not directly in line with the current one, is that in their accuracy assessments, they averaged error scores across four horizons. It is possible that, had they reported data only for the most immediate (first) horizon, and although their stimuli were different, their results would have been similar to those of Andersson et al. (2012). However, for this to happen, results from later horizons would have had to have shown the reverse pattern in order to produce the reported findings for error scores integrated across all four horizons. This leads to the question of whether the inverted U-shaped curve relating MAE to series length that I found for the immediate forecast horizon is maintained or changed (e.g., reversed) for later forecast horizons. For example, one possibility is that the peak error in the U-shaped curve is shifted to the left for more distant horizons: a peak error at series lengths of 30-40 rather than 5-10 would allow interpretation of Lawrence and O'Connor (1992) results in conjunction with all the other findings. Thus, the second experiment in this Chapter is similar to the first one, except that participants made forecasts for the third rather than for the first forecast horizon.

4.2.1 Method

Participants

One hundred and fifty participants (81 men, 69 women) were recruited from Amazon's Mechanical Turk online pool, a crowdsourcing web service commonly used for data collection by psychologists (Paolacci, Chandler and Ipeirotis, 2010). Their mean age was 33 years. They were paid 0.5 $ for their participation.

Design and Stimulus materials

Design and stimulus materials were the same as before. However, in this experiment, the vertical line indicating where the forecast had to be made was placed in the third time period after the last data point. As before, a blue dot appeared in the position of the cursor when the mouse was clicked to indicate the position of the chosen forecast.

Procedure

This experiment was web-based. The only procedural difference from the previous one was that participants were asked to provide a forecast for a more distant horizon (three steps-ahead rather than one step-ahead). Figure 4.6 shows an example of the task screen in this experiment.

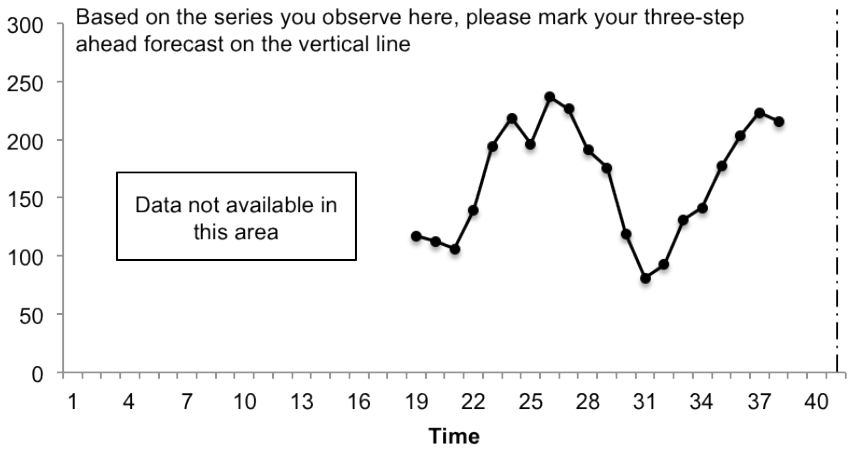

Figure 4.6 Example of the task with 20 data points of a seasonally trended series and showing the vertical bar on which participants made their forecast for the more distant (three steps ahead) forecast horizon

4.2.2 Results

Participants whose forecasts were at least 3 inter-quartile ranges from the median of each group were removed and replaced. This resulted in a total of 150 participants, thirty in each length condition.

Effects of series length on accuracy Graphs of MAE against series length are shown in Figure 4.7 for each of the four series types. An inverse U-shape function was found for all series, except for the seasonal one.

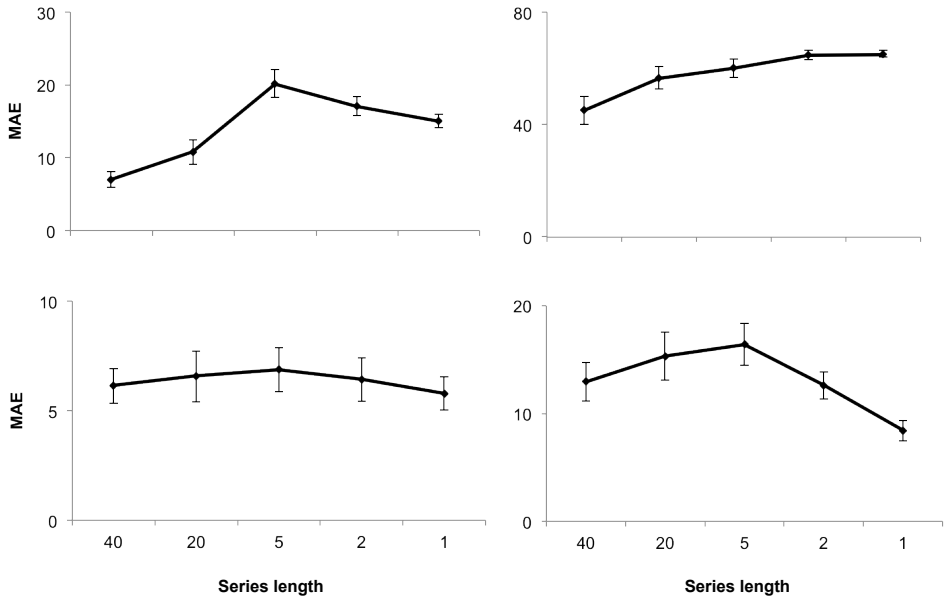

Figure 4.7 Graphs of mean values of absolute error (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

For the linearly trended series, there was a main effect of length across groups (F (4, 71.46) = 14.55; $p < 0.001$). Polynomial contrasts showed that both the linear and quadratic components were significant ($p < 0.001$). MAE described an inverted U-shaped curve with a peak value at L = 5. Independent two-sample t-tests (two-tailed), with Games-Howell corrections, showed significant differences in MAE between the pairs of lengths 40-5, 40-2, 40-1, 20-5, and 20-2. For the seasonal series, there was a main effect of length across groups (F (4, 68.48) = 4.80; p = .002). Polynomial contrasts showed the linear component to be significant ($p < 0.001$). Shorter series led to worse forecasts. Independent two-sample t-tests (two-tailed), with Games-Howell corrections, showed

significant differences in MAE only between the pairs of lengths 1-4 and 1-5 ($p < 0.05$). For the autoregressive series, there was a main effect of length across groups (F (4, 70.44) = 5.21; p = .001). Absolute error described an inverted U-shape function with polynomial contrasts showing both linear and quadratic components to be significant ($p < 0.05$). As before, peak MAE was obtained when L = 5. Independent two-sample t-tests (two-tailed), with Games-Howell corrections, showed significant differences in MAE between for 20-1 and 5-1 ($p < 0.05$). For the fractal series, the ANOVA revealed no main effects of length across groups. Polynomial contrasts analysis showed none of the components to be significant.

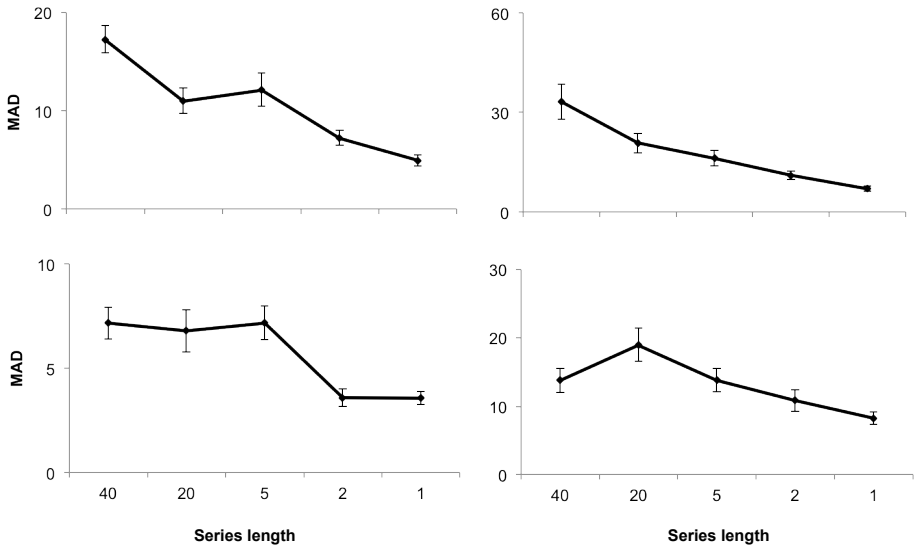

Figure 4.8 Graphs of mean values of absolute differences between forecasts and the last data point (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Distance of forecasts from the last data point. Graphs of MAD against series length are shown in Figure 4.8 for each of the four series types. For the linearly trended series, there was a main effect of length across groups (F (4, 69.48) = 20.29; $p < 0.001$). Polynomial contrasts showed the linear component to be significant ($p < 0.001$). The absolute distance of forecasts from the last data point increased with series' length. For the seasonally trended series, there was a main effect of length across groups (F (4, 67.39) = 12.93; $p < 0.001$). Polynomial contrasts analysis showed the linear component to be significant ($p < 0.001$). Absolute distance of the forecasts from the last data point increased with an increase in the series' length. For the autoregressive series, there was a

main effect of series length across groups (F (4, 69.87) = 5.90; $p < 0.001$). Polynomial contrasts showed that both linear and quadratic components were significant ($p < 0.05$). The distance of forecasts from the last data point was higher for medium lengths (L = 20) and significant pairwise differences were found between pairs 20-2, 20-1, 5-1. For the fractal series, there was a main effect of length across groups (F (4, 69.19) = 9.96; $p < 0.001$). Polynomial contrasts analysis showed that the linear component was significant ($p < 0.001$). The distance of forecasts from the last data point was higher for long and medium lengths and significant pairwise differences were found between pairs 40-2, 40-1, 20-2, 20-1, 5-1.

Analyses of signed error For the fractal and linear autoregressive series, no differences in MSE were found. For the linearly trended series, there was a main effect of length across groups (F (4, 70.44) = 17.27; $p < 0.001$). Polynomial contrasts analysis showed the linear component to be significant ($p < 0.001$). The negative sign of the MSE scores shows that forecasts were too low with this upwardly trended series. Thus trend damping occurred. However, decreasing negativity of MSE as series length increased shows that trend damping decreased as series became longer. For the seasonally trended series, there was a main effect of length across groups (F (4, 68.26) = 4.82; $p = 0.002$). Polynomial contrasts analysis showed the linear component to be significant ($p < 0.001$). The positive sign of the MSE scores show that forecasts were too high for this downward segment of the seasonal series: trend-damping occurred. However, MSE scores dropped closer to zero as series length increased, an

effect again showing that trend-damping decreased (but was not eliminated) as series became longer. The results are shown in Figure 4.9.

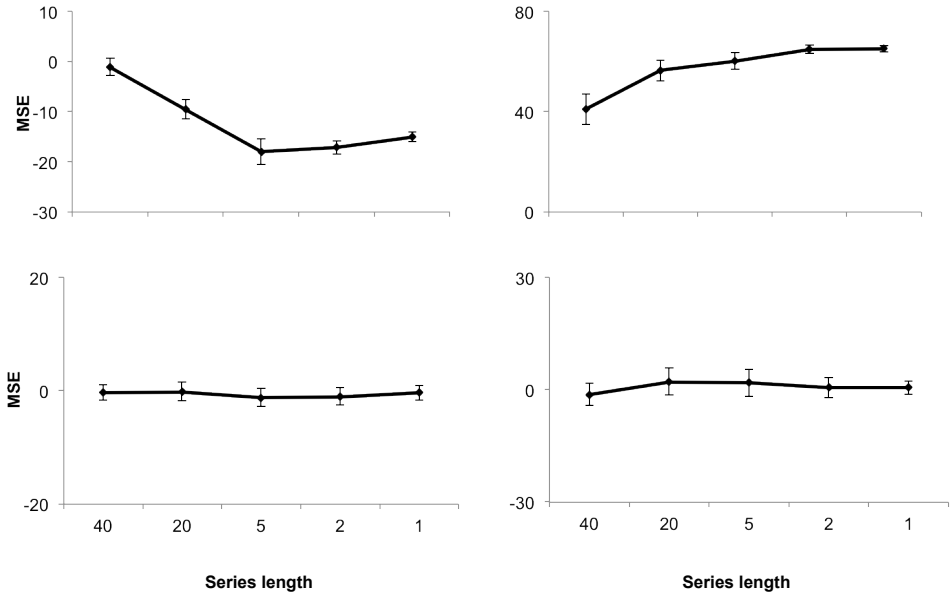

Figure 4.9 Graphs of mean values of signed error (together with standard error bars) against series length for the four different types of series, shown clockwise from the top left in the order a) linearly trended, b) seasonally trended, c) linear autoregressive, and d) fractal.

Discussion

To partly reconcile Lawrence and O'Connor's (1992) results with Experiment 3 findings and with those reported by Andersson et al. (2012) and Wagenaar and Timmers (1978), the relation between forecast accuracy and series length would have had to have been radically different from how it appeared in Experiment 3. Accuracy would have had to have been higher for L = 20 than for L = 40. This is not what the present results suggest. As Figure 4.7 shows, results were very

similar to those in Experiment 3 (Figure 4.3). MAE scores were numerically highest for L = 5 for the same three series types as before (linearly trended, autocorrelated, fractal) but, in this experiment, the quadratic component was significant for only the linearly trended and autocorrelated series. For the seasonally trended series, MAE scores failed to drop as series length was reduced from L = 2 to L = 1 in the way that they did in Experiment 3: instead they maintained the same high value. Otherwise, results were as before. Turning to the distance between forecasts and the last data point (Figure 4.8), it is again evident that results are very similar to those reported for Experiment 3 (Figure 4.4).

As before, MAD scores decreased linearly as lengths of linearly and seasonally trended series increased, indicating reduced reliance on the naïve forecast as data series became longer. With the autocorrelated series, there was again a significant quadratic component; distance of forecasts from the last data point showed a peak at L = 20. Though this peak value differed from that found in Experiment 3 (where it occurred at L = 5), it can be attributed to a similar underlying mechanisms: for short series lengths, participants tended to use the naïve forecast; for long ones, they were sensitive to the high levels of autocorrelation in the series that indicated forecasts close to the last data point were appropriate; for series of medium length, they extracted information about autocorrelation from the series but the processes that they used, in common with formal ones (Huitema and McKean, 1991), produced underestimates that resulted in forecasts being too far from the last data point. With fractal series,

forecasts were very close to the last data point when L = 1 and L = 2, indicating a strong tendency to use the naïve forecast. For longer series, forecasts were further from the last data point (Figure 4.8) but MSE remained very close to zero (Figure 4.9). This is the pattern that would be expected if participants tended to use the mean of the data as the basis for their forecasts from longer fractal series. Again, this is what one would expect if participants were insensitive to the autocorrelative structure of fractal series.

In summary, though error levels tended to be considerably higher here than they were in Experiment 3 (particularly for linearly and seasonally trended series), the way that all three dependent variables depended on series length was very similar in the two experiments: this can be seen if one compares Figures 4.3 and 4.7, Figures 4.4 and 4.8, and Figures 4.5 and 4.9. There are some minor variations but it is clear that the peak MAE did not shift to the left with the longer forecast horizon examined here. Such a shift would have allowed reconciliation of results from Experiment 3, Andersson et al. (2012), and Wagenaar and Timmers (1978) with the findings reported by Lawrence and O'Connor (1992).

4.3 Summary and General Discussion

Prior to the publication of Andersson et al.'s (2012) paper, it appeared from the work of Wagenaar and Timmers (1978) and Lawrence and O'Connor (1992) that, in contrast with forecasts produced by formal methods, judgmental

forecasts were more accurate when made from shorter series. Two accounts were proposed to account for these findings (Goodwin and Wright, 1994; Lawrence and O'Connor, 1992; Webby and O'Connor, 1996). First, human judgment may become overloaded when presented with too much data. However, for forecasting to deteriorate with longer series rather than merely fail to improve, one must assume either a) that people are unable to ignore additional data and processing more data impairs performance or b) that they are able to ignore additional data but supressing its processing incurs some cognitive penalty. Thus, people may be unable to inhibit the automatic input of older items in presented series and later controlled processing underlying forecasting may be less effective when there are more items to deal with. Alternatively, people may be able to use controlled processes to restrict input to more recent items but this may reduce the cognitive resources available to make forecasts from those items. The second proposal was that the longer a series has continued without a change in the way it has been produced, the more likely people think that such a change will occur. As a result, they would be more likely to produce forecasts that deviate from the one that they would produce on the basis of the pattern in the series.

Neither of these proposals explains Andersson et al.'s (2012) finding that judgmental forecasts improved as length of data series increased from five to 10 or 15 items. They are also inconsistent with findings from the present Experiment suggesting that forecast accuracy is related to series length via an inverted U-shaped function peaking close to five items for most series types.

The explanation that I propose for the present findings, and those of Andersson et al. (2012) and Wagenaar and Timmers (1978), is that forecasters use fairly effective pattern-independent heuristics when series are short and fairly effective pattern-based heuristics when series are long. When series are of intermediate length, they use same pattern-based heuristics that they use when series are long but do so without being aware that the effectiveness of these heuristics is increasingly compromised as series become shorter. Had they been aware of this problem, they would have used the pattern-independent heuristics that produce better performance with even shorter series.

Pattern-based heuristics can be compromised when series are short for a variety of reasons. Methods used to extract information about levels of autocorrelation may be biased. It is known that corrections need to be applied to formal methods to avoid underestimation (Huitema and McKean, 1991). Also Reimers and Harvey (2011) showed that, while judgmental forecasts indicate that people are sensitive to autocorrelation, they are insufficiently sensitive to it. Thus high levels of autocorrelation are underestimated whereas low levels are over-estimated. This is consistent with the MAD findings from the highly autocorrelated series used here (Figures 4.4 and 4.8). Why would such 'biases' increase as series become shorter? Reimers and Harvey (2011) argued that they may have a rational 'Bayesian' underpinning. Real-world series tend to show moderate levels of autocorrelation. Hence, as an a priori hypothesis, a forecaster should assume that there is a moderate level of autocorrelation in a series (say, 0.4). When they examine a series algorithmically generated to contain a high

level of autocorrelation (say, 0.8), they receive evidence that allows them to make an adjustment away from this a priori hypothesis. However, this evidence must be treated with caution because the series is noisy and not infinitely long. Hence, in producing their a posteriori hypothesis about the level of autocorrelation in the series, they move only some of the way from their a priori hypothesis (0.4) towards the level of autocorrelation indicated (with considerable uncertainty) by the series (0.8). The more certain they are of the evidence provided by the series, the more they move away from their a priori hypothesis. Thus, underestimation of high levels of autocorrelation (and over-estimation of low levels) should be greater when series are noisier. Reimers and Harvey (2011) confirmed that this was so. However, this underestimation should also be greater when series are shorter (because of such series provide lower quality evidence of levels of autocorrelation that they contain). The results that I have reported here for autocorrelated series are consistent with this (Figures 4.4 and 4.8).

Similar arguments can be made for series containing other types of patterns. For example, linear trends do not continue indefinitely. They are just parts of long-term cycles. Hence, as an a priori hypothesis, people should assume that the steepness of trends will decrease. Data from a presented series allows this a priori hypothesis to be modified. However, as presented series are noisy and not infinitely long, forecasting from steep trends still shows trend damping and this damping is greater when series are noisier (Harvey and Reimers, 2012). I should, for the same reasons, also expect trend damping also to be greater when

series are shorter; this is indeed what was found for the linearly and seasonally trended in experiments 3 and 4 (Figures 4.5 and 4.9).

Although the naïve forecast produces good estimates for most types of series, there are exceptions. Within the types of series that I examined, it produced relatively poor forecasts for seasonal series. Figures 4.4 and 4.8 show that participants made forecasts very close to the naïve forecast for seasonal series with one or two items and yet MAE scores (Figures 4.3 and 4.7) and MSE scores were very high (Figures 4.5 and 4.9). As a result, the peak of the inverted U-shaped relation between forecast error and series length shifted to the right: it was at L = 2 in Experiment 3 and at L = 1 in Experiment 4. I conclude that the peak of the inverted U-shaped relation between forecast error and series length is at five items for the other types of series because the naïve forecast is effective for those series types.

Limitations

First, Lawrence and O'Connor's (1992) findings were not reconciled with the present experiments and with those reported by Andersson et al. (2012). Lawrence and O'Connor (1992) presented their participants with graphical data and they compared performance for series with 20 and 40 points. Yet they found that the latter was worse than the former whereas I obtained the opposite result. There are some procedural differences that may help to explain these divergent findings. I used a variety of series types, including those with high levels of autocorrelation, whereas they employed series with unexpected characteristics, as discussed in the previous sections. These series might have

rendered averaging heuristics more successful. Also, I measured forecast error for a single horizon (either one step ahead or three steps ahead) whereas they integrated their error measures over four horizons.

Second, it would be useful to examine a wider range of series lengths. From the present results, it is evident that, for most series, I obtained a peak error with a series length of five because the pattern-based heuristics that people used with that series length were relatively ineffective. I would expect forecast error to be lower with series containing 10 items. If it were found to be higher, I would need to ask why forecasting is poorer from five items than from one item. One possibility would be that forecasters use less effective pattern-independent heuristics when forecasting from five items than when forecasting from a single item.

Chapter 5 Scale Effects in Judgmental Forecasting

Overview

As discussed in Chapter 1, the scale in which time series graphs are presented is another understudied and essential area in judgmental forecasting. If biases associated with the scale in which the graph is represented exist, then, the implications would be significant for all domains where judgmental forecasting is exercised in practice. For example, a variety of scale dimensions are employed to present the time series of interest in trading, managerial and other settings of the financial sector. Computer screens, monitors as well as palm-tops and mobile devices with different dimensions are used to present time series information.

Presentation scale manipulations have been investigated by Lawrence and Makridakis (1989), Lawrence and O'Connor (1992) and Lawrence and O'Connor (1993). The first study showed that the greater the space on the graph above the plot of a linearly trended time series, the higher the forecast tended to be. Contrary to expectations stemming from optimal graphical display research (Cleveland, McGill and McGill, 1988), the second study (e.g. Lawrence and O'Connor, 1992) showed that varying the scale of a graph had no effect on the accuracy of forecasts for untrended ARMA series. However, Lawrence and O'Connor (1993) showed that when the vertical scale is smaller, participants expect greater future changes in the series. Hence, they tend to forecast wider

probability distributions for smaller scales. In their paper they manipulated the vertical axis by giving their participants three different scale presentations: a small, a medium and a large one. The large scale filled three quarters of an A4 page, the medium scale halved the large presentation and the small scale halved it again.

The same scale manipulation was also employed at Lawrence and O'Connor (1992) study. They requested four horizon forecasts and calculated the average errors over these four horizons. Although larger scales exhibited smaller errors, significant differences between scales were not found. Thus, scale had no effect on accuracy of forecasts.

The significant finding on prediction intervals obtained by Lawrence and O'Connor (1993) suggests that smaller scales are more consistent with a larger range of outcomes. In fact, Lawrence and O' Connor (1993) argue that smaller scales prime forecasters to expect greater future changes. In contrast, larger scales may restrict forecasters' expectations due to boundary effects.

Increases in autocorrelation increase series variance without changing the level of the noise component in the series. Thus, on the basis of Lawrence and O'Connor's (1993) arguments, I would expect smaller scales to confer higher benefit on series with higher autocorrelation (because such scales prime forecasters to expect outcomes that are more variable). Thus, time series with widely different degrees of autocorrelation will be used here, specifically, series with zero or high autocorrelation (a = 0.9).

5.1 Scale effects in judgmental forecasting (Experimental Study 5)

In this experiment, I examine scale effects by stretching or shrinking the vertical axis. Hence, I compare two scales: a large and a smaller one. Based on Lawrence and O'Connor (1993), I tested the hypothesis that smaller scales will confer a relative advantage on forecasts made from series with higher autocorrelation.

Thus,

H_{1A}: If performance is worse with the small scale than with the large one, the advantage of the large scale over the small one will not be as great when autocorrelation is high than when it is low

Conversely,

H_{1B}: If performance is better with the small scale than with the large one, then the advantage of the small scale over the large one will be greater when autocorrelation is high than when it is low.

5.1.1 Method

In this experiment I used a within participants design in which participants made forecasts from two types of series, each of which had 30 points. They were requested to forecast the next five data points. I used two types of series

(independent and autoregressive series) and two scales (a large and a small scale). The experiment was run online.

Participants

90 participants were collected from Amazon's Mechanical Turk online pool and a total of 95 submissions were made. Participants were paid 0.2$ for their time.

Design

The experiment used a within participants design with two factors: series type (independent, autoregressive) and presentation scale (a large and a small one). There was one trial for each pair of combinations, which resulted in four trials for each participant. Series were generated uniquely for each participant and trials were randomized. Characteristics of the four types of series are described in the next section.

Stimulus materials

The two types of series were: an untrended series of independent data points and an untrended series of highly autocorrelated data points; Series were presented graphically. Untrended series were constructed by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon$, where X_{t-1} was the previous observation, μ was the mean of the series, which was set to 10 and α was the degree of autocorrelation ($\alpha = 0.9$ for autoregressive series and $\alpha = 0$ for independent series), and ε was noise produced by randomly drawing values from a uniform distribution [-3, 3] with a mean of zero and a variance of σ^2 ($\sigma^2 = 3$ for both autoregressive and

independent series). The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen. Patterned series (trended and seasonal ones) were not studied in this experiment because by manipulating the vertical axis scale would cause an alteration to the patterns in the series (i.e. a shallow trend in the small scale would become a steeper one in the large scale).

Time series were generated uniquely for each participant and the two types of series were randomly ordered separately for each of them. The task was performed within a stock price scenario, where participants were told they would observe the values of a stock price for 30 days and will be asked to forecast the next five days. Hence, the horizontal axis of the graph was labelled as days. Series were presented as line graphs. After the end of each series, five vertical lines were presented in the next five time periods to indicate where forecasts had to be made. When a forecast was made by clicking on one of the vertical lines a red dot appeared in the position of the cursor when the mouse was clicked. Two out of four trials were presented in the large scale, which was equivalent to Lawrence and O'Connor large display (three quarters of an A4 page), while in the small scale the vertical axis was halved (equivalent to the medium scale in Lawrence and O'Connor 1992, 1993 papers). The horizontal axis was kept constant. Examples of these time-series are provided in Chapters 3 and 4.

Procedure

The experiment was coded in Javascript and run online via Amazon's Mechanical Turk pool of participants. It was uploaded on to a site and subjects from the pool could participate via the web-experiment link, which was provided to them via Mechanical Turk. At the end of the task a 9 digit random number was shown to each participant and he or she had to type it back to the M-Turk site to get paid. At the start of the experiment participants read the following introductory text:

Imagine you are a trader at Wall Street premises and you are observing stock prices in this screen! Stock prices are presented in line graphs, which show the prices of the stock for 30 consecutive days! So, stock price for day 1, 2, 3,…., 30! What is the most likely stock price for the next five days, day 31, 32, 33, 34 and 35? You will mark your forecast for these days by clicking in the punctuated vertical axis! You will be presented with 4 time series in all! Instructions will be provided at the top of the screen at each stage to prompt you for any actions required. In this experiment, your time and forecasting performance is monitored. If you complete the task to quickly or produce irrelevant forecasts, your participation will be rejected automatically.

After this introductory text the trials began. To the right of the 30th observation there were five vertical lines were participants had to mark their forecasts. A label informed them about the task again. After making all five forecasts, the submit button became active for them and by clicking that they moved to the next trial. Two out of four trials were presented in the large scale, and the other

two in the small scale (these correspond to the large and medium scales in Lawrence and O'Connor's (1992, 1993) papers).

5.1.2 Results

In this section I will analyse the following variables: mean absolute error (MAE), mean absolute difference from the last data point (MAD) and the mean signed error (MSE). Mean absolute error corresponds to the difference of the forecast minus the actual value of the series and is useful to measure the forecaster accuracy. Mean absolute difference from the last data point is calculated by subtracting the forecast from the last data point. This variable is informative of the anchoring strategies participants used in each series' type and scale condition. Finally, signed errors are useful to spot elevation biases. Since this experiment is run in a stock market scenario, optimism biases might occur (e.g. Reimers and Harvey, 2011). Signed errors are calculated by subtracting the forecasting from the optimal series value. This optimal value is calculated based on the series' generating algorithm by dropping the noise component. Each of these three variables will be subjected to a three-way within-participants ANOVA, using series' type, scale' type and horizon as independent variables. Follow-up analyses will be used to clarify the nature of any obtained effects and interactions. Additional analyses will examine the degree of noise introduced into forecasts by fitting regression lines to the forecasts for the five horizons and analysing levels of residual error. Correlations between successive forecasts will also be examined. In this analysis, participants whose forecasts were at

least 3 inter-quartile ranges from the median of each group were excluded and replaced. This resulted in a total of 90 participants.

Effects on accuracy To test H_1, a three-way within participants ANOVA was employed on the mean absolute errors (MAE) for all five forecast horizons, with series type and scale type as independent variables. The three-way ANOVA yielded a main effect of horizon (F (3.23, 1152.44) = 55.81; $p < .001$), and analysis using polynomial contrasts showed that it contained a significant linear component (F (1, 356) = 123.66; $p < .001$), signifying that error increased with an increase in the forecast horizon. There was also a significant interaction between series' type and horizon (F (3.23, 1152.44) = 53.19; $p < .001$), showing the more rapid increase of error with horizon for the autoregressive series. Also, there was a main effect of series' type (F (1, 356) = 6.94; $p = .009$). Scale type yielded no significant effects. There were no other significant effects or interactions in this three-way analysis. To confirm that, a two-way analysis of variance was employed with scale type as independent variable; results confirm the three-way analysis outcomes. No significant effects or interactions were obtained. Tests of simple effects showed that MAE increase with time horizon for the independent series was not significant in either of the two scales. However, a significant increase was found for the AR series for both scales (Small scale: F (2.87, 255.65) = 55.02, $p < .001$; Large scale: F (2.48, 221.01) = 46.04, $p < .001$) and significant linear contrasts for both cases (Small scale: F (1, 89) = 114.45, $p < .001$, Large scale: F (1, 89) =

78.78, $p < .001$). Hence, error increased significantly only for the autoregressive series for both scales.

In their analysis, Lawrence and O'Connor averaged errors across the four forecasting horizons to draw their conclusions. In this section, the same will be performed for the five horizon forecasts obtained from this experiment, to produce an accuracy measure at an aggregate level. It might be the case that effects of scale operate there. A two-way ANOVA was run with average absolute error as dependent variable and scale' and series' type as independent variables. Results showed significant main effects of series' type with errors being larger for the independent series (1.58 *vs* 1.38, $F(1, 356) = 6.94$; $p < .05$). Scale type effects have not reached significance although errors were numerically larger for the small scale (1.51 *vs* 1.47). The interaction between series' and scale type have not reached significance either.

However, comparisons for each scale condition show that participants in the small-scale condition exhibit significantly larger absolute errors when forecasting for the independent series than for the autoregressive one (1.62 *vs* 1.39, $F(1, 178) = 4.04$, $p < .05$). The same is not true for the large scale (1.55 *vs* 1.38, $F(1, 178) = 2.90$, $p = .09$); in the large-scale condition, participants' performance for the two types of series is not any more distinguishable. This finding seems to occur here because in the large scale, participants' errors for the independent series are decreased. This analysis provides some (weak) evidence to support H_1.

Signed errors analysis A three-way ANOVA yielded a main effect of horizon (F (3.13, 1115.22) = 9.18; $p < .001$), and analysis using polynomial contrasts showed that it contained a linear component (F (1, 356) = 18.42; $p < .001$), signifying that error increased with an increase in the forecast horizon. Errors were always positive, with an increase with time horizon, suggesting a positive elevation bias. There were no other significant effects or interactions in this three-way analysis. Tests of simple effects in the signed error showed significant increase of MSE with time horizon for the independent series for the small scale (F (3.78, 336.6) = 3.78, $p < .05$). Errors were always positive. The same manipulation in the signed error showed significant increase of MSE with time horizon for the independent series for the large scale (F (3.91, 348.55) = 4.32; $p < .05$). Errors had a positive value for horizons 2, 3, 4 and 5. There were also main effects of time horizon to the MSE for both scales (Small Scale: F (2.30, 204.94) = 2.88; $p = .05$, Large Scale: F (2.20, 195.79) = 2.45; $p = .083$). Linear contrasts analysis showed again significant linear components for the small scale (F (1, 89) = 4.79; $p < .05$) but not for the large scale. This positive elevation bias may be associated with the fact that in this particular task subjects were forecasting stock prices where higher values are better (see also Figure 5.1).

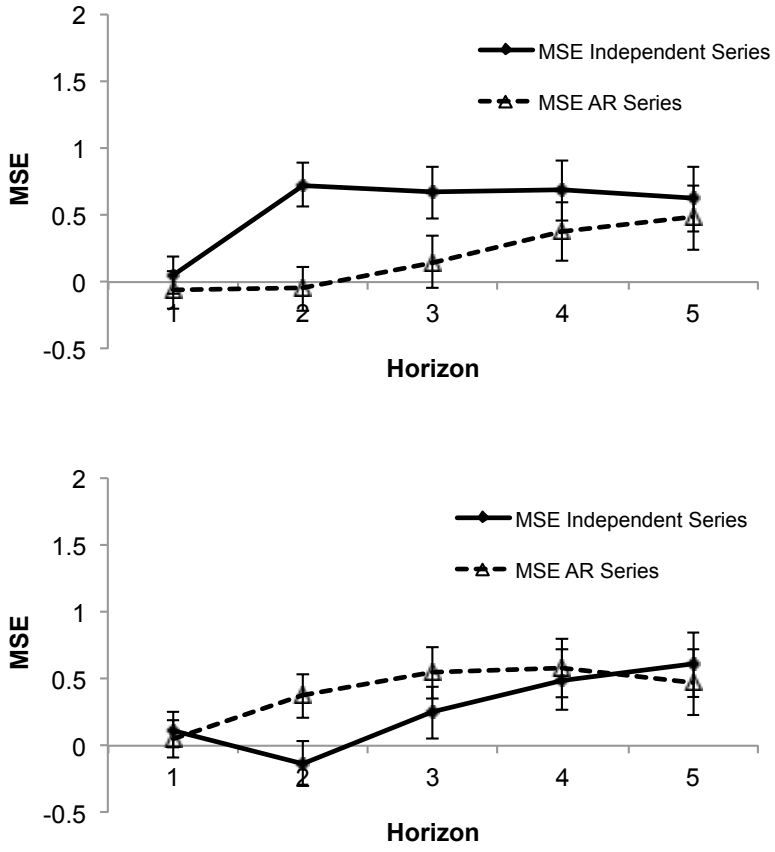

Figure 5.1 Graphs of mean values of signed error (together with standard error bars) for the independent series (continuous line) and for the autoregressive series (dashed line), for each scale condition (Small scale: upper panel, Large scale: lower panel). A positive elevation bias is present in both series' types and scales.

To complement the main analysis of the results, I fitted regression models to each one of the four sequences of five forecasts produced by each participant. For each sequence, I fitted the model: forecast = a + b (horizon) + error. Mean values of constants and trend coefficients in each condition, together with optimal values derived from the generating equations are shown in Table 5.1.

Table 5-1 Linear regressions of forecast sequences for each series type: mean values (variances in parentheses) of constants and trend coefficients. Actual values in the generating equations are shown for comparison.

		Constant (*a*)	*Trend* (*b*)
Random Series	Actual	10	0
	Small scale	10.58 (3.38)	-0.05 (0.2)
	Large scale	9.79 (3.96)	-0.19 (0.32)
Autoregressive Series	Actual	10	0
	Small scale	9.35 (17.22)	0.17 (0.50)
	Large scale	10.91 (19.18)	-0.03 (0.44)

A two-way ANOVA with dependent variable the slope of the regression equation and independent variables the scale and series type revealed a significant interaction between scale and series' type (F (1, 356) = 7.38; p = .007). For the independent series, the regression line slope coefficient was greater for the large scale, while the opposite was true for the autoregressive series. No other effects were found. The same analysis for the regression line intercept indicated that there was a significant interaction between scale and series' type (F (1, 356) = 11.39; p = .001); a positive elevation bias occurred in

the small scale for the independent series, while, for the autoregressive series, a small elevation bias was associated with the large scale display. `

Mean absolute distances I compare differences between mean absolute differences (MAD) from the last data point. A three-way ANOVA was run and yielded only an effect of series' type (F (1, 356) = 28.08; $p < .001$), signifying that anchoring mechanisms were significantly different in the two types of series. Specifically, forecasts were closer to the last data point in the autocorelated series (Figure 5.2): this is to be expected if participants are sensitive to series' autocorrelation. Hence, to follow this up, I carried out a supplementary analysis. Correlations between successive forecasts for each of the series' type and scale conditions were examined. For autoregressive series, in both scales, high correlations between successive points are observed. Their magnitude was similar to the series' autocorrelation. For the independent series, the average correlation between successive points was, on average, around 0.5 (Figure 5.4). Hence, participants were sensitive to series autocorrelation but insufficiently so. This replicates Reimers and Harvey's (2011) findings.

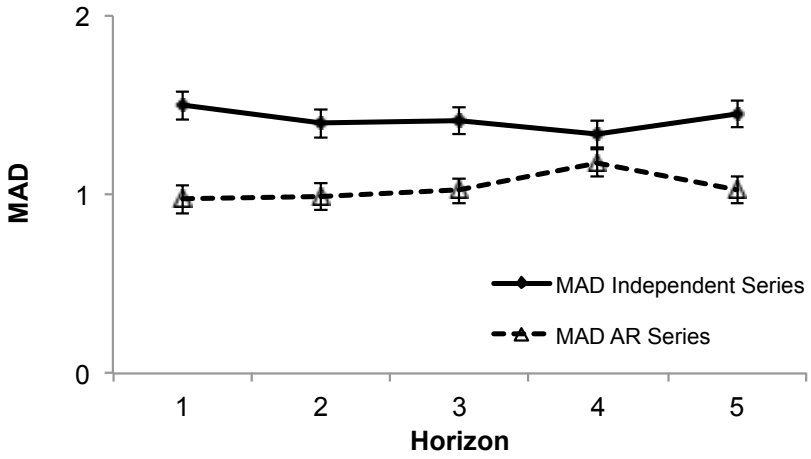

Figure 5.2 Graph of mean values of absolute differences (together with standard error bars) for the independent series (continuous line) and for the autoregressive series (dashed line), irrespectively of scale condition. Significantly different anchoring strategies are observed for the two types of series.

Tests of simple effects in the mean absolute differences confirmed these results. There were no significant scale effects or interactions between the two scale types (Figure 5.3).

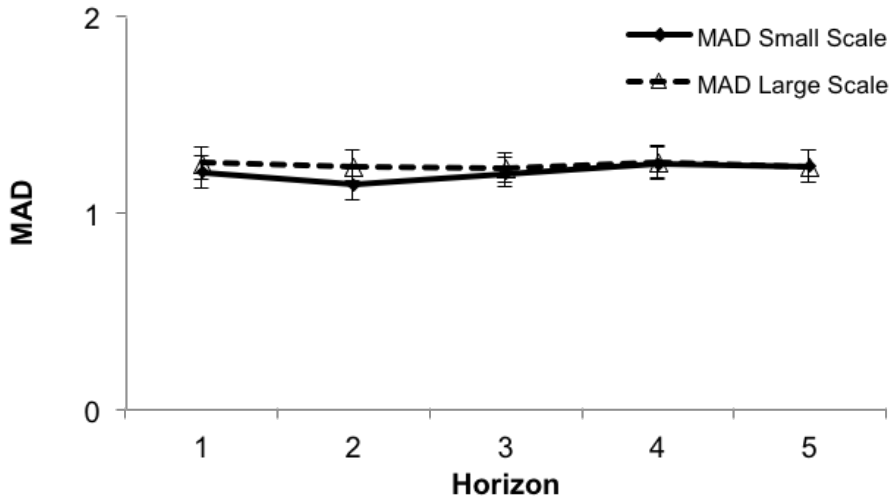

Figure 5.3 Graph of mean values of absolute distances between successive points (together with standard error bars) for the small scale (continuous line) and for the large scale (dashed line), irrespectively of series' type.

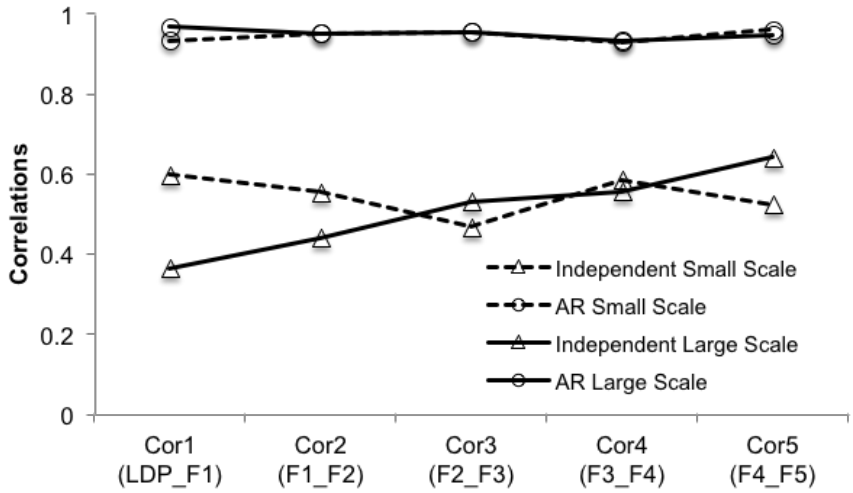

Figure 5.4 Graph of correlation values between successive points for the two series' types, for large scales (continuous line) and for small scales (dashed line).

Discussion

The experiment provided some evidence to support the hypothesis. Specifically, when the scale was small, participants produced lower errors when the series were highly autocorrelated than when they were random but, when the scale was large, this difference did not appear. Thus, the small scale selectively advantaged forecasting from the autocorrelated series. Series' variance was higher in such series: the range of values in the presented series and the range of values to be forecast were higher in the autocorrelated series. On the basis of Lawrence and O'Connor's (1993) suggestion that smaller scales lead forecasters to expect a greater range of outcomes, I argued that reducing the scale would selectively benefit forecasting from the autocorrelated series. This is what occurred.

Note that, the variance of the noise component was the same in both the independent and the autocorrelated series. Hence the scale effect cannot be attributed to this factor.

5.2 Scale effects in judgmental forecasting (Experimental Study 6)

In this experiment, scale effects were examined by manipulating again the vertical axis as before. Thus, comparisons are again performed between two scales: the large and the small one. The experiment was run as before.

The same hypothesis (H_1) as before is tested. However, in addition, a cross-experiment comparison might allow further examination of Lawrence and O'Connor's (1993) suggestion. Gaussian noise allows for greater perturbations and shifts in the series. The possibility (though small) of more extreme values would be more consistent with small scale presentation if Lawrence and O'Connor's (1993) suggestion holds (i.e. small scales lead people to expect more extreme values). Hence I also run tests to examine the following hypothesis:

H_2: Small scale displays will selectively benefit series with Gaussian noise over series that contain uniform noise of the same magnitude.

5.2.1 Method

The method was the same as in Experiment 5 except for the difference in the series noise distribution, which was Gaussian this time.

Participants

110 participants were collected from Amazon's Mechanical Turk online pool. They were paid 0.2$ for their participation.

Design

The design was exactly the same as in Experiment 5.

Stimulus materials

Same as in Experiment 5 but noise distribution was now Gaussian with a mean of zero and variance $\sigma^2 = 9$. In other words, in this experiment, I changed the noise distribution of the series. The range of perturbations produced by the Gaussian noise will be now greater than those produced with uniform noise.

To get a sense of the difference in perturbations that Gaussian noise will introduce in this experiment, I will present here simulated outcomes from the series under investigation with uniform and Gaussian noise. To measure the perturbations in the series, I calculate incrementally the absolute differences between the series' points X_n and $X_{(n + h)}$, where h is the horizon of h steps ahead. These differences will provide a measure of perturbations in the series and, hence, in the most recent segments. Here, absolute difference (AD1) corresponds to the absolute difference between successive points in time steps n and (n + 1), AD2 corresponds to the absolute difference between points in time steps n and (n + 2) and so on, until AD10, which corresponds to the absolute difference between point n of the series and point (n + 10). The absolute differences were calculated by simulating autoregressive series of 4000 points.

After simulating 4000 time steps of the series of interest with both uniform and Gaussian noise, I calculate the averages of AD1, AD2, AD3, AD4, AD5, AD6, AD7, AD8, AD9, AD10 from 100 simulated outcomes. Figure 5.5 shows the average ADs for the autoregressive series (a = 0.9) with uniform and Gaussian noise.

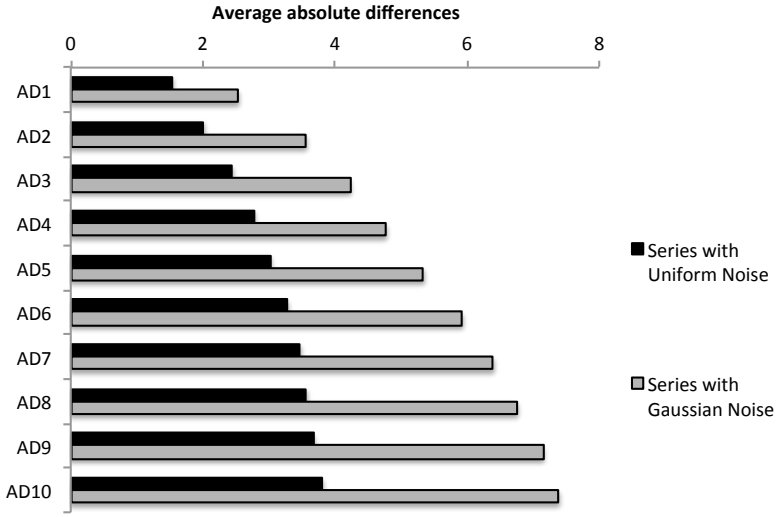

Figure 5.5 Graph of simulated mean values of absolute differences for the autoregressive series with uniform (black bars) and Gaussian (grey bars) noise.

Figure 5.5 shows differences in perturbations encountered by the subjects in autoregressive series (a = 0.9) with uniform and Gaussian noise. It is evident that the series with Gaussian noise used here produces greater perturbations for all time steps. These perturbations increase with an increase of horizon. The same applies for random series (a = 0). There, average perturbations are again greater for series with Gaussian noise (3.50 *vs* 1.90) but there is no increase

with horizon. Thus, I expect greater forecast errors in this experiment as well as less anchoring to the last data point.

Procedure

The procedure was exactly the same as before.

5.2.2 Results

I excluded participants whose forecasts were at least three inter-quartile ranges from the median of each group. This resulted in a total of 110 participants.

Effects of scale In this section, the same analysis will be performed, as before. To test H_1, a three-way within participants ANOVA was employed on the mean absolute errors (MAE) for all five forecast horizons, with independent variables those of series and scale type. Graphs of MAE are shown in Figure 5.6 for each series' type and scale. They show accuracy decreasing with increasing horizon for the autoregressive series. Scale doesn't seem to influence accuracy much for either type of series, as before. Further analysis will provide evidence about the significance of these effects.

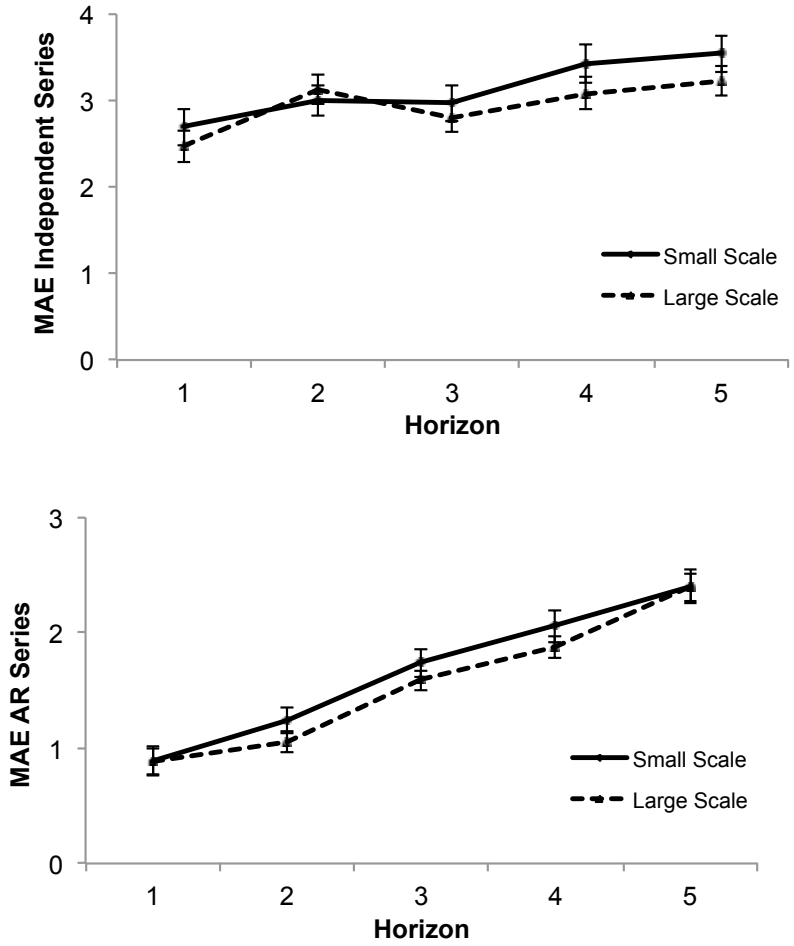

Figure 5.6 Graphs of mean values of absolute error (together with standard error bars) for the independent series (top panel) and for the autoregressive series (lower panel), for the small scale (continuous lines) and for the large scale (dashed lines).

The three-way ANOVA yielded a main effect of horizon *(F* (3.79, 1653.05) = 45.58; $p < .001$), and analysis using polynomial contrasts showed that it contained a significant linear component *(F* (1, 436) = 129.25, $p < .001$) signifying that, overall, error increased with an increase in the forecast horizon. There was also a significant interaction between series' type and horizon *(F*

(3.79, 1653.05) = 7.91; $p < .001$), showing the more rapid increase of error with horizon for the autoregressive series (Figure 5.6). Also, there was a main effect of series' type *(*F (1, 436) = 184.13; $p < .001$).

Scale type yielded no significant effects and no other significant effects or interactions occurred in this three-way analysis. To confirm that, a two-way analysis of variance was employed with scale type as independent variable; results confirm the three-way analysis outcomes. No significant effects or interactions were obtained. Tests of simple effects showed significant increase of MAE with time horizon for both series in both scales; for the independent series, both the small scale, F (4, 436) = 4.13, $p = .003$, and the large scale, F (4, 436) = 4.29, $p = .002$) exhibited significant increase with horizon. Linear contrasts were significant for both the small (F (1, 109) = 13.3, $p < .001$) and the large scale (F (1, 109) = 9.08, $p = .003$). A significant increase of MAE with horizon was found also for the AR series for both scales (Small scale: F (3.03, 330.40) = 37.18, $p < .001$, Large scale: F (2.37, 259.22) = 43.47, $p < .001$) and significant linear contrasts in both cases (Small scale: F (1, 109) = 76.38, $p < .001$, Large scale: F (1, 109) = 70.04, $p < .001$). Hence, error increased significantly in all cases (See also Figure 5.7).

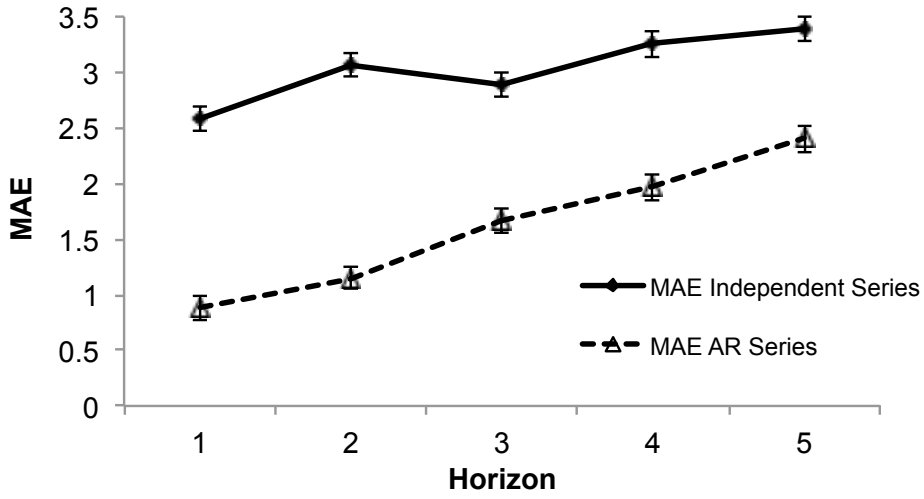

Figure 5.7 Graph of mean values of absolute error (together with standard error bars) for the independent series (continuous lines) and for the autoregressive series (dashed lines), irrespectively of scale condition.

Analysis of average errors across the five horizons was performed, as before. A two-way ANOVA was run with average absolute error as dependent variable and scale' and series' type as independent variables. Results showed significant main effects of series' type with errors being larger for the random series (3.03 *vs* 1.61, F (1, 436) = 222.63; $p < .001$). Scale type effects have not reached significance although errors were numerically larger for the small scale (2.39 *vs* 2.25). The interaction between series' and scale type have not reached significance either.

Signed errors analysis A three-way ANOVA yielded no effects but errors were again positive on average (Figure 5.8).

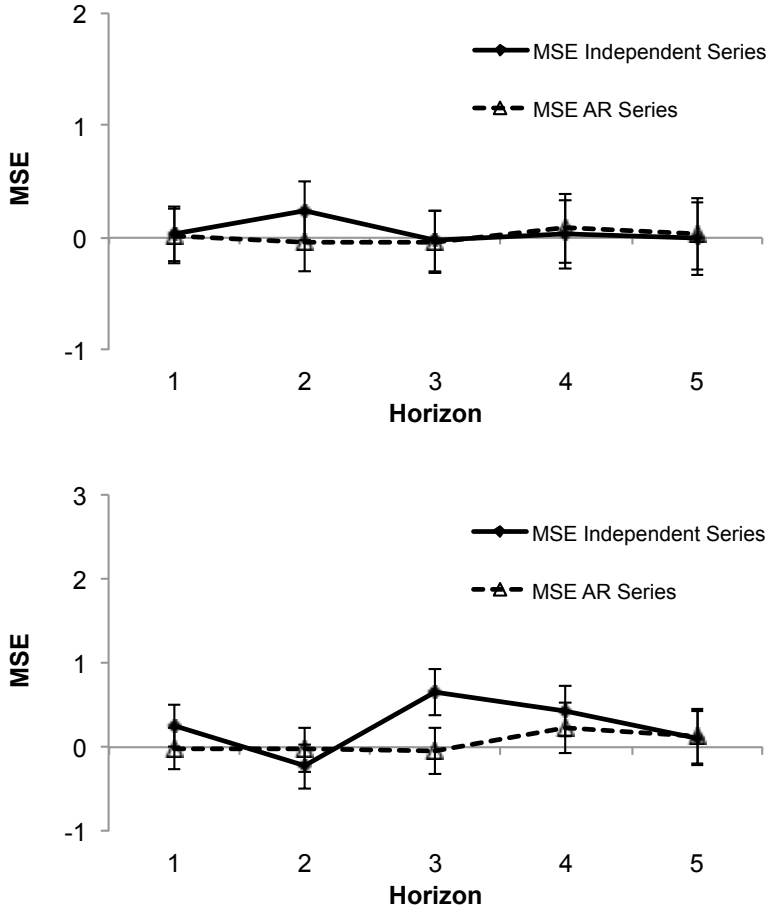

Figure 5.8 Graphs of mean values of signed error (together with standard error bars) for the random series (continuous line) and for the autoregressive series (dashed line), for each scale condition (Small scale: upper panel, Large scale: lower panel).

To complement this analysis of the results, I fitted regression models, as before. Mean values of constants and trend coefficients in each condition, together with optimal values derived from the generating equations are shown in Table 5.2.

Table 5-2 Linear regressions of forecast sequences for each series type: mean values (variances in parentheses) of constants and trend coefficients. Actual values in the generating equations are shown for comparison.

		Constant (a)	*Trend (b)*
Random Series	Actual	10	0
	Small scale	10.66 (7.13)	-0.05 (0.71)
	Large scale	9.98 (7.57)	-0.19 (0.50)
Autoregressive Series	Actual	10	0
	Small scale	9.67 (18.33)	0.17 (0.19)
	Large scale	10.91 (17.24)	-0.03 (0.21)

A two-way ANOVA with dependent variable the slope of the regression equation and independent variables the scale and series type revealed a significant interaction between scale and series type (F (1, 436) = 7.65; p = .006). For the random series, the regression line slope coefficient was again greater for the large scale, while the opposite was true for the autoregressive series. No other effects were found. The same analysis for the regression line intercept indicated that there was a significant interaction between scale and series type (F (1, 436) = 45.98; $p < .001$); a positive elevation bias occurred in the small scale for the random series, while, for the autoregressive series, a

small elevation bias was associated with the large scale display. Differences in residuals haven't reached significance.

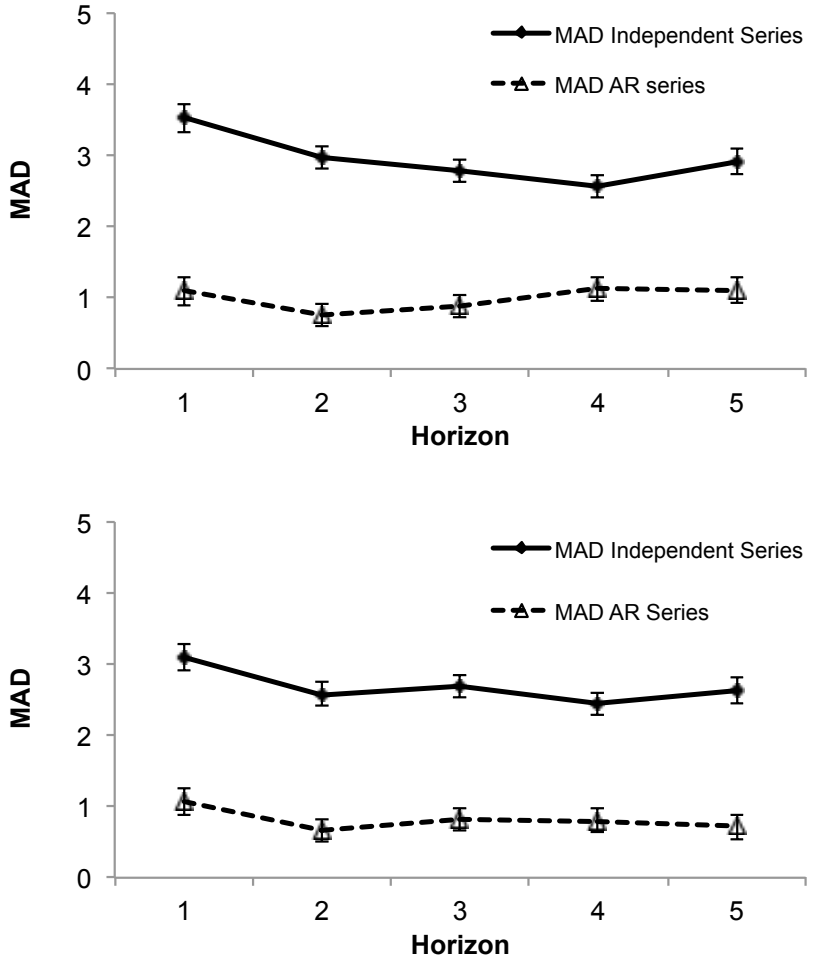

Figure 5.9 Graphs of mean values of absolute differences (together with standard error bars) for the random series (continuous line) and for the autoregressive series (dashed line), for each scale condition (Small scale: upper panel, Large scale: lower panel).

Effects of mean absolute difference I compare differences between mean absolute differences (MAD) from the last data point. Same procedure was followed as before; a three-way ANOVA was run and yielded main effects of

horizon (F (3.56, 1554.40) = 18.29; $p < .001$), a between subjects effect of series' type (F (1, 436) = 243.43; $p < .001$) and a significant interaction between horizon and series' type (F (3.56, 1554.40) = 3.76; $p = .007$). This is also shown in Figure 5.9. There were no other significant effects or interactions in this three-way analysis. Tests of simple effects showed no significant increase of MAD with time horizon for the random and autoregressive series for both scales. Again, these results show participants were sensitive the differences in autocorrelation in the two series types.

Tests of simple effects in the mean absolute differences confirmed these results. There were no significant scale effects or interactions between the two scale types.

To follow this finding up, I examined correlations between successive forecasts for each of the series' type and scale conditions. For autoregressive series, in both scales, high correlations between successive points are observed. For the independent series, the average correlation between successive points is on average around 0.26 (Figure 5.10). This again replicates Reimers and Harvey's (2011) finding that people are sensitive to autocorrelation in series but insufficiently so.

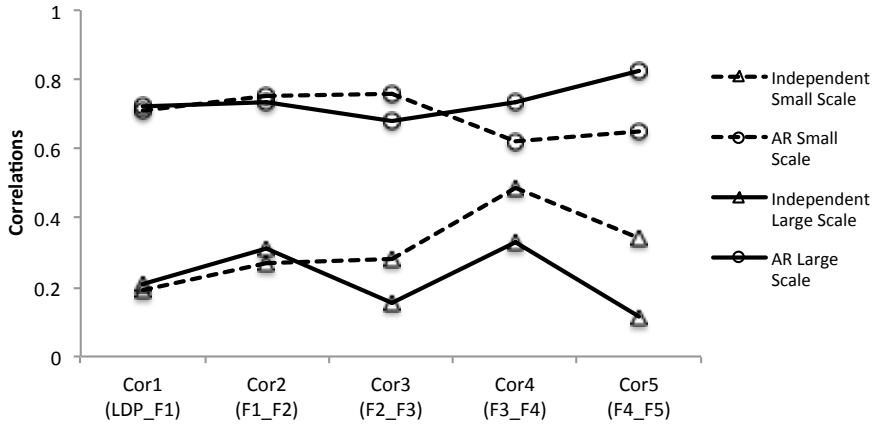

Figure 5.10 Graph of correlation values between successive points for the two series' types, for large scales (continuous line) and for small scales (dashed line).

Fisher's z transformation showed a significant difference in all pairs formed between the random and the autoregressive series. This means that for the different types of series, participants perceived a different degree of autocorrelation.

Cross-experimental comparisons

A four-way within participants ANOVA was employed on the mean absolute errors (MAE) for all five forecast horizons, with independent variables those of series' and scale type as well as experiment type (Experiment 5, Experiment 6). A graph of the overall MAE for the two experiments is shown in Figure 5.11 for each series' type and scale. This shows accuracy decreasing with increasing horizon in both experiments, while errors are constantly higher for Experiment 6, where Gaussian noise was used as a noise term.

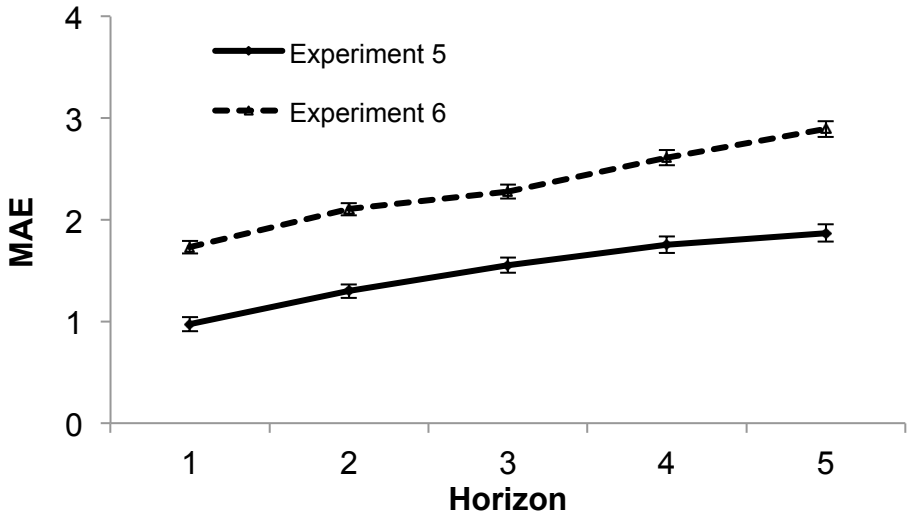

Figure 5.11 Graph of mean values of absolute error (together with standard error bars), for Experiment 5 (continuous lines) and for Experiment 6 (dashed lines).

The four-way ANOVA yielded a main effect of horizon (F (3.69, 2926.98) = 88.94; $p < .001$), and analysis using polynomial contrasts showed that it contained a significant linear component (F (1, 792) = 236.38; $p < .001$), signifying that error increased with an increase in the forecast horizon. There

was also a significant interaction between horizon and Experiment type (F (3.69, 2926.98) = 36.28; $p < .001$), showing the more rapid increase of error with horizon for the second experiment. A three-way significant interaction between horizon, series type and experiment type (F (3.69, 2926.98) = 7.38; $p < .001$) was found, denoting the differences in forecasting performance, which stemmed from the noise type introduced in the different experiments; error increase with horizon was more rapid for both series in the second experiment. Finally, there were main effects of experiment type (F (1, 792) = 151.41; $p < .001$), series' type (F (1, 792) = 144.22; $p < .001$). Scale type yielded no significant effects or interactions. Thus there was no support for H_2.

A four-way within participants ANOVA was employed on the mean absolute distances (MAD) for all five forecast horizons, with independent variables those of series' and scale type as well as experiment type (Experiment 5, Experiment 6). A graph of the overall MAD for the two experiments is shown in Figure 5.12 for each series type and scale. This shows absolute distances being constantly higher for Experiment 6, where Gaussian noise was used as a noise term.

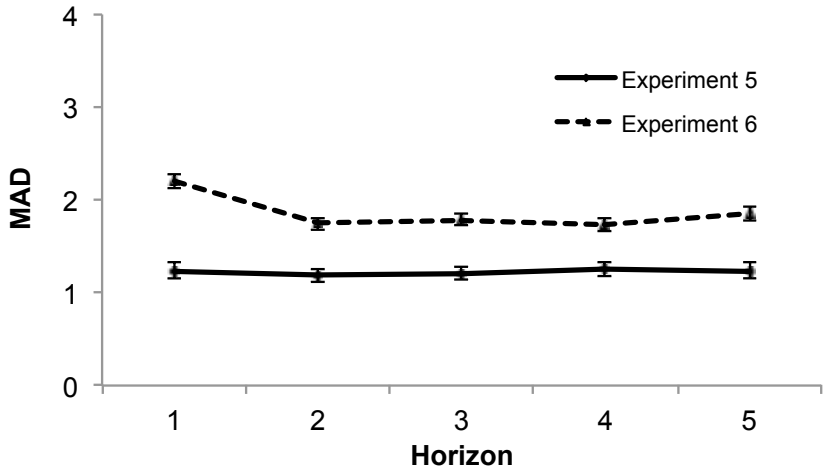

Figure 5.12 Graph of mean values of absolute distances (together with standard error bars), for Experiment 5 (continuous lines) and for Experiment 6 (dashed lines).

The four-way ANOVA yielded a main effect of horizon (F (3.74, 2963.18) = 5.60; $p < .001$), and analysis using polynomial contrasts showed that it contained a significant linear component (F (1, 792) = 4.36; $p < .05$). There was also a significant interaction between horizon and Experiment type (F (3.74, 2963.18) = 5.11; $p < .001$). A two-way significant interaction between horizon and series type (F (3.74, 2963.18) = 5.29; $p < .001$) was found, denoting the differences in forecasting behaviour between random and autocorrelated series. Finally, there were main effects of experiment type (F (1, 792) = 70.59; $p < .001$) and series' type (F (1, 792) = 232.72; $p < .001$). Scale type again yielded no significant effects or interactions.

Discussion

In this experiment; there was no support for H_1: in contrast to Experiment 5, there was no evidence that a smaller scale selectively benefitted forecasting from autoregressive series over forecasting from independent series. Neither was there any evidence to support H_2: there was no evidence that the smaller scale selectively benefitted forecasting from series with Gaussian noise over forecasting from series with uniform noise.

Nevertheless, the experiment did produce a number of significant findings; for example, the findings confirmed that people are sensitive to series autocorrelation but are insufficiently sensitive to it.

5.3 Summary and General Discussion

In the current chapter, I examined the influence of presentation scale on forecasting accuracy and corresponding anchoring behaviours. This was achieved by employing the scale dimensions used in a previous study by Lawrence and O'Connor (1992). This allowed for comparisons between the two studies. In the current chapter, different time series (autoregressive and independent ones) were used to examine the generalizability of previous findings.

Based on Lawrence and O'Connor's (1993) suggestion that smaller scales lead forecasters to expect a wider range of outcomes, I tested two hypotheses. First, smaller scales should selectively benefit forecasting from autoregressive over

forecasting from independent series containing the same underlying error variance. This was expected because the former series have a wider range of outcomes. Although there was some evidence to support this hypothesis from Experiment 5, there was no support for it from Experiment 6. Second, smaller scales should selectively benefit forecasting from series with Gaussian noise over forecasting from series with uniform noise. Again this was expected because the former series have a wider range of outcomes. A cross-experiment comparison produced partial support for this hypothesis. Nevertheless, the effect of distribution shape was not successfully isolated via the current manipulations because the underlying distributions of noise did not have the same variance, confounding, thus, the variables of distribution variance and shape. Thus, in order for robust conclusions to be achieved, variance should have been equal between experiments. The current manipulation only allows for conclusions between scale and series' types.

Both experiments produced evidence that there are effects of scale type that depend on the type of series being forecast. In both studies, regressions were fitted to individual forecast sequences that participants produced. These showed that constants were higher with independent series when the scale was smaller but were higher with autoregressive series when scale was higher. In addition, forecasting from independent series showed a negative trend when the scale was large (but minimal trend when it was small) whereas forecasting from autoregressive series showed a positive trend when scale was small (but minimal trend when it was large). These findings are intriguing but the reasons

for them are unclear. They do not appear to be amenable to being explained via the sort of effects that Lawrence and O'Connor (1993) propose.

Both experiments confirmed that forecasters are sensitive to the level of autocorrelation in the data series but that they are insufficiently sensitive to it (cf Reimers and Harvey, 2011).

It is noteworthy, though, that in all conditions and experiments, higher errors were associated with the small presentation scale, but these yielded only non-significant numerical differences. If the data contain more abrupt shifts than those used in this experiment, that these numerical differences might attain significance. Thus, further research with data that contain shifts and extreme values, such as, for example, anti-persistent fractal series (see Koutsoyiannis, 2000), could be useful in this respect.

Chapter 6 Judgmental Forecasting from Experience

Overview

Forecasting from real-time experienced streams of interrelated data is a task encountered both in professional and in everyday life forecasting situations. In practice, this is a typical task for traders and other financial experts, who observe time series in real-time dynamic displays; they use a combination of graphs with dynamic input of new prices of stock market variables that appear in real-time in the screen. Then, they go on to take their investment decisions relying on their anticipations about market developments. Of particular interest are the so-called "rally periods", where prices are constantly rising or falling in real-time.

Managers operate in an experiential manner as well. Their core competence is the ability to forecast crucial developments at a very early stage (see for example, Nuthall, 2001). This is why research related to strategic planning acknowledges the importance of the need for prompt and efficient assimilation of incoming information (Armstrong, 1982; Straatemeier, Bertolini and Brommelstroet, 2010).

Policy makers receive real-time information for indicators such as GDP growth and inflation indexes and then come up with their decisions for future policies. Monetary policy decisions, for example, are taken in real-time with the use of

judgment and models that assess current and future economic conditions (for relevant nowcasting econometric models, see for example Giannone, et al., 2004). Weather forecasters also make use of their judgment in real-time settings (see for example a relevant experiment by Lusk and Hammond, 1991).

It is not only in professional environments where people experience real-time sequential changes in a variable but also in everyday life. People often need to deal with streams of information that they receive over time (for example, the weather, prices in the supermarket, and so on). This kind of dynamic input may produce judgments arising from specific forecasting strategies. Research in the domain of risky choice from sequential sampling suggests that big discrepancies exist between decisions from experience and decisions from description (Hertwig, Barron, Weber and Erev, 2004). The experimental paradigm of experience-based decisions presents values to the observer, which are received in a sequential manner and a decision is made on the basis of this information. Nevertheless, these values are not necessarily interrelated as in the time series paradigms found in the forecasting literature: they are typically independent. However, two of the most important findings in this area could be of interest in forecasting tasks from experience. First, the likelihood of rare events is often underestimated and, second, recency effects operate in most cases (e.g. Fiedler and Juslin, 2006, p.6).

Here I ask whether similar effects arise with experiential forecasting from a stream of sequentially presented and interrelated stimuli. There has been no published research on this issue in the forecasting literature. In judgmental

forecasting studies static approaches involving graphs or tables have been examined (for tabular versus graphical format effects see Bolger and Harvey, 1996) but experiential forecasting paradigms have been scarce. Only Wagenaar and Timmers (1979) introduced a novel experiential setting, where subjects had to forecast the growth of a series via the pond and the duckweed paradigm (i.e. the representation of duckweed multiplying itself in a pond). This unique paradigm in the forecasting literature tested judgmental forecasting performance in an experiential, non-numerical way with the use of blocks instead of graphs or tables. Results showed that participants damped the exponential trends as in static paradigms. This finding suggests that in experiential forecasting paradigms, trend damping and anti-damping biases still operate as they do in graphical and tabular displays (e.g. Harvey and Bolger, 1996).

There is loosely related research in areas such sequential learning and perceptual choice. In sequential learning tasks (Gureckis and Love, 2010), humans have been found to use simple associative mechanisms (i.e. based on direct associations) to learn, for example, a sequence of numbers. This finding suggests that in experiential forecasting tasks, the autocorrelation illusion (e.g. Reimers and Harvey, 2011) may still operate as in judgmental forecasting tasks from description (i.e. graphs and tables). This account is further strengthened by research in perceptual choice (Tsetsos et al., 2012), which provides evidence for recency effects. Tsetsos et al's (2012) data showed that observers base their estimations on the last set of observations, thereby producing recency effects. Finally, other researchers within the risky choice domain have proposed that the

representativeness heuristic can account for their experimental findings (see Juslin and Fiedler, p. 137, 2004; Juslin et al., 2004). Thus, trend-damping, sensitivity to autocorrelation and representativeness phenomena seem to operate in experientially based settings. These are the same phenomena that are generally acknowledged to operate in judgmental forecasting from static displays.

Thus, these findings suggest that a forecaster, when encountering an experiential forecasting task might be still prone to the biases identified in the classical literature of judgmental forecasting with static tasks. These forecasting tasks have revealed several robust phenomena in forecasters' performance, namely, trend damping, sensitivity to autocorrelation and noise introduction.

Here, I propose a new way to directly investigate whether judgmental forecasting biases from graphs are present when the forecaster is experiencing a time series. I introduce a simple task, where the forecaster experiences time series instead of observing them in static displays. In this task, successive values of the series are presented individually as a sequence of bar charts. At the end of this presentation, the observer has to make forecasts for the next values. The structure of underlying time series can be modified to investigate the three robust phenomena outlined above. Therefore, for the investigation of trend damping, participants will be presented with trends of different directions, gradients and noise levels; then, noise introduction effects will be assessed using series of various noise levels; and, lastly, the exploration of sensitivity to autocorrelation will employ series with different autocorrelations.

As in every novel experimental set-up, there is a number of variables which will require parameterisation by the experimenter. Unfortunately, there are no suggestions as to what the optimal values for these parameters are since there is no previous similar research. For example, there might be screen margin effects that affect forecasting performance, such as those found in Lawrence and Makridakis' (1989) research in a static setting. Alternatively, there might be an optimal speed at which successive data should be displayed for the forecaster to perform well. In these experiential settings, the effect of display time will be of particular interest. Will reduced time between successive stimuli enhance or impair forecasting accuracy? Hypotheses about this must be built on research from other fields where successive presentation of stimuli has been examined tested. Alvarez and Cavanagh (2004), for example, used a visual search task to determine optimal speed presentation. They showed that participants reached maximum accuracy at 450 milliseconds. More recently, Kiani, Hanks and Shadlen (2008), who studied direction discrimination tasks in monkey populations, suggested that accuracy levelled off from 500 milliseconds onwards. Their subjects' performance was not significantly different when stimuli were presented for 500 milliseconds and 1000 milliseconds. This finding reinforces that of Alvarez and Cavanagh (2004). In addition, Woodman, Vogel & Luck (2001) showed that visual search for 500 millisecond displays remained efficient even when visual working memory is fully occupied. These findings from perceptual studies suggest the 500 milliseconds benchmark as a threshold between slow and fast displays.

Another strand of research in decision-making has investigated whether fast or slow displays change accuracy. There, fast displays were found to decrease judgement accuracy (for a review, see Edland & Svenson, 1993). On the other hand, other studies have shown that stress, and, thus, fast displays, can improve performance (e.g. Harvey et al. 1992) because subjects use their cognitive resources more efficiently. In the specific experiential task reported here, fast displays could impair perception of changes in series: for example, the time steps in a trended series. On the other hand slow displays could cause problems for participants in remembering a set of previous data points (e.g. to judge the mean of a series). The effects of speed of display are likely to depend on the characteristics of the task and the underlying series. However, in the simple displays used here, high-speed displays are likely to impair forecasting performance.

In summary, experiments reported in the current chapter were designed to investigate whether well-documented phenomena in the forecasting literature still appear when forecasters experience data points individually via dynamic bar chart displays.

Here I test the hypothesis (H_1) that the phenomena found with static displays will be also obtained with dynamic ones and the hypothesis (H_2) that faster speeds will impair performance.

6.1 Experiential forecasting from upward trends (Experimental Study 7)

This study was explicitly designed to test the forecaster sensitivity to trend damping and speed display in an experiential setting. The setting was designed according to the specifications described in Chapter 2; sequential bar charts were used to present sequentially the data points of a time series to the forecaster. Both the gradient of the trend as well as the speed of stimuli presentation were manipulated. On the basis of previous reports (e.g., Harvey and Reimers, 2013), participants were expected to dampen steep trends and anti-dampen shallow ones. In terms of the time interval between successive stimuli, the 500 ms benchmark was used to distinguish fast and slow displays.

Thus, in the next section, the following hypothesis will be tested:

H_1: Subjects will exhibit trend damping for steep trends and trend anti-damping for shallow trends.

H_2: Presentation speed will affect forecasting performance: fast displays will impair accuracy.

6.1.1 Method

In this experiment, I used the experiential forecasting task described in Chapter 2. Participants produced two forecasts at the end of a sequence that had 30 points.

Participants

A total of 120 undergraduate students, (45 male, 75 female, age M = 21.46, SD = 3.22), took part in the experiment. Participants were recruited from University College London. Participants were not paid for their time. Instead, they were told the five most accurate participants would receive £5.

Design

The study employed a 2 trend gradients (shallow, steep) x 2 speed displays (slow, fast) x 2 time-periods (forecast 31, forecast 32) (see Table 6.1). A total of 120 undergraduate students participated in the experiment, thirty in each of the four conditions. Each participant was tested on one trial and gave two successive forecasts.

Table 6-1 Experimental design for Experimental study 7

Speed/Trend gradient	Interval = 900 *ms* (*Slow*)	Interval = 300 *ms* (*Quick*)
Steep	Condition 1	Condition 2
Shallow	Condition 3	Condition 4

Stimulus materials

To construct the shallow trended series the equation $X_t = 2t$ was used. The steep gradient trended series was constructed by using the equation: $X_t = 4t$. So, each step of the shallow gradient series was equal to 2 and of the steep gradient series was equal to 4. Thus, in a 0 to 150 vertical axis chart, the last data point for the shallow trend was found at a value of 60 and for the steep trend at a

value of 120 (see Figure 6.1 for a screenshot of the experiment). The series presented to participants were noise free and series' data points were presented graphically in a sequential manner, with time intervals between successive data points equal to 300ms or 900ms, depending on the speed condition to which the participant belonged to (fast or slow).

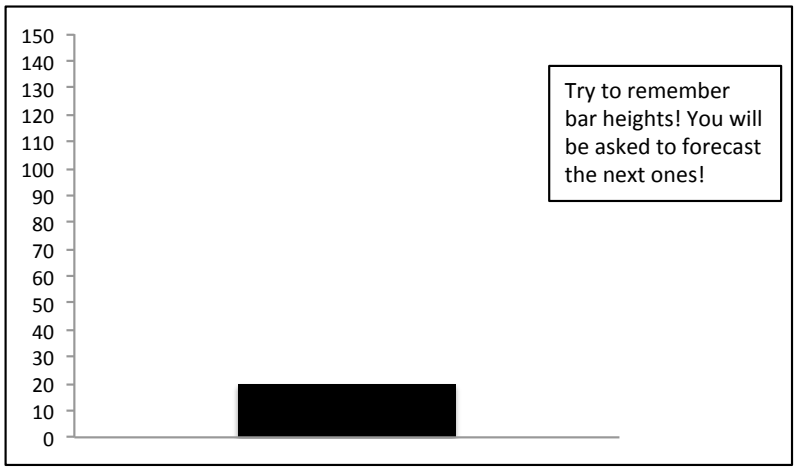

Figure 6.1 Illustration of the experiment: screenshot of the 5th data point, were the bar-height is at a value of 20 for the steep gradient condition.

Procedure

After participants had agreed to take part in the experiment, they were asked to enter their age and gender in MATLAB. Participants were randomly assigned to one of the four conditions: shallow/slow, shallow/quick, steep/slow, and steep/quick. They then read the following instructions:

"Imagine you are a trader… You are now at Wall Street premises and you are observing a specific stock price in this screen! The stock price values are not presented in numbers. Instead,

they are presented with the use of bar charts. The greater the height of the bar is, the larger the price of the stock. A first bar appears in your screen with the initial price. When the stock price changes (it does within seconds in the stock market), the next bar appears in your screen. The previous one disappears! At the end of the task and after observing approximately 30 consecutive stock price changes, you will have to predict the height of the next two bars (i.e. stock prices) by mouse-clicking the height of the bar. Will the price of the stock increase or decrease? Will it remain the same? Your prediction will show whether you are appropriate to become a trader! If you are among the top 5 traders, then, you will receive a 5 pound award!"

Once they had finished reading the instructions, they pressed the space bar for the experiment to begin. Each participant saw 30 data points in a sequential manner. The goal of the experiment was for them to forecast points 31 and 32 by using the mouse to click at the height they thought the next points would be. After participants had indicated their predictions, they were debriefed and thanked for their time.

6.1.2 Results

To measure forecasting performance, mean absolute error was calculated (MAE). Participants whose MAE values were more than three standard deviations from the mean of the group were excluded and replaced. To determine whether trend damping occurred, I used the methodology associated with the exploitation of the mean signed error measure (MSE), which is calculated for each forecast as the difference between the forecast and the corresponding trend value. Trend damping occurred when significantly higher errors were associated with the more distant horizons. Signed error was again

calculated by subtracting the forecast from the optimal value of the series (in this case the optimal and the real values coincide).

Forecasting performance Participants' MAE scores were used as an input to a 2 trend gradients (shallow, steep) x 2 speed conditions (slow display, fast display) x 2 horizons (forecast 31, forecast 32) repeated-measures ANOVA. Overall, participants displayed effects of horizon (F (1, 116) = 8.77, p < .001), suggesting, thus that overall MAE for horizon 2 were larger than MAE for horizon 1 ($M_{Horizon1}$ = 7.71 *vs* $M_{Horizon2}$ = 8.48). A main effect of trend gradient was also found (F (1, 116) = 5.96, p = .016), with post-hoc tests showing that those in the shallow gradient produced a larger MAE overall compared to those in steep gradient trends ($M_{Shallow\ Trend}$ = 9.52 *vs* $M_{Steep\ Trend}$ = 6.67). There was a horizon x trend gradient interaction (F (1, 116) = 6.42, p = .013). For shallow trend gradients, participants' MAE increased faster overall from period 1 to period 2 ($M_{Shallow,Horizon1}$ = 8.80 and $M_{Shallow,Horizon2}$ = 10.24 *vs* $M_{Steep,Horizon1}$ = 6.61 and $M_{Steep,Horizon2}$ = 6.73), suggesting the possibility that, with shallow trends, participants had more space to mark their forecasts and, in line with Lawrence and Makridakis (1989), this affected forecasting performance (Figure 6.2).

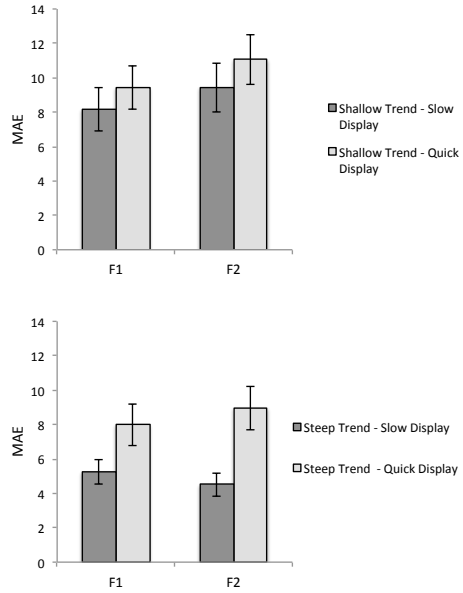

Figure 6.2 Graphs of mean values of absolute error (together with standard error bars) against forecast horizon for the two different types of trended series, shown from upper to lower panels in the order a) shallow trend b) steep trend. In shallow conditions, participants produced larger MAEs than in steep conditions.

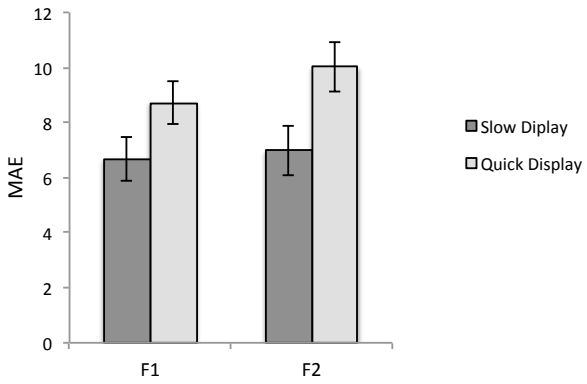

Figure 6.3 Marginal means of mean absolute error (together with standard error bars) for slow (dark grey) and quick (light grey) displays against forecast horizon. Overall, in fast speed conditions participants produced larger MAEs.

Effects of speed A main effect of speed was found (F (1, 116) = 4.71, p = .032), with post-hoc tests showing that those in the high-speed conditions produce larger MAE overall as compared to those in low-speed conditions (M_{Slow} = 6.83 *vs* M_{Fast} = 9.36) (Figure 6.3).

Signed errors analysis Participants' MSE were also used as an input to a repeated-measures ANOVA, same as before. Overall, participants displayed effects of horizon (F (1, 116) = 18.51, p < .001), suggesting, thus that overall MSE for horizon 2 was higher than MSE for horizon 1 ($M_{Horizon1}$ = 1.51 *vs* $M_{Horizon2}$ = 2.74). A main effect of trend gradient was also found (F (1, 116) = 39.45, p < .001), with post-hoc tests showing that those in the shallow gradient produced a positive MSE whereas those in steep gradient trends produced negative signed errors ($M_{Shallow\ Trend}$ = 7.20 *vs* $M_{Steep\ Trend}$ = -2.95). There was a marginally significant horizon x trend gradient interaction (F (1, 116) = 2.86, p = .09). For shallow trend gradients, participants' MSE was positive and increased faster from period 1 to period 2, whereas in steep trends, MSE was negative, decreasing with horizon ($M_{Shallow,Horizon1}$ = 6.34 and $M_{Shallow,Horizon2}$ = 8.06 *vs* $M_{Steep,Horizon1}$ = -3.32 and $M_{Steep,Horizon2}$ = -2.57), suggesting that, with shallow trends, the effect could be characterized as anti-damping. The same was not true for steep trends, where trend damping was expected to occur (see also Figure 6.4).

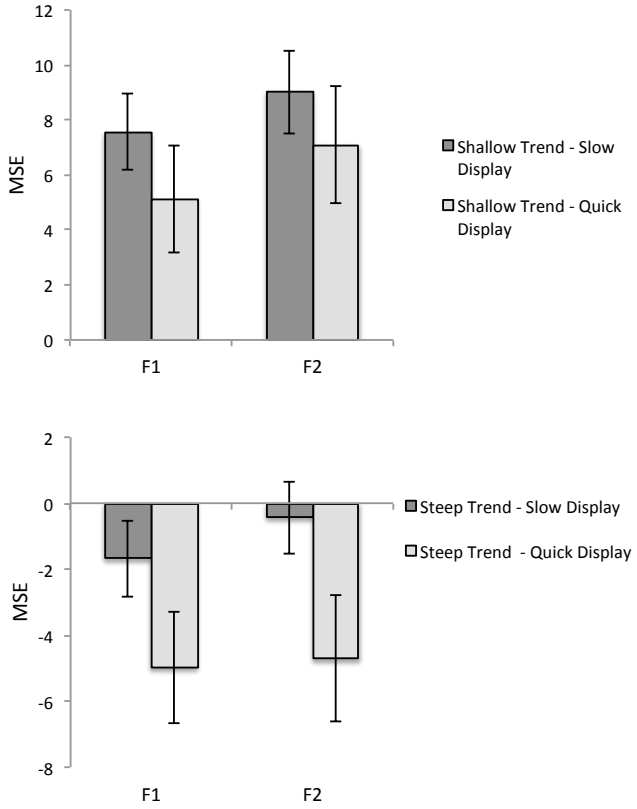

Figure 6.4 Graphs of mean values of signed error (together with standard error bars) against forecast horizon for the two different types of trended series, shown from upper to lower panels in the order a) shallow trend b) steep trend. In shallow conditions participants produced positive errors while for steep conditions errors were negative. While for the shallow trend errors increased from horizon 1 to horizon 2, which is evidence for anti-damping, the same was not true for steep trends, where error decreased with time horizon.

Discussion

Summarizing the results of this experiment, one could say that horizon, gradient and speed variables all had effects on forecasting performance.

Forecasting performance MAE analysis showed overall greater errors for the most distant horizon and the shallow trend gradient, with errors increasing faster with horizon in the case of shallow trends. Larger errors for the most distant horizon were expected; this finding is in accordance with forecasting research with the use of graphs (see for example Bolger and Harvey, 1993). Nevertheless, larger and faster growing errors for the shallow rather than the steep trends condition was an unexpected result; in tasks where graphs are used, larger errors are associated with steeper trends.

Why did participants in shallow trend conditions produce larger errors? Possibly, it was because, in these conditions, participants had more space to mark their forecasts, allowing, thus, more error to be introduced in their performance. If absolute error is correlated with available space, then, it should be the case that as gradient decreases, average absolute error increases. Research by Lawrence and Makridakis (1989) revealed boundary biases in a judgmental forecasting task with the use of graphs. They showed that the greater the space on the graph above the plot of a linearly trended time series, the higher the forecast tended to be. So, boundary biases are likely to have been responsible for the effect found here.

Effects of presentation speed The MAE analysis also showed that those in the fast conditions produced larger errors compared to those in slow conditions. Faster speed impaired performance. This finding is in accordance with hypothesis 2. One can suppose that participants attempted to capture the nature of the data generation process, then formed a representation of the characteristics of that process, and, finally, attempted to generate an accurate forecast from this representation. Participants experiencing the data slowly had more time to detect the nature of the data generation process and, thus, their judgement extrapolations were more representative of the data series (e.g. Alvarez and Cavanagh, 2004; Kiani et al., 2008). As a result, they benefited from increased accuracy.

Elevation and damping biases MSE analysis showed overall positive errors for shallow trend gradients and negative errors for steep trend gradients; forecasts were higher than the real values in shallow trends and lower than the real values in steep trends. This means that in both cases elevation effects were present.

Were the usual damping effects present as well? It is clear that positive signed errors increased with horizon for shallow trends, which is evidence for anti-damping (error increase with horizon). However, with steep trends, the negative signed error decreased marginally with horizon. This result can be explained in two ways. Either there were confounding elevation as well as damping effects between horizons 1 and 2, which eventually masked damping for steep gradients or trend damping never occurred. It is difficult to disentangle

elevation from damping effects in this case in order to decide which of these two accounts is true.

A method that distinguishes elevation and damping effects more clearly is needed. Here, I introduce a new measure to tackle this issue: the measure of implied time-steps. To obtain implied time steps for each horizon, I calculate the first horizon implied time step as first horizon forecast minus last given data point and the second horizon implied time step as second horizon forecast minus first horizon forecast. According to this calculation, participants in the shallow condition produced a first forecast far from the trend ($\delta F_{\text{Shallow (Horizon1-Last Datapoint)}} = 8.34$, much greater than the given series step, which was equal to 2) and then implied (with their second forecast) that the trend step was smaller and comparable to that of the given series ($\delta F_{\text{Shallow (Horizon2-Horizon1)}} = 3.72 > 2$, but 3.72 is much lower than the first implied time step, which was equal to 8.34).

Did the two horizon forecasts differ cognitively in the way those where produced? Was the first horizon forecast just an approximate estimate of the height (influenced by the margins), and the second horizon forecast the implied step of the trend (also influenced by the margins but significantly less than F1)? By looking at the steep trend results as well, one should be able to confirm whether an account like that could be used to interpret these findings. For steep trends, thus, the first horizon forecast is placed according to the same rational: now, it is shifted below the trend due to the upper margin effects ($\delta F_{\text{Steep (Horizon1-Last Datapoint)}} = 0.57$, much lower than the given series step, which was equal to 4).

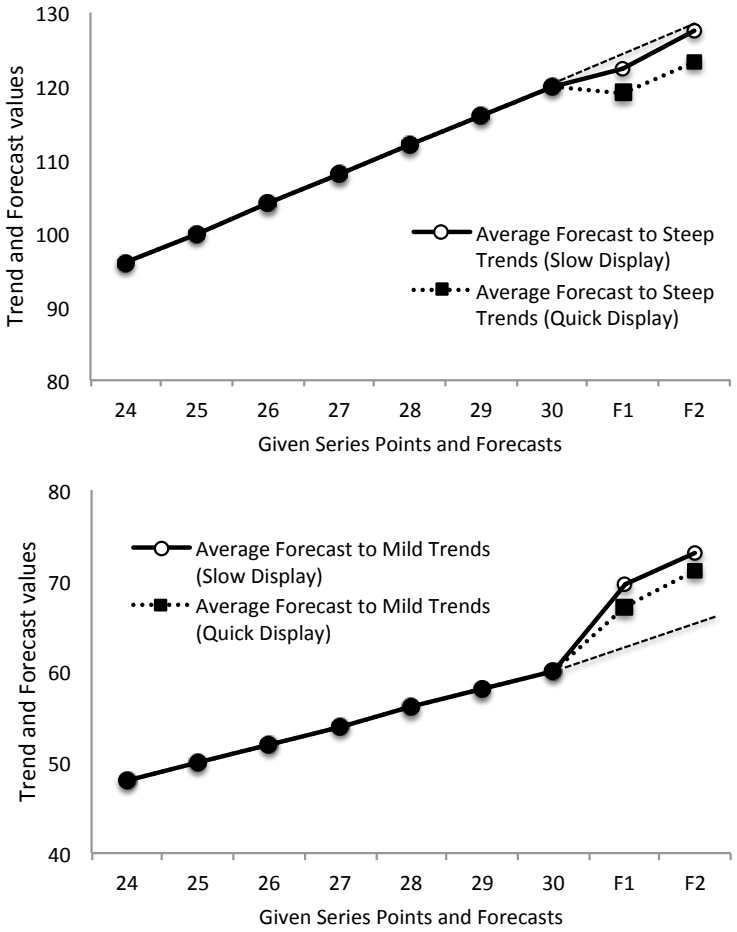

Figure 6.5 Graphs of average forecasts against forecast horizon for the two different types of speeds, shown from upper to lower panels in the order a) steep trends b) shallow trends.

Table 6-2 Implied time steps for each horizon

Real time step of the given series	Implied time step of the trend *F1*	Implied time step of the trend *F2*
Steep trend (*Real step = 4*)	0.57	4.72
Shallow trend (*Real step = 2*)	8.34	3.72

Participants' first horizon forecast is produced as if participants damped the trend for the first forecast enough so to allow space for a larger second forecast: the second horizon forecast now implied a comparable time step to that of the given series ($\delta F_{Steep\ (Horizon2\text{-}Horizon1)} = 4.71 > 4$, but 4.71 much greater than 0.57). Table 6.2 presents the average implied time steps for each series type and forecast horizon and Figure 6.5 the average forecasts in conjunction with the given series.

To sum up, implied time step differences in direction between series of different gradients might be related to the screen margins and space availability. Moreover, implied time step differences between forecast horizons 1 and 2 might be related to qualitative differences between horizon 1 and 2 forecasts. Results from this first experiential experiment are inconclusive as to whether trend-damping biases occurred. Nevertheless, according to the official definition of trend damping and anti-damping, only trend anti-damping occurred.

6.2 Experiential forecasting from downward trends (Experimental Study 8)

In this experiment, the direction of the trend is opposite to the one used before. Thus, downward trends were investigated. According to the literature (see for example Harvey and Reimers, 2013), one should expect even more pronounced damping effects for these types of trends. Also, according to findings from Experiment 7, fast displays should impair performance. Thus, hypotheses for this study (H_{2A} and H_{2B}) were the same as in Experimental Study 7.

6.2.1 Method

The method was the same as in Experiment 7 except for the difference in the series direction, which was downwarding this time.

Participants

A total of 120 undergraduate students, Age (M = 22.03, SD = 3.57) 57 male, 63 female took part in the experiment. Participants were recruited again from University College London. They were not paid for their time. Instead, they were told the five most accurate participants would receive £5.

Design

The study employed a 2 downwarding trend gradients (shallow, steep) x 2 Speed conditions (slow, fast) x 2 time-periods (forecast 31, forecast 32) mixed design. A total of 120 undergraduate students participated in the experiment,

thirty in each of the four conditions. Each participant was tested in a single experiment comprising one trial, as before.

Stimulus materials

There were two types of series: a downward trended linear series with a steep gradient and a one with a shallow gradient. To construct the shallow trended series, I used the equation: $X_t = -2t$. The steep gradient trended series was constructed by using the equation: $X_t = -4t$. So, the step of the shallow gradient series was equal to 2 and the steep gradient series step was equal to 4. The series presented to participants were noise free and data points were presented graphically in a sequential manner, as before. Thus, in a 0 to -150 vertical axis chart, the last data point for the shallow trend was found at a value of -60 and for the steep trend at a value of -120 (see also Figure 6.6 for a screenshot of the experiment). The series presented to participants were noise free and series' data points were presented graphically in a sequential manner, with time intervals between successive data points equal to 300ms or 900ms, depending on the speed condition to which the participant belonged to (fast or slow).

Procedure

The procedure was exactly the same as before.

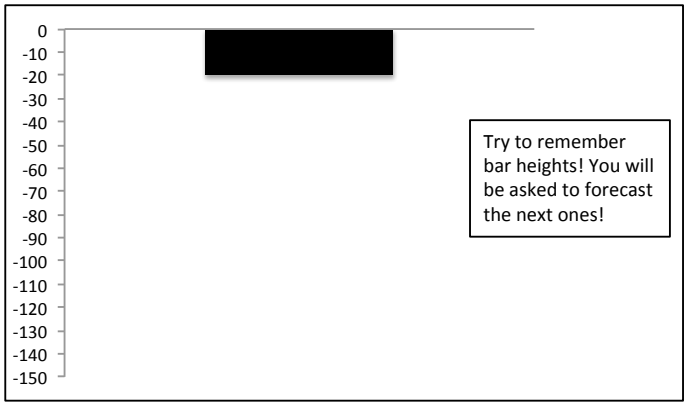

Figure 6.6 Illustration of the experiment: screenshot of the 5th data point, where the bar-height is at a value of -20 for the steep gradient condition.

6.2.2 Results

In this section, effects of MAE and MSE will be studied. Cross-experimental comparisons between Experimental Studies 7 and 8 will be performed as well.

Effects of trend gradient Participants' MAE scores were used as an input to a 2 trend gradients (shallow, steep) x 2 speed conditions (slow, fast) x 2 horizons (forecast 31, forecast 32) repeated-measures ANOVA. Overall, participants do display effects of horizon (F (1, 116) = 116, p = .026), suggesting, that overall MAE for horizon 2 was larger than MAE for horizon 1 ($M_{Horizon1}$ = 7.80 *vs* $M_{Horizon2}$ = 8.39). A main effect of trend gradient (F (1, 116) = 4.62, p = .034) was revealed, with post-hoc tests showing that the shallow trends produced larger MAE overall compared to the steep gradient trends ($M_{Shallow\ Trend}$ = 9.39 *vs* $M_{Steep\ Trend}$ = 6.80). There was a horizon x trend gradient interaction (F (1, 116) = 116, p = .011). For shallow trend gradients, participants' MAE scores

increased overall from period 1 to period 2 whereas in steep trend gradients, MAE decreased marginally ($M_{Shallow,Horizon1}$ = 8.75 and $M_{Shallow,Horizon2}$ = 10.02 *vs* $M_{Steep,Horizon1}$ = 6.84 and $M_{Steep,Horizon2}$ = 6.75), again suggesting that, with shallow trends, participants had more space to mark their forecasts and this affected forecasting errors (Figure 6.7).

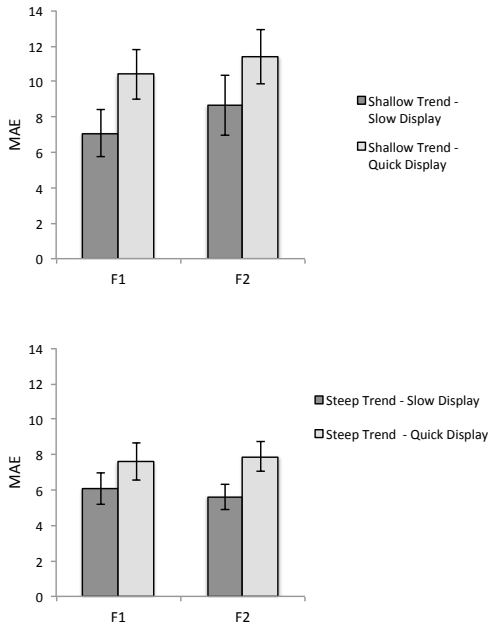

Figure 6.7 Graphs of mean values of absolute error (together with standard error bars) against forecast horizon for the two different types of trended series, shown from upper to lower panels in the order a) shallow trend b) steep trend. In shallow conditions, participants produced larger MAE than in steep conditions.

Effects of speed A main effect of speed was also found (F (1, 116) = 4.24, p = .042), with post-hoc tests showing that participants in the fast conditions

produce larger MAEs overall compared to those in low-speed conditions (M_{Slow} = 6.85 *vs* M_{Fast} = 9.33) (Figure 6.8).

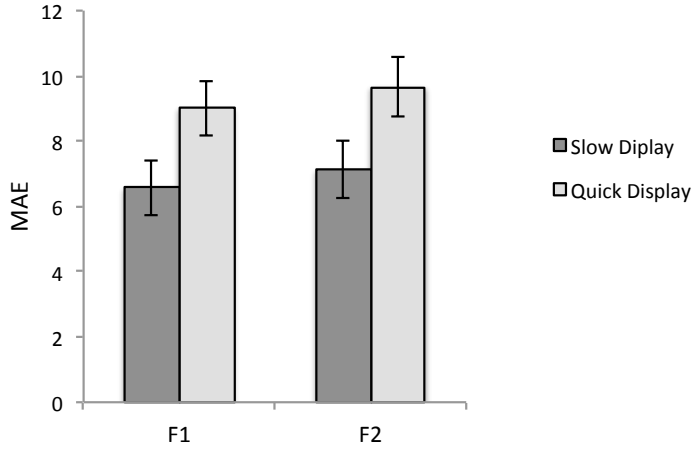

Figure 6.8 Marginal means of mean absolute error (together with standard error bars) for slow (dark grey) and quick (light grey) displays against forecast horizon. Overall, in fast speed conditions participants produced larger MAEs.

Signed errors analysis In the previous experiment mean signed error was calculated by subtracting the optimal value from the forecast, providing negative values for damping in upward trends but positive for damping in downward ones. Here, for the purposes of the cross-experimental comparisons, I reverse the coding of this error. This will allow making direct comparisons of the size of the damping or antidamping effects for upward and downward trends in the two experiments. Thus, here (and for this experiment only) mean signed error is calculated by subtracting the forecast from the optimal value. This way, MSE will show damping for downward trends as a negative value (like it is for upward ones in Experimental Study 7) and antidamping for both upward and

downward trends will be signalled by a positive MSE. Participants' MSE scores were used as an input to a repeated-measures ANOVA, as before. Overall, participants did display effects of horizon (F (1, 116) = 32.79, $p < .001$), suggesting, thus that overall MSE scores for horizon 2 were higher than MSE scores for horizon 1 ($M_{Horizon1} = 1.77$ *vs* $M_{Horizon2} = 3.31$). A main effect of trend gradient was also revealed (F (1, 116) = 49.58, $p < .001$), with post-hoc tests showing that trials with the shallow gradient produced positive MSE whereas trials with the steep gradient produced negative signed errors ($M_{Shallow,Trend} = 8.05$ *vs* $M_{Steep,Trend} = -2.96$) (see Figure 6.9).

Table 6-3 Implied time steps for each horizon

Real time step of the given series	Implied time step of the trend *F1*	Implied time step of the trend *F2*
Steep trend (*Real step = -4*)	-0.47	-5.12
Shallow trend (*Real step = -2*)	-9.07	-3.95

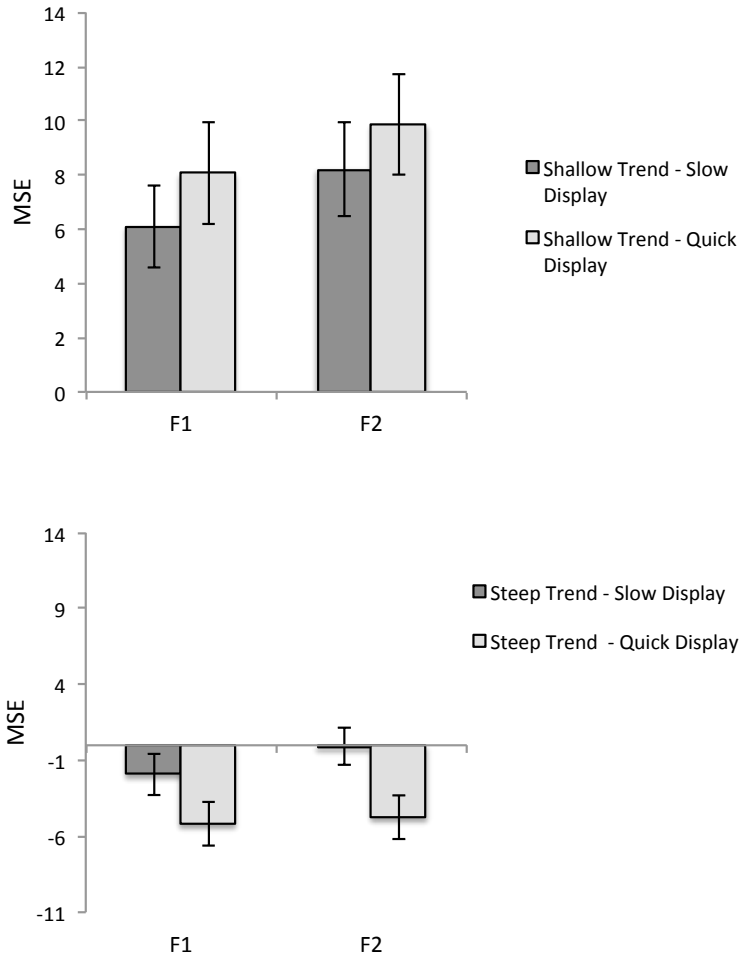

Figure 6.9 Graphs of mean values of signed error (together with standard error bars) against forecast horizon for the two different types of trended series, shown from upper to lower panels in the order a) shallow trend b) steep trend. In shallow conditions participants produced positive errors while for steep conditions errors were negative, as before.

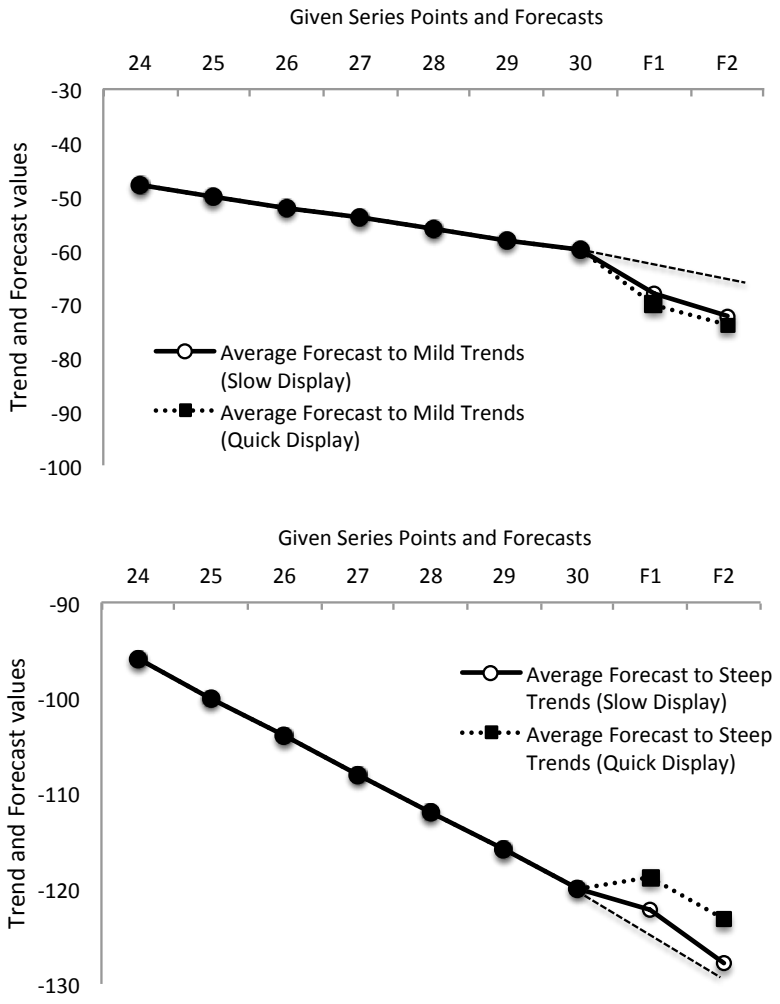

Figure 6.10 Graphs of average forecasts against forecast horizon for the two different types of speeds, shown from upper to lower panels in the order a) shallow trends b) steep trends

Table 6.3 presents the average implied time steps for each series type and forecast horizon and Figure 6.10 the average forecasts in conjunction with the given series.

Cross-experimental comparisons

Participants' MAEs were also used as an input to a 2 directions (upward, downward) x 2 trend gradients (shallow, steep) x 2 speed conditions (slow, fast) x 2 horizons (forecast 31, forecast 32) repeated-measures four-way ANOVA. For MAE, no interactions were found between upward and downward directions suggesting that the average absolute effect of boundaries is the same for both directions. Participants' MSE were also used as an input to a 2 directions (upward, downward) x 2 trend gradients (shallow, steep) x 2 speed conditions (slow, fast) x 2 horizons (forecast 31, forecast 32) repeated-measures four-way ANOVA. Overall, main effects of experiment were found (or alternatively, direction), as expected (F (1, 232) = 17.22, $p < .001$), suggesting, thus that overall MSE for experiment 8 was higher than MSE for experiment 7 (M_{Exp1_Up} = 2.12 *vs* M_{Exp2_Down} = 2.54). Also, on the between participants factor, a main effect of Experiment x trend gradient was found (F (1, 232) = 88.56, $p <$ .001, $M_{Exp1Shallow}$ = 7.20 *vs* $M_{Exp2Shallow}$ = 8.05; $M_{Exp1Steep}$ = -2.95 *vs* $M_{Exp2Steep}$ = -2.97), as expected. For MSE, horizon x direction interactions were found to be significant, (F (1, 232) = 49.77, $p < .001$, $M_{Exp1Horizon1}$ = 1.50, $M_{Exp1Horizon2}$ = 2.74; $M_{Exp2Horizon1}$ = 1.77, $M_{Exp2Horizon2}$ = 3.31). Thus, greater anti-damping occurred in the second experiment (Figure 6.11). A significant three-way interaction for horizon x experiment x trend gradient was also significant (F (1, 232) = 5.29, $p < .022$).

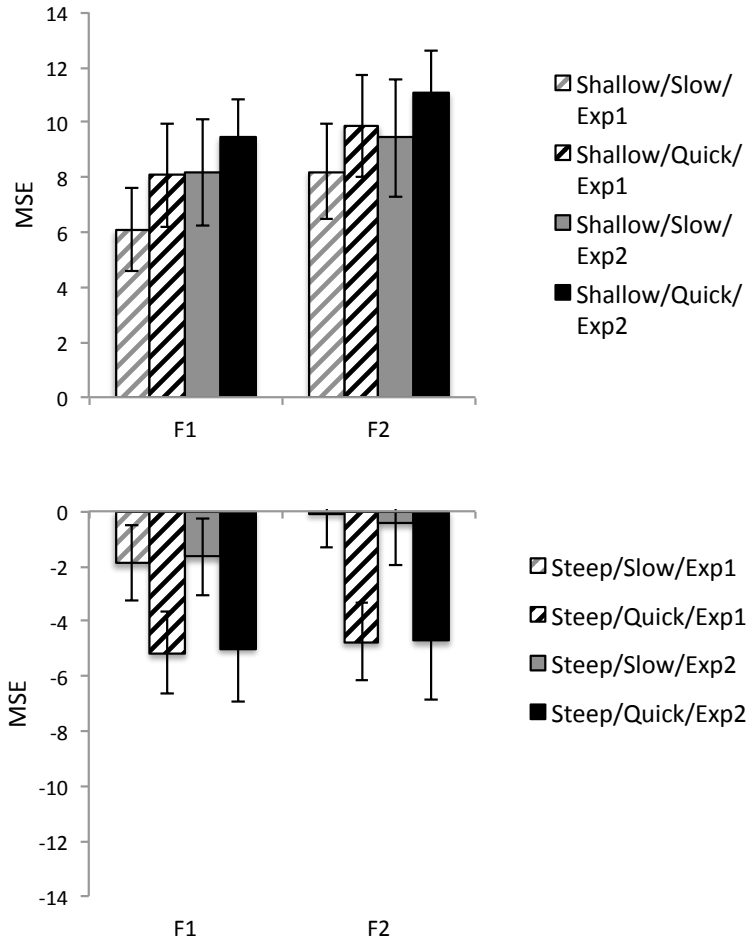

Figure 6.11 Graphs of mean values of signed error (together with standard error bars) against forecast horizon for the two types of trended series shown from top to bottom panels in the order a) shallow trend b) steep trend. Solid bars are from Experiment 7 MSEs, while patterned bars refer to MSEs from Experiment 8. Lighter bars correspond to MSEs for slow displays, while darker bars are for quick display MSEs. In shallow trend conditions, participants produced higher anti-damping for downward trends (Experiment 8) for both speed displays. The same was not true for steep trends where MSEs were of comparable magnitude.

Discussion

Summarizing the results of this experiment, there was confirmation that horizon, gradient and speed variables all have effects on forecasting performance. The MAE analysis showed again overall greater errors for the most distant horizon and the shallow gradients. It also showed that fast display conditions produced larger errors than slow display conditions.

The MSE analysis showed evidence for anti-damping for shallow gradients, while, with steep gradients, trend damping has not occurred. Effects were of comparable magnitude to those in the previous experiment but with trend-antidamping being more pronounced for downward trends in Experiment 8. This is in accordance with evidence from judgmental forecasting from graphs (Harvey and Reimers, 2013; Harvey and Bolger, 1996; Lawrence and Makridakis, 1989; O'Connor et al., 1997). There, effects related to downward trends were found to be more pronounced than for upward ones. This is attributed to the optimism bias (Weinstein, 1980).

Again, it appears that participants were influenced by the screen margins, especially when the first horizon forecast was produced. The effects observed can be interpreted as a combination of elevation and damping biases and they significantly impaired accuracy. It is clear now that upper and lower margins affect forecasting performance in a significant way. It is likely that trend biases were more pronounced for shallow trends because of the greater space, which was available to the forecaster. If this is the case, optimal performance should

be obtained for intermediate trends. To confirm this result, in the next experiment, I will compare shallow, intermediate and steep trends to determine whether biases related to trend are eliminated for intermediate trends.

6.3 Experiential forecasting from intermediate trends (Experimental Study 9)

Speed of the display was set to 1 second in this experiment, because experimental studies 7 and 8 provided evidence that slow displays enhance performance. Three forecasts were requested this time to investigate further the perceived trends that participants' forecasts implied.

H_3: Participants will exhibit optimal performance for intermediate trends.

6.3.1 Method

Participants

A total of 99 participants, 43 male, 56 female (Age M = 30.01, SD = 9.40), took part in the experiment. Participants were recruited from Amazon Mechanical Turk and were paid 0.5$ each.

Design

The study employed a 3 upward trend gradients x 3 time-periods (forecast 21, 22 and 23) between-participants design. A total of 99 participants were

recruited for the experiment (33 in each of the three conditions) from online sources as before. Speed was set to 1 sec between successive stimuli.

Stimulus materials

There were three types of trended linear series, which were constructed by using the equations: $X_t = 3t$, $X_t = 4.5t$, $X_t = 6t$. On a 0 to 150 vertical axis, the last data point, for each of the corresponding trends, was found at 60, 90 and at 120 (for an illustration of those trends, see Figure 6.12). The series presented to participants were noise free and series' data points were presented graphically in a sequential manner, as before.

Procedure

Procedure was the same as before, only this time participants saw 20 data points and were requested to produce three forecasts. This time I used 20 data points for the given series to accommodate for all types of trends in the same screen display.

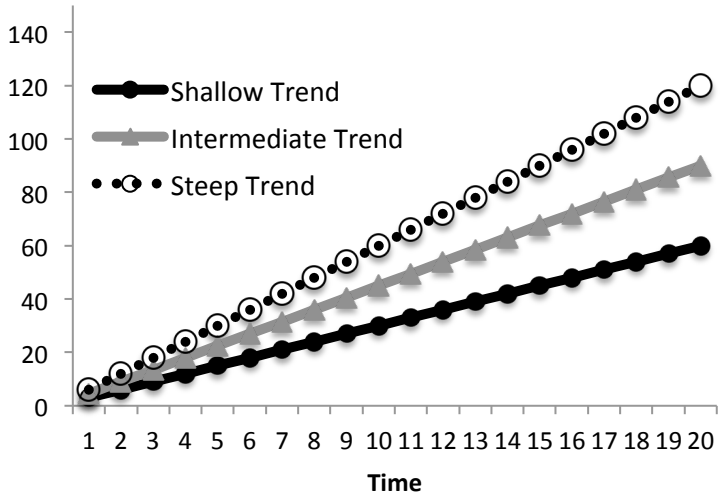

Figure 6.12 Graphical representation of the three gradient conditions in Experimental Study 9.

6.3.2 Results

Forecast performance was measured by MAE and MSE, as before. Here, I also calculated implied time-steps for all horizons, as before; first horizon time step = first horizon forecast – last given data point, second horizon time step = forecast second horizon – forecast first horizon, third horizon time step = forecast third horizon – forecast second horizon. I will first present the analysis for the implied time steps for the first horizon, where I obtained elevation effects in experiments 7 and 8. Figure 6.13 shows the first horizon implied time steps in conjunction with the real time steps of the series for all three gradient conditions. Implied step for intermediate trends approximates more the real time step of the series. This is what was expected according to hypothesis H_3.

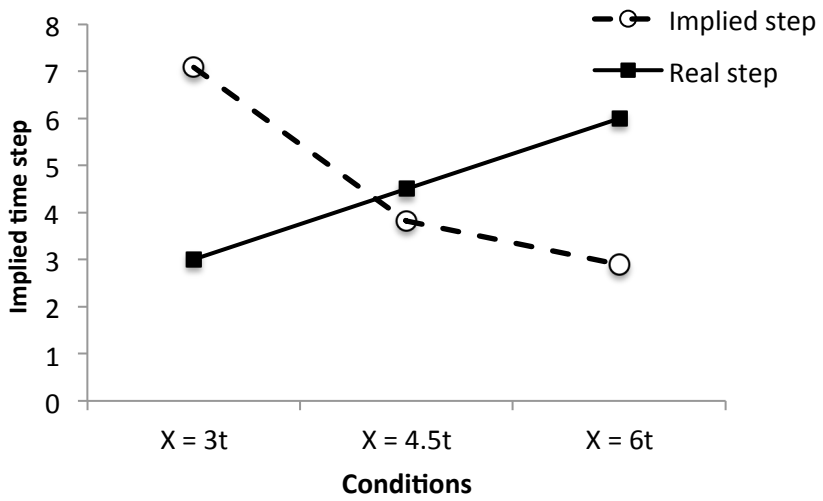

Figure 6.13 Real and implied time steps for the first horizon forecast for all conditions. Participants average implied time step for horizon 1 decreases as trend gradient increases. Best approximation between implied and real time steps is obtained for intermediate trend series X=4.5t.

To confirm whether these differences in implied time steps were significant, I ran a trend gradient x implied time step between-subjects ANOVA with implied time step as a dependent variable. Main effects of trend gradient were marginally significant (F (2, 96) = 2.88, p = 0.06). I will now turn to analyse all forecast horizons implied time steps. Table 6.4 and Figure 6.14 present the averages of implied steps for all horizons.

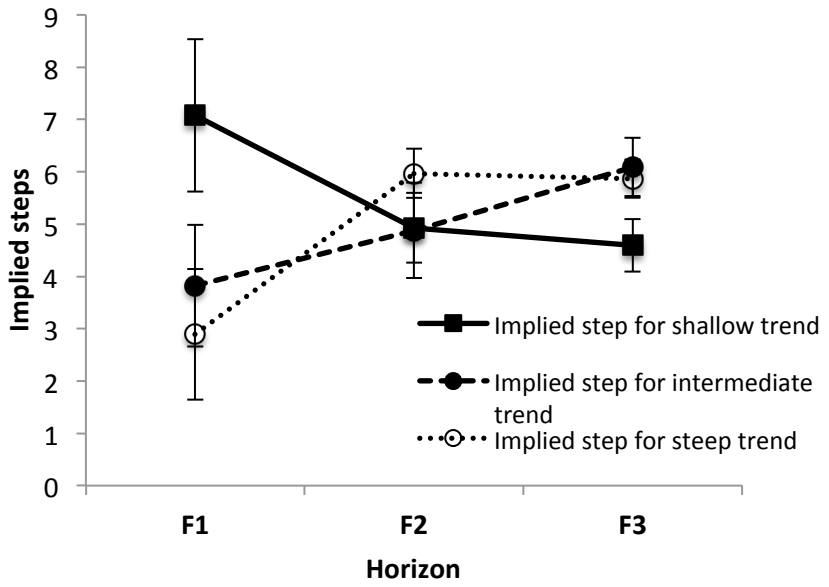

Figure 6.14 Implied time steps for all three horizons. For time steps 1, 2 and 3, implied time step decreased for the shallow and intermediate trends and increased for the steeper ones.

Table 6-4 Implied time steps for each horizon and each gradient

Real time step of the given series	Implied time step of the trend *F1*	Implied time step of the trend *F2*	Implied time step of the trend *F3*
Real step = 3	7.08	4.93	4.59
Real step = 4.5	3.82	4.88	6.09
Real step = 6	2.89	5.97	5.87

These findings suggest that for shallow trends (real time step = 3), participants misplaced their forecasts for horizon 1, implying a step equal to 7.08. They then decreased their time step estimations for horizons 2 and 3 to 4.93 and 4.59 respectively. Participants' performance for shallow trends has been characterized by elevation and anti-damping effects (horizon 1), while anti-damping characterized the rest of the forecasts for horizons 2 and 3. For

intermediate trends (real time step = 4.5), while participants approximated the real time step for horizon 1, they then increased their implied time step estimations for horizons 2 and 3, to 4.88 and 6.09 respectively. Thus, they showed no elevation or anti-damping effects for horizon 1, while anti-damping characterized the rest of the forecasts for horizons 2 and 3. This is in accordance with hypothesis 3. Finally, for steeper horizons (real time step = 6), participants again misplaced their forecasts for horizon 1, implying a time step of 2.89 but increasing this for their estimations for horizons 2 and 3 to 5.97 and 5.87 respectively. There was no evidence of damping. This is in accordance with findings for experiments 7 and 8.

One-sample t-tests were used to compare each participant's implied time steps within the criterion values (i.e. 3, 4.5, 6). For shallow trends, participants' implied time steps were significantly higher than the actual time-step (t (32) = 2.80, p = .009, for implied time step 1; t (32) = 2.88, p = .007, for implied time step 2; t (32) = 3.15, p = .003, for implied time step 3). For intermediate trends one-sample t-tests with 4.5 as a criterion value showed no significant differences for implied time steps 1 and 2 but, for implied time step 3, the time step was found to be significantly higher than the criterion value (t (32) = 2.85, p = .007, for implied time step 3). This means that for intermediate trends, participants were accurate in their implied time step predictions for the first horizons; in the last horizon they showed significant anti-damping. Finally, for steep trends, implied time steps for horizon 2 and 3 did not differ significantly from the actual time step. It was only in horizon 1 where forecasts lied

significantly lower than the actual time step (t (32) = -2.48, p = .018, for implied time step 1). Thus, damping did not occur for horizons 2 and 3 but misplacement took place for horizon 1.

Signed errors analysis To further examine these effects, and their significance, participants' MSEs were entered to a 3 trend x 3 horizons repeated-measures ANOVA. MSEs correspond to the difference between the forecast and the real value of the series in this case (see Figure 6.15). Positive signed errors show anti-damping behavior and negative ones show damping behavior; signed errors, which are close to zero, show no effects. Overall, main effects of horizon were found (F (2, 192) = 10.18, p < .001), suggesting, thus that overall signed error was higher as horizon increased (M_{F1} = -0.11 *vs* M_{F2} = 0.55 *vs* M_{F3} = 1.91). A main effect of trend gradient was found (F (1, 96) = 9.11, p < .001), (M_{3t} = 7.26 *vs* $M_{4.5t}$ = -0.53 *vs* M_{6t} = -4.37), with post hoc tests showing significant differences in signed errors between shallow and intermediate, shallow and steep but not between intermediate and steep trends. For MSE, there was also a significant horizon x trend gradient interaction (F (4, 192) = 3.97, p = .004), suggesting faster increase of error for the shallow trends.

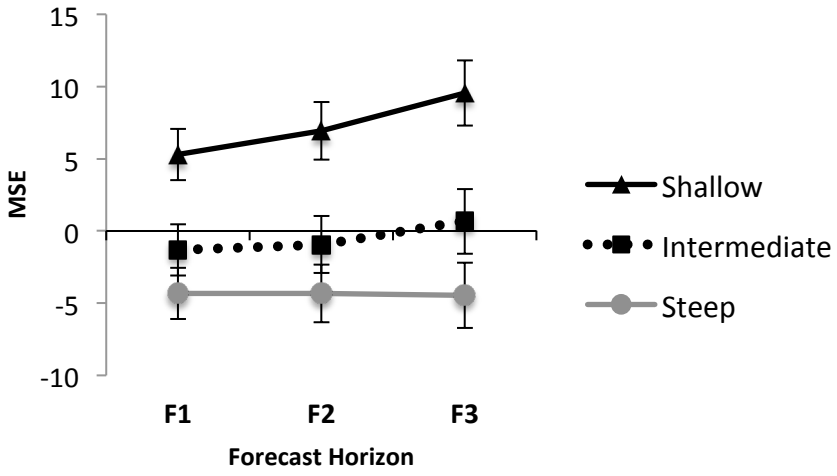

Figure 6.15 Graph of mean values of signed error for each trend gradient. Intermediate trends produce the best estimates in terms of MSE for all horizons.

Absolute errors analysis Participants' MAEs were also used as an input to 3 trend gradients x 3 horizons repeated-measures ANOVA. Overall, main effects of horizon were found (F (2, 192) = 17.66, p < .001), suggesting, thus that overall MAE was higher as horizon increased (M_{F1} = 6.72 *vs* M_{F2} = 7.02 *vs* M_{F3} = 7.18). There was no main effect of trend. But, there was a significant horizon x trend gradient interaction (F (4, 192) = 3.00, p = .002), with absolute errors increasing faster for the shallow trends (Figure 6.16).

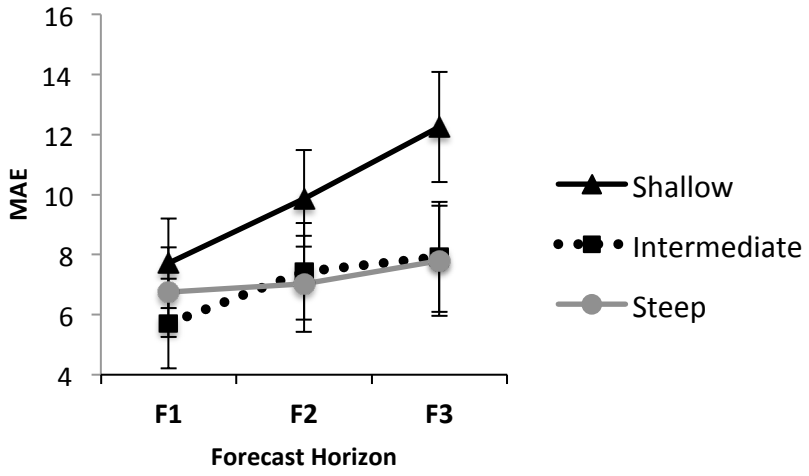

Figure 6.16 Graphs of mean values of absolute errors for all trend conditions. MAE is larger and increases faster for the shallow trends. Intermediate and steep trends outperform shallow trends in terms of accuracy.

Discussion

In this experiment I confirmed that best performance is found for intermediate trend gradients. Across the three trend gradient conditions, participants were found to exhibit a significantly different behavior in terms of signed error for shallow trends: their errors for all horizons were larger and increased with horizon rapidly while the same was not true for the other conditions. This strongly suggests that greater space availability impairs forecasting performance. Perhaps lack of available space excludes the response options that otherwise characterize forecasting biases.

Overall signed error for all conditions was higher as horizon increased. This finding shows participants' propensity to anti-damp trends as forecast horizon increases, especially for shallow and intermediate trends. For steep trends no effects were found. Perhaps the limited space available to make forecasts prevented the usual processes that produce damping from occurring.

The previous experiments employed noise free series. Environmental time-series, however, are noisy series. As discussed in Chapter 1, noisy series impair participants' judgmental performance when series are presented in graphical format (Harvey, 1995). Will the same happen in experiential tasks? I will now turn to examining participants' performance with noisy trends.

6.4 Experiential forecasting from noisy trends (Experimental Study 10)

Harvey (1995), in a forecasting task with graphs, found damping to be greater for steeper gradients, in an experiment where forecasts were made from near-linear segments of high frequency cyclical series. Will the same happen in experiential tasks?

H_4: Noise introduction will impair performance in trended series; the higher the noise the less the accuracy in participants' forecasts. Damping effects are expected to be more pronounced than those found in noise free series and to be greater with higher noise levels.

6.4.1 Method

Participants

A total of 114 participants, 53 male, 61 female (age M = 28.94, SD = 6.04), were recruited for the experiment (38 in each of the three conditions) from online sources and were paid 0.5$ each.

Design

The study employed a 3 noise levels x 3 time-periods (forecast 21, 22 and 23) between-subjects design. Speed was set to 1 sec between successive stimuli, as before.

Stimulus materials

There were three types of trended linear series, which were constructed by using the equation: $X_t = 4.5t + \varepsilon_t$, where ε_t was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2. The series time-step was the same for all three series and was selected according to the findings of the previous experiment because it eliminated misplacement effects for the first forecast. Thus, the last data point of each series was near the vertical mid-point of the screen. The noise term, ε, changed according to the noise condition the participants belonged to. For low noise conditions, ε had a mean of zero and a variance of 9, for medium noise conditions ε had a mean of zero and a variance of 20.25 and for high noise conditions ε had a mean of zero and a variance of 36 (for a graphical illustration of differences between low and high noise conditions, see Figure 6.17). High noise series display larger differences between successive data points. Medium noise series successive changes are on average lower than the high noise series but higher than the low noise series. Figure 6.17 comprises only high and low noise series for clearer illustration.

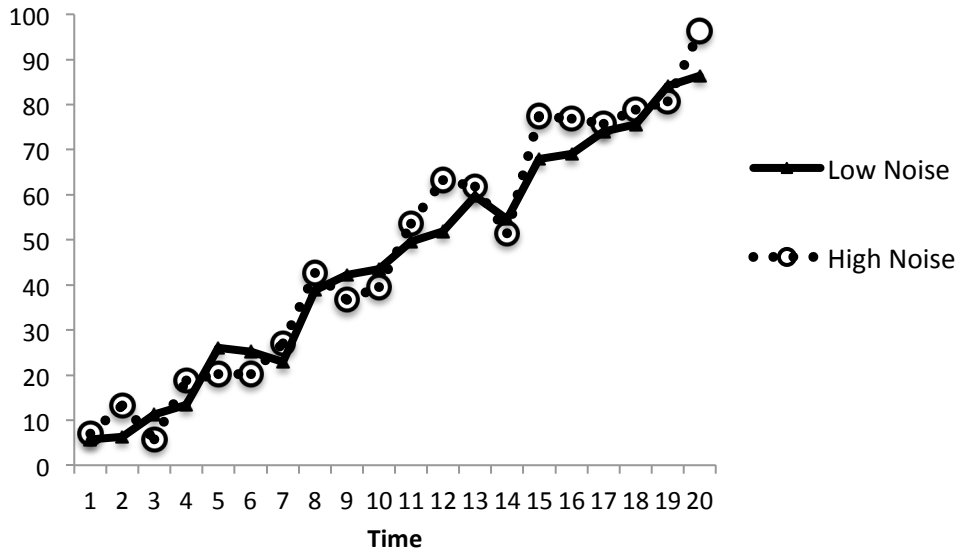

Figure 6.17 Graphical representation of high and low noise series in Experimental Study 10.

Series were presented to participants, as in the previous experiments.

Procedure

Procedure was the same as before: participants saw 20 data points and were requested to produce 3 forecasts.

6.4.2 Results

Here, I calculated the implied slope by entering the three horizon forecasts provided by participants to a regression model. I then compared those values with the actual slope of the series (i.e. 4.5). According to the methodological analysis provided in Chapter 2, this is considered a good measure to deal with trend-damping biases in noisy series. This measure assimilates elevation effects

as well (e.g. see Harvey and Reimers, 2013). Table 6.5 and Figure 6.18 summarize those findings.

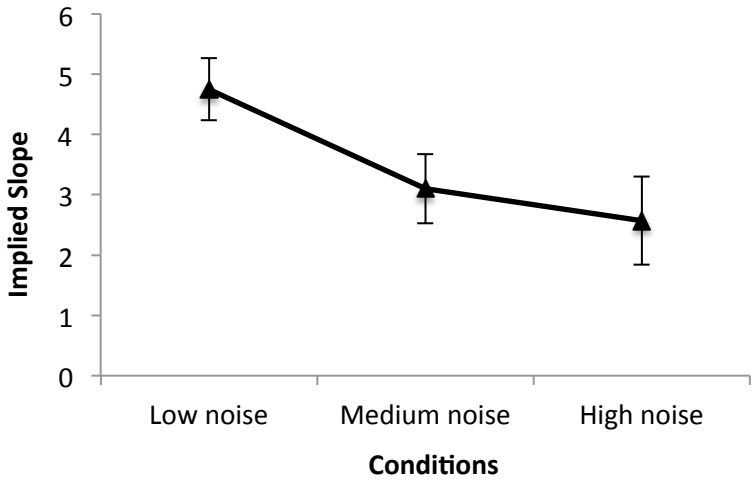

Figure 6.18 Implied slope for the three noise conditions together with standard errors. The higher the noise, the lower the average implied slope.

Table 6-5 Implied slope and associated standard errors for each noise condition

	Implied slope
Low noise	4.75
	(0.52)
Medium noise	3.10
	(0.58)
High noise	2.57
	(0.73)

To see whether differences between implied slopes were significantly different from each other I run a one-way ANOVA for the 3 noise levels. The dependent variable was implied slope. Overall, main effects of noise level were found (*F*

(2, 111) = 3.43, $p < .05$), indicating that overall implied slope decreased as noise increased. Significant differences were found between low and medium and low and high noise conditions.

Discussion

This experiment indicated that noise impairs forecasting performance by increasing damping effects. This is in accordance with hypothesis 4. With these findings, results from the previous experiments that failed to show damping seem to make sense. It is clear that damping appears only when there is significant noise in the signal. However, the fact that noise introduction impairs subjects' performance, is not new; it is in line with forecasting experiments with graphical representation of the series; the higher the noise, the more participants damp trended series. Harvey and Reimers (2013) have shown exactly that; in their experiments, damping and antidamping increased with noise in graphically presented series. Interestingly, they posed a question for future directions in this line of research: "would the effects have been the same if the temporal patterns providing the context had been presented as tables of numbers or as sequences of events experienced in real time?" Judgmental forecasting research using tabular formats confirms the notion that damping and antidaping increases with noise (see for example, Keren 1983). Also, Experimental Study 10 of the current thesis presents evidence that this might be the case for real-time experiential settings as well.

A dominant explanation, for this phenomena which appears to be true for those different settings is that of adaptation and ecological knowledge, which posits that humans have adapted to the environment, in which natural trends tend to be damped. Harvey & Reimers (2013), in a large scale online forecasting experiment from graphs tested 1020 participants on a single shot experiment. Trend-damping and anti-damping effects were obtained, even though participants were completing only a single trial. This provided evidence that these phenomena could not be attributed to experimental artefacts. They therefore proposed that damping and anti-damping arise from long-term adaptation to the natural environment. It is true that in our environment, growth tends to accelerate positively because resources are sufficiently available to allow it to continue. However, this growth becomes unsustainable when the resources for it are no longer available. At this point, the original pattern of growth becomes damped, and the series that initially showed positive acceleration becomes sigmoidal. This sigmoidal growth has been shown to be characteristic of many time series in the environment (see Tsoularis and Wallace, 2002). In these cases, growth curves appear to be typically sigmoidal.

As mentioned above, Keren (1983) also provided indirect evidence that adaptation to the environment seemed to be the cause of trend-damping, even when the data was presented in a tabular format. He asked Canadian and Israeli participants to forecast food prices based on data from the previous four years. Both of the groups were prone to trend damping, however, Israeli participants damped less. This effect was initially attributed to the fact that Israeli

participants were using higher numerical values, but when Israeli participants made forecasts from prices in post-1980 Israeli Shekels rather than pre-1980 Israeli Pounds (worth one tenth of an Israeli Shekel), the results were the same. Keren (1983) proposed the effect was due to the experience of higher food prices by Israelis. There were also other studies which have examined forecasting in a tabular presentation mode (Harvey & Bolger, 1996). When participants viewed tables of numbers from which they have to produce forecasts, they were still prone to damping. This indirect evidence that ecological knowledge is not format dependent along with the present results, suggests damping and anti-damping of noisy series appears regardless of the presentation mode.

6.5 Experiential forecasting from untrended noisy series (Experimental Study 11)

In the previous experiment, I showed that damping biases are more pronounced with higher levels of noise in the series. This was expected based on findings from classical judgmental forecasting literature (e.g Harvey, 1995). It should be interesting to examine whether noise effects are also present when untrended series are presented to the forecaster. Findings from forecasting tasks from graphs suggest that the forecaster introduces noise in an attempt to represent the given data series (e.g. Harvey, 1995). If this is the case in experiential settings as well, then noise introduction effects can also be generalised as biases that appear regardless of the presentation mode. Therefore, I test the following hypothesis:

H_5: Noisier untrended series will produce noisier forecasts

6.5.1 Method

Participants

A total of 73 participants with a mean age of 29 years, (SD = 8.1), took part in the experiment (there were 37 in a low noise level condition and 36 in a high noise level condition). Forty-three were male and thirty were female. Participants were recruited from online sources and were paid 0.5$ each.

Design

The study employed a 2 noise level (low, high) x 3 time-periods (forecast 21, 22 and 23) between-subjects design. Speed was set to 1 sec between successive stimuli, as before.

materials

Untrended linear series, were constructed by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon$, where X_{t-1} was the previous observation, μ was the mean of the series, α was the degree of autocorrelation ($\alpha = 0.5$ for both conditions), and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2 ($\sigma^2 = 14$ for condition 1 and $\sigma^2 = 225$ for condition 2). The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen ($\mu = 75$).

Series were presented to participants, as before.

Procedure

Procedure was the same as before

6.5.2 Results

To test whether participants introduced more noise in the high noise condition, I followed Harvey's (1995) methodology, which is explained in depth in Chapter 2; I fitted linear regression lines to the forecasts and compared residuals in each

condition. Those were significantly different (F (1, 217) = 19.18, $p < .001$), with residuals for the low noise condition ($M_{lownoise}$ = 6.55) being significantly lower than residuals in the high noise condition ($M_{highnoise}$ = 12.68). Thus, subjects introduced more noise when presented with the noisier series. These results confirm the noise introduction bias in experiential tasks.

I also calculated the mean absolute distance of each point from its proceeding one to measure the degree of association between those points. This can be perceived as a combined measure signalling noise introduction and implied autocorrelation from the participant. It also corresponds to the degree of anchoring to the preceding points. The mean absolute distance between the first point and the last given data point of the series (i.e. MAD_1) is calculated by subtracting the first point forecast from the last data point and taking its absolute value. This is done similarly for the second horizon point forecast by subtracting the first from the second forecast and taking its absolute value (i.e. MAD_2). The same is done for the third forecast (i.e. MAD_3). As mentioned before, this can be perceived as a measure indicative of the noise introduced but also of the autocorrelation implied by the forecaster. The implied autocorrelation measure introduced by Reimers and Harvey (2011) is not used here because the amount of data collected does not allow that. Graphs of MAD for the two noise conditions can be seen in Figure 6.19.

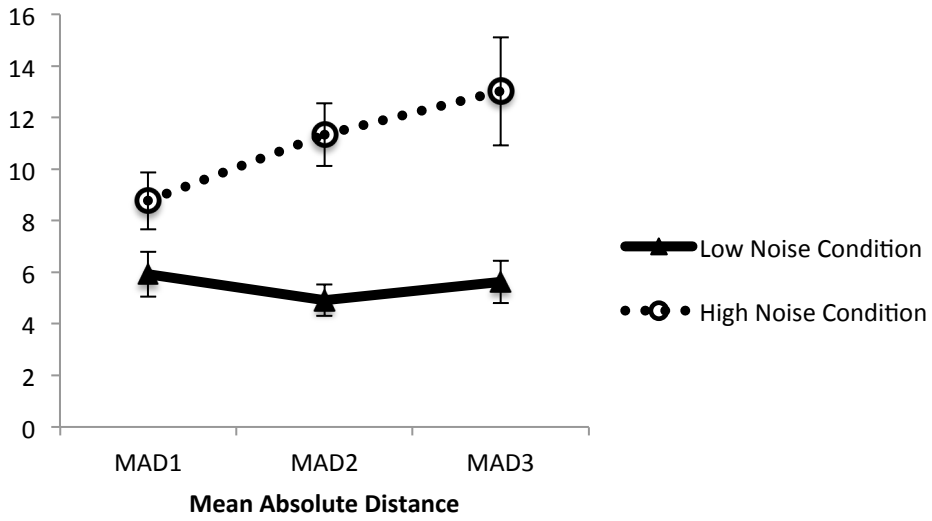

Figure 6.19 Graphs of Mean Absolute Distance (MAD) between successive forecasts for low noise (bar charts in black) and high noise (bar charts in grey) conditions. Mean absolute distances for the high noise condition are significantly higher than those for low noise condition.

To see whether differences between mean absolute differences were significantly different from each other, MAD for high and low noise conditions were entered into a one-way repeated-measures ANOVA. Overall, differences reached significance. The between-subjects factor of noise was found to have a significant main effect on mean absolute distance (F (1, 71) = 28.72, $p < .001$). High noise condition absolute differences were significantly higher than low noise ones. Also, a significant interaction between condition and horizon was found to be significant (F (1.86, 132.13) = 1.80, $p < .05$), showing a faster increase of MAD with horizon in the high noise condition.

To obtain a measure of the magnitude of these differences in comparison with the actual series, I simulated 20000 actual values of both the low and the high noise series and calculated the average absolute differences between successive points and their standard deviations. This way, a benchmark was generated for comparative purposes. Thus, for the low noise series, MAD between successive points had a mean of 3.50 and a variance of 7, while, for the high noise series, the mean absolute differences had a mean of 14 and a variance of 110. By comparing the MAD of each of the forecasts with the MAD from the series using one-sample t-tests, I found that all differences were significant except for MAD_3 in the high noise series. This means that MAD between the second and the third forecast in high noise conditions, was not significantly different from the MAD in the series. In all other cases, differences were significant (Low noise MAD_1, t (36) = 2.79, $p < .05$; Low noise MAD_2, t (36) = 2.31, $p < .05$; Low noise MAD_3, t (36) = 2.59, $p < .05$; High noise MAD_1, t (35) = -4.91, $p < .05$; High noise MAD_2, t (35) = -2.24, $p < .05$). Hence, low noise series' forecasts were always significantly higher than the series' average MAD; the opposite was true for high noise series: forecast MADs were always significantly lower than the series' average MAD, except for the case of MAD_3.

Discussion

These results confirm the noise introduction bias in experiential tasks. This bias was first highlighted as a robust one in graphical settings as well (e.g. Harvey, 1995). It is attributed to the effort of the forecaster to represent the environment (i.e. the given series here) in the best way possible. Thus, the forecaster

introduces more noise in noisier series. The same behaviour seems to be true in experiential settings as well. The forecaster introduces noise when a series is noisy. The noisier the data presented to the forecaster, the noisier the forecasts produced in an attempt to represent the given series. The analysis also shows that mean absolute differences between successive points, which are indicative of the implied noise and autocorrelation in the series, is higher than the series' MAD for low noise series and lower for high noise series. Taking into account that series in both conditions had the same levels of autocorrelation, means that this effect must be strongly associated with noise introduction. Nevertheless, autocorrelation in the series was found to influence robustly the forecaster behaviour. Thus, it would worth investigating the same phenomena with series having different autocorrelation values but same noise levels. This is the scope of the next experiment.

6.6 Experiential forecasting from series with different autocorrelations (Experimental Study 12)

The robust biases of damping and noise introduction, which are exhibited in judgmental forecasting studies from graphs, are now confirmed for judgmental forecasting tasks from experience (Experimental Studies 10 and 11). The last robust finding, which will be examined here in experiential settings, is sensitivity to autocorrelation (e.g. Reimers and Harvey, 2011). This will be achieved by presenting participants with series having various autocorrelation levels. Taking into account the findings from the previous experiments regarding the magnitude of noise introduction effects, and keeping the noise levels constant, I should be able to determine the effect of autocorrelation levels on forecasting behaviour. The hypothesis (H_6) is that forecasting behaviour will vary with autocorrelation levels; the higher the autocorrelation in the series, the closer the distances between successive forecasts. This is in line with the work by Reimers and Harvey (2013).

6.6.1 Method

Participants

A total of 75 participants, Age (M = 30, SD = 8.5), 28 male, 47 female took part in the experiment. Participants were recruited from online sources and were paid 0.5$ each.

Design

The study employed a 3 autocorrelation levels (0, 0.4, 0.8) x 3 time-periods (forecast 51, 52 and 53) within-subjects design. Autocorrelation levels were selected to match exactly those used in Reimers & Harvey (2011) in order to obtain comparable results. The series' length was now set to 50 points in order to make sure that participants experienced the different series for a long duration, enough to perceive the autocorrelation level of the series. The time interval between successive stimuli was set to 0.7 sec to ensure that participants maintained their interest to the task (it might have been boring to the forecaster to observe 50 data points in low speed). The design was chosen to be within participants, in order to increase the statistical power.

Stimulus materials

Untrended linear series, were constructed by inserting appropriate parameters into the following generating equation: $X_t = \alpha X_{t-1} + (1 - \alpha) \mu + \varepsilon_t$, where X_{t-1} was the previous observation, μ was the mean of the series, α was the degree of autocorrelation ($\alpha = 0$ for condition 1, $\alpha = 0.4$ for condition 2 and $\alpha = 0.8$ for condition 3), and ε was noise produced by randomly drawing values from a Gaussian distribution with a mean of zero and a variance of σ^2 ($\sigma^2 = 324$ for all conditions). The mean value, μ, was selected to ensure that the final data point was close to the vertical mid-point of the screen ($\mu = 75$).

Series were presented to participants, as before.

Procedure

Procedure was the same as before

6.6.2 Results

To test the autocorrelation illusion bias, I implemented the methodology used in the previous experiment. Thus, mean absolute differences between successive points were calculated again. Graphs of MAD for each condition are shown in Figure 6.20.

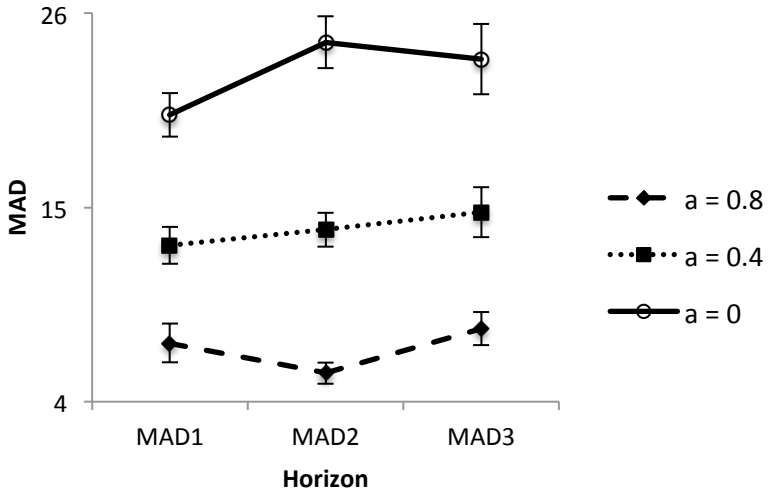

Figure 6.20 Graph of mean absolute distances for all correlations and horizons

To see whether differences between mean absolute differences were significantly different from each other, MADs for high, medium and low autocorrelation conditions were entered into a one-way repeated-measures ANOVA. Overall, differences reached significance for autocorrelation level (*F*

(1, 222) = 69.97, $p < .001$). The high autocorrelation condition contained absolute differences that were significantly lower than medium and low autocorrelation conditions. Linear contrasts were significant (F (1, 222) = 3.95, $p = .048$). Post-hoc tests showed that for all pairs of MAD scores between the different conditions, differences were significant.

To obtain a measure of the magnitude of these differences in comparison with the actual series, I simulated again 20000 values of low medium and high autocorrelation series and calculated the average absolute differences between successive points and their standard deviations. This way, a benchmark was generated as before. Thus, for the low autocorrelation series, MAD between successive points had a mean of 20 and a variance of 247, for the medium autocorrelation series MAD between successive points had a mean of 17 and a variance of 173, while, for the high autocorrelation series, the mean absolute differences had a mean of 15 and a variance of 132. By comparing MADs of each of the forecasts with MADs from the series using one-sample t-tests, I found that all differences for the high autocorrelation series were significant (MAD_1, t (74) = -14.11, $p < .001$; MAD_2, t (74) = -15.41, $p < .001$; MAD_3, t (74) = -7.34, $p < .001$). This means that participants anchored more in the high autocorrelation condition, implying higher autocorrelation in the series (i.e. autocorrelation overestimation). This finding is somewhat different than that found in Reimers and Harvey (2011). There, participants slightly underestimated the autocorrelation of 0.8. For the medium autocorrelation series, MADs of the forecasts were significantly lower than the MADs in the

series for MAD_1 and MAD_2 (MAD_1, t (74) = -3.98, $p < .001$; MAD_2, t (74) = -3.39, $p < .001$). MAD_3 was not significantly different than the series MAD. This means that participants anchored more than required in the medium autocorrelation condition. This finding is in line with Reimers and Harvey (2011) findings, where subjects overestimated the autocorrelation of 0.4. Finally, for the low autocorrelation series, MADs in the forecasts were not significantly different from the MADs in the series, except for the case of MAD_2 (MAD_2, t (74) = 2.96, $p < .05$). This means that participants anchored equally or less than the series' MAD. This finding is not in line with Reimers and Harvey (2011) findings, where subjects overestimated the autocorrelation of 0. Here, they underestimated the autocorrelation in the second time step, implying negative autocorrelation. This might be due to the fact that in this experiment, the noise introduced was higher than that employed in Reimers and Harvey's (2011) experiments.

Discussion

In this experiment, the autocorrelation illusion was partially validated for forecasting tasks from experience. Three series of different autocorrelation levels (same autocorrelation conditions as the ones used in Reimers and Harvey, 2011) but same noise levels were used to avoid effects such as those shown in the previous experiment. Significant differences were found in the way the forecaster anchored in those three different autocorrelation conditions. For high autocorrelation series, participants anchored significantly more in their three successive forecasts than in the other conditions, suggesting that they took into

account the autocorrelation in the series; for low autocorrelation series, they anchored significantly less than the other three conditions, suggesting that they adjusted away from the preceding points to accommodate for the low autocorrelation in the series. Finally, for medium autocorrelations, MADs where found to be between those two extreme cases. The MAD benchmark suggests that they overestimated autocorrelation for high and medium autocorrelations. For low autocorrelation conditions, they appeared to imply equal autocorrelation with that of the series for time steps 1 and 3 and a slight underestimation for time step 2.

This finding in conjunction with that found in static judgmental forecasting settings (e.g. Reimers and Harvey, 2011; Eggleton, 1982, Bolger and Harvey 1993), confirms the sensitivity of naïve forecasters to the degree of autocorrelation in the series.

Using the MAD as a benchmark I found that subjects anchored more than they had to in the medium and high autocorrelation series but not in the low autocorrelation series. This finding suggests that in experiential settings there might be an "either or" strategy where participants either decide to anchor conservatively to the series or not to anchor at all. Also, the high noise level chosen for this experiment might have influenced the anchoring process. Although this experiment confirmed participant's sensitivity to the autocorrelation in the series, more research with various noise levels is required to confirm exactly the forecasting behavior and specifically to identify when overestimation or underestimation of the autocorrelation in the series appears.

Larger samples would provide a better opportunity to employ the autocorrelation measure introduced by Reimers and Harvey (2013)

6.7 Summary and General Discussion

In the current chapter, I examined the theme of forecasting from time-series that were experienced by the participants in real-time. A sequential bar-charts experiential paradigm was created and used for the purposes of this research, inspired by research in the area of mental representations. Overall, bar chart formats serve as way to investigate forecasting from real-time experience. Significant biases were found to operate in this setting, similar to those found in judgmental forecasting from graphs, but sometimes more extreme than those found when the experimenter uses graphs. For example, effects related to the screen margins appeared to be higher than those found in other settings (e.g. Lawrence and Makridakis, 1989). These effects were scrutinised in Experimental Studies 7 and 8 of this Chapter, where subjects were found to significantly anti-damp shallow trends from both upward and downward series, with the effect being more pronounced for downward trends in accordance to relevant research with graphical representations of the series (e.g. Harvey and Reimers, 2013). The greater the space on the graph above the plot of a linearly trended time series was, the higher the forecast tended to be. So, boundary biases were responsible for this effect. This is an important finding for forecasting and it should be further investigated in graphical settings as well.

In Experimental studies 7 and 8, I also tested the effect of the speed of the display between successive data points; this showed that those in the fast display conditions produced larger errors compared to those in slow conditions. Faster speeds impaired performance in forecasting tasks from experience more than slow speeds. This finding can be attributed to the fact that participants experiencing the data slowly had more time to identify the nature of the data generation process, and specifically the size of the time step of the noise free trended series. Hence, their judgement extrapolations were more representative of their experiences (e.g. Kiani et al., 2008).

This result should be further investigated more with more complex series. Here, I used series where patterns were easily perceived by the forecaster. It might be the case that significant memory effects are at play in more complex settings. These memory-related effects might change the direction of the speed display effect I obtained here. For example, if a noisy seasonal series is presented to participants in a slow display, they might not be able to capture the underlying seasonal signal, whereas a fast display might enhance the mental representation of the signal in the series.

Boundary effects, such as those found in Studies 7 and 8, were further scrutinised in Study 9, where I used as stimuli materials trends of different gradients. It was confirmed that optimal performance for the first horizon, where forecasts are most often misplaced, was achieved with intermediate trends. Final data points for series with intermediate trends were close to the mean of the vertical axis. Accuracy was greater in these conditions.

Experimental studies 7, 8 and 9 employed noise free series. These were useful in order to investigate systematic errors uncontaminated by additional effects produced by noise. Nevertheless, investigations with noisy series are of special interest to forecasting research. Hence, noisy trends of intermediate gradient were used in Experimental Study 10. The design utilized knowledge from the previous experimental findings. Thus, different noise terms were imposed on intermediate trends. Results showed that participants introduce noise into their forecasts; the noisier the series, the noisier the forecast sequence. This confirmed the noise introduction bias found in judgmental forecasting settings with graphs (e.g. Harvey 1995).

Trend damping was found using the implied slope measure introduced in Experimental Study 10. The higher the noise, the more pronounced the trend-damping in the series. This finding is in accordance to research in trend damping (e.g. Harvey and Reimers, 2013). The next experiment (Experimental Study 11) confirmed noise introduction effects using untrended series One interpretation is that participants are prone to noise introduction because of their attempts to represent the data series in their forecast sequence.

Finally, participants were found in Experimental study 12 to be sensitive to the levels of autocorrelation in the series. A within-participants experiment was designed with series of different autocorrelation levels but the same noise. In high autocorrelation series, participants anchored significantly more than in medium autocorrelation series and low autocorrelation series. Thus, these findings suggest that the forecaster is prone to similar (though not identical)

biases to those found in static settings where graphs are used as stimulus material.

Participants showed trend damping for noisy series, noise introduction for series with noise and sensitivity to the level of autocorrelation in the series. Only boundary effects were found to be more pronounced than those found in static settings. Thus, these findings partially confirm the accounts of adaptation to the environment (e.g. trend damping and sensitivity to autocorrelation) and the account of representativeness in forecasting processes (e.g. noise introduction).

The experiential experiments of the current Chapter were aimed at investigating robust biases of the classical judgmental forecasting literature (e.g. Harvey, 1995; Reimers and Harvey, 2011; Harvey and Reimers 2013) using a new type of display. Instead of viewing all points simultaneously, participants experienced data points individually. Results have shown similar patterns to those from the judgmental forecasting literature.

Chapter 7 Summary and Conclusions

The aim of this thesis was to examine understudied areas of judgmental forecasting research. The themes that I pursued, were mainly associated with the way presentation format can improve forecasting processes with the input of judgment. In a series of experimental studies, by controlling the presentation of forecasting information, I obtained order, end-anchoring, length and dynamic presentation effects, while scale manipulation confirmed previous findings of scale invariance in graph perception. The underlying mechanisms were explored by closely examining the context sensitive, anchoring and adjustment heuristic. Forecasting accuracy was improved in a number of ways by introducing longer lengths to the forecaster, by reversing the order of the forecasting task and introducing end-anchors to it, and by presenting time series dynamically to the forecaster. That way, I demonstrated that forecasting can be improved in simple ways, which can be easily introduced in practice. In this Chapter, I will summarize the main findings of each of the previous chapters, discuss the implications of these findings in practice, and consider future directions.

.

7.1 Summary of findings

In Chapter 3, I examined order effects in judgmental forecasts for multiple time periods. Here, I introduced the notion of end-anchoring, where the forecaster makes his prediction first for the most distant horizon and then for the proximal ones. I constructed a forecasting task where participants provided their forecasts for various series types with or without the use of an end-anchor, in normal or reverse direction (*Experimental Study 1*). Results showed that end-anchoring can be a useful tool. It primes the forecaster to take into account global patterns of the series in a more deliberate way than traditional heuristic-driven forecasting strategies, which are strongly influenced by noise in the series. The use of an end-anchor increases forecasting accuracy for the most distant horizon but also enhances forecasting for the rest of the forecast sequence. Further evidence for the usefulness of the end-anchoring strategy in higher noise environments was provided in *Experimental Study 2*. Forecast direction was also investigated in this set of experiments. Results show that forecast direction has an effect on accuracy only when the ideal sequence of forecasts is strongly nonlinear.

In Chapter 4, I examined the effect of series' length on forecasting accuracy and on underlying cognitive mechanisms. By manipulating the series' length, I distinguished the series lengths that enhance accuracy from lengths that severely impair it for various types of time series (*Experimental Study 3*). Results showed that forecast error describing an inverted U-shaped function:

mean average errors are lower for longer series and increase as series length decreases, taking a maximum value for a few data points series (two to five, depending on the series' type). Then absolute error decreases again for very short lengths. In terms of the underlying cognitive mechanisms, it appeared that pattern-based heuristics that are effective for long lengths continue to be used for forecasting shorter series (e.g., five items) where they are less appropriate without modification. As a result, accuracy is lower than it would have if the forecasts had been produced by the naïve forecast. This is because the pattern parameters used in those heuristics tend to be inappropriate when series are short. In these circumstances, performance can be improved by reducing series length still further and thereby forcing forecasters to use the naïve forecast. Length effects were further explored for more distant forecasting horizons (*Experimental Study 4*) and similar results were obtained.

In Chapter 5, another understudied area of judgmental forecasting was scrutinized: scale effects. In a study by Lawrence and O'Connor (1992), no scale effects were obtained when using ARMA series. The same was generally true in two experimental studies (*Experimental Study 5* and *6*) where autoregressive series with uniform and Gaussian noise were employed. Different time series (autoregressive and independent ones) were used to examine the generalizability of previous findings. In *Experimental Study 5*, I investigated scale effects by using uniform noise in the series' generation algorithms. The findings yielded only weak evidence that scale effects depended on degree of autocorrelation. These results were further examined in

Experimental Study 6 using a Gaussian noise term this time. In this study, presentation scale did not affect accuracy. Noise type affected accuracy significantly but did not interact with scale effects.

Finally, in Chapter 6, a novel, forecasting paradigm, the experiential forecasting task, was designed with the aim to present stimuli in a dynamic fashion to the forecaster with the use of sequential bar charts. Questions in this set of experimental studies concerned robust phenomena of judgmental forecasting. Thus, trend damping, noise introduction and the autocorrelation illusion were studied in a set of six experimental studies (*Experimental Studies 7-12*). Experiential judgments showed common features with those found in descriptive formats. Trend damping, noise introduction and sensitivity to autocorrelation were confirmed for dynamic settings as well, rendering the adaptation accounts that explain these behaviours more plausible.

7.2 Implications and Limitations

Forecasting plays an essential role in business planning and in many other areas of life, as discussed in Chapter 1. Although development of formal methods of forecasting continues apace, many surveys have shown that most forecasting within businesses is based on judgment (e.g., Sanders and Manrodt, 2003). Adoption of formal techniques appears to have reached an asymptote (Lawrence, 2000). Given the importance of forecasting within business and other areas and the fact that much of it continues to be largely based on

judgment, research into judgmental forecasting has real potential for increasing business effectiveness. There is now a large corpus of such research (Lawrence et al., 2006) and findings have been used to develop principles of good practice (Armstrong, 2001; Armstrong et al., 2013).

The outcomes of this thesis provide promising suggestions for how the forecaster can improve forecasting performance, especially in cases where the presentation format of the forecasting task plays an important role. Additionally, this research can be used in conjunction with findings from the cognitive science field to provide the foundations of future anticipation processes.

Implications for practice when forecasting for multiple time periods ahead

A wide variety of techniques have been developed for improving judgmental forecasts. They include feedback-based training (Goodwin and Fildes, 1999; Benson and Önkal, 1992), decomposition (Edmundson, 1990), combining forecasts from a number of forecasters (Clemen, 1989), and use of advisors (Lim and O'Connor, 1995). However, all of these approaches require quite heavy investments of time, money, or effort. In the present thesis, I have shown that significant gains in forecast accuracy can be achieved simply by changing the order in which forecasts are made. In particular, requiring the forecast for the most distant horizon to be made first is an effective way of increasing the accuracy of forecasts, especially those for more distant horizons. This research can be applied in a variety of applied and academic settings where distant horizon forecasting is of special interest. For example, for managerial but also

entrepreneurial purposes of strategic planning (Fildes et al., 2009; Goodwin et al., 2010; Syntetos et al. 2009; Armstrong et al., 2013), where distant horizon forecasts are of paramount importance, this approach offers a new way of thinking, prioritizing and organising an efficient strategic plan. A requirement for an initial distal forecast can be also directly introduced in techniques such as Delphi and group-forecasting (Rowe and Wright, 1999; Önkal, Lawrence and Sayım, 2011). Optimism research is another area where this finding can have important implications (Harris & Hahn, 2011); the experimental paradigms used in optimism research are currently conducted without the use of forecasting knowledge although temporal effects in corresponding likelihood judgments can be attributed to the format of the experimental design.

Implications for practice on dimensional aspects of presentation format

Retaining and retrieving data for forecasting purposes is often expensive. Hence, it is reasonable to ask whether it is worth the effort. Also, when businesses change hands, historical data may be lost. Then, important questions arise concerning the way optimal forecasting can be achieved by the new owners. The present thesis produced results relevant to both these practical questions in the studies in which I varied series' length. Retaining and retrieving data for judgmental forecasting purposes can be useful: the longer the series presented to the forecaster, the more accurate judgmental forecasts or judgmental adjustments are likely to be. Findings of the present thesis show that forecasting from series of intermediate length can impair judgmental forecasting accuracy. Thus, it is worth making an effort to increase series'

length to improve accuracy. (Of course, the cost of the effort must be weighed against the benefits accruing from the gain in accuracy). If no more than five items are available and logistics, costs or data unavailability prevent series length from being increased, then shortening the series to, say, one item will improve accuracy of judgmental forecasts for most series types and not impair it for others.

A variety of length and scale formats are employed to present time series of interest in trading, managerial and other settings in the financial sector. Computer screens and monitors, as well as palm-tops and mobile devices, serve as a way to obtain this information. Then a judgmental forecast and a decision can be made regarding subsequent investments. This thesis provides evidence that while time series lengths should be long enough for the forecaster to achieve optimal performance, the scale in which time series graphs are presented is relatively independent of the forecasting accuracy. Thus, there is little difference if one consults big or small screens to reach a final decision. Nevertheless, when prediction intervals are also crucial determinants of a final decision, as in the case of weather forecasting, then larger scales should be favoured (Lawrence and O'Connor, 1993).

Implications for practice when forecasting from real-time experience

Judgmental forecasting from experience, or experiential forecasting can be useful in a number of ways. First, it can be used by cognitive scientists in the study of sequential mechanisms of evidence integration in perceptual and preferential choice (Tsetsos et al, 2012) and in studies of sequential learning

(Gureckis and Love 2010). In these areas, there are no experimental paradigms, which use spatial representations of sequential stimuli.

The real-time experiential forecasting paradigm can also have implications in the financial world. Traders typically experience series in real-time and have to forecast their future prices. Their tasks are different in many ways: they have larger working memory loads, they experience distractions and can often see multiple data points at once. It is possible that studies using the experiential paradigm could be used to produce findings that enhance forecasting performance by fine-tuning timing and boundary variables. This technique also has potential for investigating differences in decision speed and accuracy between experts and novices, for which much conflicting evidence has been found (see for example Muradoglu and Önkal, 1994; Thomson, Pollock, Henriksen & Macaulay, 2004; Thomson, Önkal, Avcioglu and Goodwin, 2004; Önkal, Yates, Şımga-Muğan and Öztin, 2003). If experts are faster at extracting critical information for series than novices, their accuracy will reach an optimal performance earlier.

Limitations

Although these implications are important in both applied and theoretical settings, it must be acknowledged that the experiments reported here have certain limitations, which would benefit from future research and use of modern technology. The set of experiments reported in this thesis serve mainly as evidence of underlying biases that might operate in settings where future judgments are to be made. Improvements are suggested under the minimal

experimental context used here. Nevertheless, this context might be different for real-world tasks, where much forecasting is carried out in environments where there is a wealth of domain information. Managers, for example, take into account a variety of domain information pieces when making forecasts; future macroscopic trends of the market, future promotions and competitors' strategies are just a few of them. This suggests that perceptual and motivational biases might operate in these cases creating distortions (see for example Goodwin, 2005; Goodwin and Fildes, 1999). Undoubtedly, one of the major obstacles to accurate predictions of future outcomes is the way in which humans introduce distortions. One apparent pervasive example is optimism bias. The impact of this kind of distortion on forecasting was acknowledged in the British government's 2003 'Green Book' intended for HM Treasury as a guide for Central Government. The Green book identified optimism bias as one of the key factors to be mitigated. Optimism bias has also been referred to by the IMF (International Monetary Fund) as a basis for error prone predictions of National Bodies handling of the current monetary crisis in the Eurozone. Clearly, this psychological phenomenon is thought to have a severe bearing on several important real world issues, which are closely associated with accurate predictions. Such considerations were not taken into account in the current experiments, where the aim was to isolate underlying biases with minimal information other than the shape of the given series itself.

Moreover, contextual information can be provided in business settings in the form of judgmental prediction intervals or density forecasts, which were not

investigated here. Only point forecasts were assessed in this set of 12 experiments. Nevertheless, the estimation of future prediction intervals and probability distributions is important in a variety of applied fields. For example, it is consistently used by those responsible for providing insurance against hurricane damage to property. Every year, insurers look at past records of hurricane occurrences and use advice from mathematical models to make judgments about the number of hurricanes that will strike the Atlantic seaboard of the USA. Insurers use these forecasts to set insurance rates. Practitioners in this sector take into account past historical values of hurricane counts, formal model-based hurricane forecasts from official sources, such as NOAA, from catastrophe risk modelers, and from in-house modeling outputs. All model information is provided in the form of prediction intervals rather than point forecasts in this important business sector, which influences thousands of human lives, and future estimations are also sketched in a prediction interval canvas. There is a wealth of interesting findings associated with biases associated with prediction intervals estimation (see for example O'Connor and Lawrence, 1989), which suggests that estimated prediction intervals tend to be too narrow. This important issue was only loosely related with the hypothesis drawn in Chapter 5 but, nonetheless, requires further investigation in light of the present findings.

Except from the contextual considerations discussed above, the current thesis would have benefited from some methodological improvements as well. For example, experiments could have taken significant advantage from web

crowdsourcing tools with the aim to collect thousands of participants in each condition in the way Reimers and Harvey (2011) collected data. Modern technology provides nowadays various means of collecting data from large pools and this is especially important for judgmental forecasting studies, where noise is an important factor. The collection of data of this magnitude would have eliminated concerns related to the size of the sample, the type of series used, the understanding and generalizability of results and the potential influence of the sample characteristics or the individual differences associated with that (Eroglu and Croxton, 2010).

7.3 Future directions

Designing superior Forecasting Support Systems (FSS)

Effective forecasting is a vital component of commercial competitiveness. Companies that produce effective forecasts can have competent supply chains, superior product availability and lower production costs. Thus, predictions elicited via the forecasting process affect all core functional areas of a firm because forecasts are used as input to inform decisions of these functional areas. In the supply chain domain, for example, predictions are accomplished via Forecasting Support Systems (FSS), which not only provide valuable information and statistical forecasts for the next periods but also allow for judgmental input from the forecaster (Armstrong, 2001). This input mainly concerns components that the statistical model cannot take into account such as promotions. In other cases, there might be insufficient data available for the

statistical model to predict data regularities. It is crucial, thus, to continuously improve the FSS design so that integration of management judgment can be carried out in an efficient way. Much existing research on judgmental forecasting has focused on the properties of the time series data. However, potentially equally important is the way in which data are presented to the user of the FSS.

Towards that direction, a variety of findings stemming directly from this thesis can be used as potential recommendations for better integration of management judgment into Forecasting Support Systems: a first recommendation, stemming from results obtained in Chapter 4, would concern the usefulness of presenting to the forecaster historical data of adequate length ($n > 40$). In cases where this is not possible, it is advisable to present only the last value of the series, in order to avoid biases introduced by inappropriate use of pattern-based heuristic mechanisms when series of intermediate length are shown to the forecaster. Another suggestion, stemming from Chapter 6 findings (Experiment 6.1), concerns the screen margins; it is advisable to adjust the screen frames, especially when trended series are to be presented to the forecaster. Vertical screen borders might affect the judgmental process, rendering, thus, trend damping biases more pronounced if screen margins are not adjusted accordingly. Additionally, Chapter 3 findings suggest that it might be worth adding an option in forecasting systems where the forecaster is requested to produce distant horizon forecasts first and then proximal horizon ones; judgmental biases introduced via short-term noise introduction mechanisms could be reduced by adopting this approach. Finally, Chapter 6 findings suggest

that, in some cases, it might be beneficial for the forecaster to observe information in a dynamic rather than a static mode. These recommendations are valid for a variety of series and noise levels in the series, as it is evident from the data reported in this thesis.

However, there are quite a few research ideas related to the above-mentioned presentation format findings that would benefit from future research. It would worth studying, for example, whether the findings of this thesis are still valid for series where rare events or extreme irregularities occur. These have a special importance in operational settings. Moreover, and since there are no specific guidelines on FSS design, it would certainly be advisable to focus on all aspects of FSS design, including those not mentioned in this thesis. Within graphical presentation, there are many ways in which time series can be presented - line graphs, bar charts, tables or scatterplots. There is clear evidence that the choice of graphical display format can affect the way data are perceived (e.g. Harvey and Bolger, 1996). So, research on display effects in forecasting that is designed to eliminate cognitive biases and, thus, make forecasts more efficient would certainly make sense.

Judgmental Forecasting and visual perception research

Work being conducted in the field of judgmental forecasting, can be now enhanced with new techniques developed for research in related areas. Eye-tracking or mouse-tracking methodologies are two such techniques. These could be used to measure information seeking and exposure time for each of the series' elements. Eye-tracking techniques could provide a useful tool because

they capture the gaze of the forecaster thereby revealing systematic fluctuations of visual attention (see for example, Wills, Lavric, Croft and Hodgson, 2007). This way, eye-tracking studies can reveal the way cognitive resources are allocated and can locate the elements to which special attention is paid. (In these settings, gaze is usually directed towards items of interest). Eye-tracking techniques would, thus, provide evidence of whether the forecaster is focused on global or local patterns in the series, of whether his attention is focused more on the last segments of a series or rather on initial segments of it or even on rare events. Questions related to pattern-based heuristics used during the forecasting process or fast heuristics that exploit differences found in the last segments of the series should perhaps be reconciled with the use of such techniques. Thus, such evidence would be crucial in unveiling the underlying mechanisms determining the most important factors of forecasting mechanisms.

Drawing parallels between research in judgmental forecasting and the cognitive science of sequential processing in the accumulation of evidence

The experiential forecasting paradigm created in this thesis shares common characteristics with tasks employed in the cognitive science of sequential micro-processing of perceptual evidence over time (e.g. Tsetsos et al., 2012; Gureckis and Love, 2010; Summerfiled and Tsetsos, 2012). Nevertheless, investigations in these domains, as well as proposed models, remain distinct. Forecasting with the use of judgment in dynamic or static settings is concerned with how observers detect and use time series information to make predictions of future outcomes whereas research in the cognitive science of evidence integration

investigates how the observer detects, categorizes and assimilates information over time. It is possible that common mechanisms operate in these two areas. For example, in both fields, recency or anchoring mechanisms are found to operate when people are presented with series of stimuli (Epley and Gilovich, 2006; Harvey 2007; Lawrence and O'Connor, 1995; Tsetsos et al., 2012; Fiedler and Juslin, 2006). Similarly, in both domains, rare events are often underestimated (O'Connor and Lawrence, 1989; Hertwig et al., 2004; Goodwin and Wright, 2010). Thus, behavior in these two types of tasks might stem from common mechanisms of adaptation (e.g., Harvey, 2011).

Nevertheless, it is important to note here that cognitive methodologies used to examine perception and subsequent judgment of sequential stimuli in time typically involve micro-processing tasks, where subjects are presented with sequential evidence, which accumulate in a micro-time scale (e.g. seconds). In fact, studies of perception and subsequent judgment or choice in time rarely use larger temporal scales (e.g. day, months, years), as those found in judgmental forecasting from graphs. This renders regularities and biases found in judgmental forecasting from graphs even more valuable if a holistic framework of time perception and subsequent judgment is to be constructed. The few examples where larger time scales are used to unveal biases can be found in somewhat separate research fields of cognitive science. For instance, in affective forecasting, evidence suggests that people are inaccurate in predicting large time scale future outcomes and their reactions to those outcomes (e.g., Wilson & Gilbert, 2003). Inconsistencies related to large time scale future

predictions are also found in the intertemporal choice literature (e.g. Loewenstein et al., 2003). Finally, Trope and Liberman (2010) propose that large time scale judgmental problems are consistent with their Construal Level Theory (CLT), which proposes that the temporal location of a future event influences how we construe it; events that are distant are construed at a higher level, whereas those proximal to us are construed in a more thorough way. Therefore, the time scale for which people are making estimates of likelihood of future outcomes matters. This, in turn, would have implications for a unified framework of time perception and judgment. It might be possible to build parallels between findings in these separate domains with the aim to come up with complete accounts of sequential integration of stimuli in different scales in time. This would provide a common framework for understanding judgments made about future outcomes, under which an agent has to assimilate numerical information representing serial interrelated cues.

Judgmental forecasting and risky choice with the use of modern technologies

Findings from the judgmental forecasting literature are rarely used in practice in other than the financial and business settings although various domains could benefit from findings from them. Research in the domain of risky choice, is one such area. Most research into the psychology of risk has focussed on choices with immediate consequences, generally monetary gambles (e.g. Payne, 2005). However, there are various settings where decisions are made using sequential stimuli rather than discrete ones.

One such example is cumulative risk, with its pervasive implications for everyday life (Hogarth, Portell and Cuxart, 2007). People choose to expose themselves to many different risks in their everyday lives. Some have the potential for immediate catastrophic consequences, such as drink-driving and illicit drug use. For others, such as smoking and unhealthy eating, the immediate risk is vanishingly small, but the cumulative risks over years and decades can be very grave. Cumulative risk, thus, is perceived as less threatening relative to more immediate, isolated, short-term risks (Svenson, 1984). A critical component of everyday risky behaviours is myopia for distant consequences which may appear intangible. One factor might be noise in such series. This suggests that cumulative risks are particularly susceptible to underestimation (as is the continuation of a trend in judgmental forecasting (Harvey and Reimers, 2013)). Nowadays, applications for smartphones are dedicated at helping people to monitor and control these cumulative risks, such as smoking, drinking and overeating (Froehlich, Chen, Consolvo, Harrison and Landay, 2007). Within these applications, people interact on their mobile phones on a daily basis, entering various attributes of their current states (see myCompass, www.mycompass.org.au, for self-report criteria & standards). Data are collected and retained according to protocols such as the Experience Sampling Method (see for example Hogarth et al., 2007). Historical data in the form of series are presented to the users (for example, weight loss or amounts of cigarettes used within a day) from which the users can extrapolate to form their anticipations for the future. Judgmental forecasting findings could be used in practice in such settings in an attempt to design interventions that mitigate risky

behaviours. For example, a simple model could provide corrections to the user forecasts (i.e. damping of a trend), enhancing, thus, accuracy of self-predictions. The end-anchoring option could be suggested to a user as the optimal strategy to extrapolate to the future. Such an option to the menu of an application could prevent the user from common damping extrapolation biases, as discussed in Chapter 3. This is one example where judgmental forecasting findings can be applied to the risky choice domain by using new technologies. Also, data collected from cumulative risk studies can be used to build a forecasting model to determine the factors that predict lapses in people's control, with the use of hierarchical linear models (Bryk & Raudenbush, 2002) as well as fractal algorithms, which detect persistency in time-series (Koutsoyiannis, 2002). It may be the case that morning tiredness, for example, is predictive of smoking or overeating risk later in the day, or that boring activities are more likely to encourage lapses than interesting activities. An option of the corresponding application highlighting such links with the associated trend of the variable of interest (e.g. smoking or eating levels) could as well enhance the judgment of a person interested in eliminating influences of cumulative risks in their lives.

7.4 Conclusion

In this thesis, in a series of experimental studies, I investigated how judgmental input influences forecasting decisions when the presentation format of the task is manipulated. Forecasting decisions with the aid of human judgment are of paramount importance in a number of areas such as Finance, Supply Chain

Management, Environmental Operations and so on. It is noteworthy that nowadays, the area of Behavioural Operations, which incorporates applications for all the aforementioned fields, is becoming a recognized domain of research (e.g Journal of Operations Management's special issue on Behavior Issues in OM, 2013). The area of Behavioural Operations aims at understanding the decision-making of managers under various settings and at using this understanding to generate interventions that would improve operations. The findings of the current thesis demonstrate that appropriate presentation of a task can enhance performance in tasks where effective forecasting is crucial for an organisation. Specifically, two factors were found to significantly improve accuracy in graphical presentations: the use of an end-anchor, the presentation of a sufficiently long series. Scale manipulations did not yield major effects, confirming findings of invariance in the graph perception literature. Furthermore, the findings revealed that judgmental forecasting from a simple dynamic paradigm, which simulates real-time experience of time series, elicits phenomena similar to those found in the classical literature of judgmental forecasting from static graphs; the forecaster was found to damp trends from noisy series, to introduction noise in the forecasts in an attempt to represent the series and to be sensitive to autocorrelation. These common characteristics underlying forecasting from graphs and from experience suggest that an integrated approach involving common cognitive mechanisms could be applied to all types of judgmental forecasting tasks. Overall, the results of this thesis shed light in understudied areas of judgmental forecasting, refining existing knowledge of the way people use time series information to predict future

outcomes and revealing possibilities for an integrated framework for judgmental forecasting research, which could be proven useful both in applied settings, such as those associated with the emerging field of Behavioural Operations, but also in more theoretical settings investigating the way the human mind produces estimations about the future when encountering streams of stimuli..

References

Allen, P. G., & Fildes R. (2001). Econometric forecasting. In: Armstrong J. S. (Eds). *Principles of Forecasting* (pp. 303–362). Norwell, MA: Kluwer Academic Publishers.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2)*,* 106-111.

Andersson, M., Gärling, T., Hedesström, M., & Biel, A. (2012). Effects on stock investments of information about short versus long price series. *Review of Behavioral Finance, 4*, 81-97.

Andersson, P., Edman, J., & Ekman, M. (2005). Predicting the World Cup 2002 in soccer: Performance and confidence of experts and non-experts. *International Journal of Forecasting, 21*, 565–576.

Anderson, N., (1981). *Foundations of Information Integration Theory*. New York: Academic Press.

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

Andreassen, P. (1990). Judgmental extrapolation and market overreaction: on the use and disuse of news. *Journal of Behavioral Decision Making, 3*, 153-174.

Armstrong, J. S. (1982). Strategies for implementing change: An experiential approach. *Group and Organization Studies, 7*, 457-475.

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

Armstrong, J. S., & Collopy. F. (1993). Causal forces: Structuring knowledge for time series extrapolation. *Journal of Forecasting, 12*, 103-115.

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

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

Armstrong, J. S. (2001). The forecasting dictionary. In J. S. Armstrong (Eds.), *Principles of forecasting* (pp. 761-824). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001). *Principles of forecasting.* Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S., Green, K.S., & Graefe A. (2013). *Golden rule of forecasting: Be conservative.* Working Paper Draft. Retrived from:

http://forecasters.org/wp/wp-content/uploads/gravity_forms/7-2a51b93047891f1ec3608bdbd77ca58d/2013/07/Green_Kesten_ISF2013.pdf

Bagnoli, F., Guazzini, A., & Lio P. (2008)). *Human heuristics for autonomous agents,* in Bio-Inspired Computing and Communication, Lecture Notes in Computer Science 5151 (pp. 340-351). Berlin: Springer.

Barkoulas J. T., & Baum C. F. (1998). Fractional dynamics in Japanese financial time series. *Journal of Pacific-Basin Finance, 6*, 115-124.

Batchelor, R., & Dua, P. (1990). Forecaster ideology, forecasting technique and the accuracy of economic forecasts. *International Journal of Forecasting, 6*, 3-11.

Benson, P. & Onkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting, 8*, 559-573.

Bloomfield, P. (1992). Trends in global temperature. *Climatic Change, 2*, 1-46.

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. (1995). Judging the probability that the next point in an observed time-series will be below, or above, a given value. *Journal of Forecasting,* 14, 597-607.

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

Box, G. E. P., & Jenkins, G. M., (1976). *Time Series Analysis: Forecasting and Control*: San Francisco: Holden-Day.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (2005). Processing abstract sequence structure: Learning without knowing, or knowing without learning? *Psychological Research, 69* (5-6), 583-398.

Bromiley, P. (1987). Do forecasts produced by organizations reflect anchoring and adjustment? *Journal of Forecasting, 6*, 201-210.

Byrk, A. S., & Raudenbush, S. W. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage Publications.

Bubic, A., Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 4*(25), 1-15.

Bunn, D. W., & Wright, G. (1991). Interaction of judgmental and statistical forecasting methods: Issues and analysis. *Management Science, 37*, 501-518.

Busemeyer, J., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432-459.

Cajueiro, D. O., & Tabak, B. M, (2008). Testing for long-range dependence in world stock markets. *Chaos, Solitions and Fractals, 37*, 918-927.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science, 32*, 36-67.

Chater N., & Brown, G.D.A. (1998). Scale invariance as a unifying psychological principle. *Cognition, 69*, B17-B24.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559-583.

Clements, M. P. (1995). Rationality and the role of judgement in macroeconomic forecasting, *The Economic Journal, 105*, 410-429.

Cleveland, W., McGill, M., & McGill, R. (1988). The shape parameter of a two variable graph. *Journal of the American Statistical Association, 83*, 289-300.

Cohen, B. L., & Wallsten, T. S. (1992). The effect of constant outcome value on judgments and decision-making given linguistic probabilities. *Journal of Behavioral Decision Making, 5,* 53–72.

Cont R., (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance, 1*, 223-236.

Crandall, V. J., Solomon, D., & Kelleway, R. (1955). Expectancy statements and decision times as functions of objective probabilities reinforcements. *Journal of Personality*, *24*, 192-203.

Delignières, D., Fortes, M., & Ninot, G. (2004). The fractal dynamics of self-esteem and physical self. *Nonlinear Dynamics in Psychology and Life Sciences, 8*, 479-510.

Diebold, F. X., & Rudebusch G .D. (1989). Long memory and persistence in aggregate output. *Journal of Monetary Economics, 24*, 189-209.

Edmundson, R. (1990). Decomposition: A strategy for judgmental forecasting. *Journal of Forecasting, 9*, 305– 314.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision-making* (pp. 27-40). New York: Plenum Press.

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

Eltahir, E. A. B. (1996). El Nino and the natural variability in the flow of the Nile River. *Water Resources Research, 32*(1), 131-137.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6*, 1-27.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12,* 391-396.

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

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

Evans, T. E. (1996). The effects of changes in the world hydrological cycle on availability of water resources. In: Global Climate Change and Agricultural Production (Eds.) Bazzaz F., & Sombroek W. *Direct and Indirect Effects of Changing Hydrological Pedological and Plant Physiological Processes*, (Chapter 2). Chichester, West Sussex, UK: FAO & John Wiley.

Fiedler, K., Freytag, P., Unkelbach, C., Bayer, M., Schreiber, V., Wild, B., & Wilke, M. (2004). *Subjective validity judgments of fictitious research findings: A paradigm for investigating sensitivity to sampling bias.* Unpublished manuscript, University of Heidelberg.

Fiedler, K., & Juslin, P. (2006). *Information sampling and adaptive cognition.* New York: Cambridge University Press.

Fildes, R., & Stekler, H. (2002). The state of macroeconomic forecasting. *Journal of Macroeconomics, 24*(4), 435-468.

Fildes R., Goodwin P., Lawrence M., & Nikolopoulos K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*, 3–23.

Froehlich, J., Chen, M., Consolvo, S., Harrison, B., & Landay, J. (2007). *MyExperience: A System for in situ tracing and capturing of user feedback on mobile phones*, Proceedings of MobiSys*,* 57-70.

Giannone, D., Reichlin L., & Sala L. (2004). Monetary Policy in Real Time. In *NBER Macroeconomics Annual*, (Eds.) Gertler M., & Rogoff K. (pp. 161–200). Boston: MIT Press.

Gigerenzer, G. (2006). Bounded and rational. *Contemporary debates in cognitive science,* 115–133.

Gilden, D. L., Schmuckler, M. A., & Clayton, K. (1993). The perception of natural contour. *Psychological Review, 100,* 460-478.

Gilden, D. L. (2009). Global model analysis of cognitive variability. *Cognitive Science, 33,* 1441-1467.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press.

Glaser, M., Langer, T., & Weber, M. (2007). On trend recognition and forecasting ability of professional traders. *Decision Analysis, 4*(4), 176-193.

Goodwin, P., & Wright G. (1993). Improving judgmental time series forecasting: a review of the guidance provided by research. *International Journal of Forecasting, 9,* 147-161.

Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time-series forecasting. *Omega: International Journal of Management Science, 22,* 553-568.

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

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

Goodwin, P., Önkal-Atay, D., Thomson, M.E., Pollock, A.C., & Macaulay, A. (2004). Feedback-labeling synergies in judgmental stock price forecasting. *Decision Support Systems, 37*, 175-186.

Goodwin, P. (2005). Providing support for decisions based on time series information under conditions of asymmetric loss. *European Journal of Operational Research, 163*(2), 388-402.

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

Goodwin, P., Önkal, D., & Thomson M. (2010). Do forecasts expressed as prediction intervals improve production-planning decisions? *European Journal of Operational Research, 205,* 195-201.

Goodwin, P., Önkal, D., & Lawrence M. (2011). Improving the role of judgment in economic forecasting. In M. P. Clements & D. F. Hendry (Eds.), *The Oxford Handbook of Economic Forecasting*, (pp. 163-189). Oxford: Oxford University Press.

Gottschalk, A., Bauer, M.S., & Whybrow, P.C. (1995). Evidence of chaotic mood variation in bipolar disorder. *Archives of General Psychiatry, 52*, 947-959.

Gureckis, T. M., & Love, B. C. (2010). Direct associations or internal transformations? Exploring the mechanisms underlying sequential learning behavior. *Cognitive Science, 34*, 10–50.

Harris, A. J. L., & Hahn, U. (2011). Unrealistic optimism about future life events: a cautionary note. *Psychological Review, 118*, 135–154.

Harvey, N. (1988). Judgmental forecasting of univariate time series. *Journal of Behavioral Decision Making, 1,* 95-110.

Harvey, N., Bolger, F., & McClelland, A. (1991). *Judgmental forecasting within and across correlated time series*. Department of Psychology, University College London Working paper.

Harvey, N., Bolger, F., & McClelland, A. (1994). On the nature of expectations. *British Journal of Psychology, 85*, 203-229.

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

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

Harvey, N., Ewart T., & West R. (1997). Effects of Data Noise on Statistical Judgment. *Thinking and Reasoning, 3*(2), 111-132

Harvey, N., & Fischer I. (1997). Taking Advice: accepting help, improving judgment and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*(2), 117–133.

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & reasoning, 13*(1)*,* 5-24.

Harvey N. (2011). Learning judgment and decision making from feedback: an exploration-exploitation trade-off? in *Judgment and Decision Making as a Skill: Learning, Development, and Evolution*, (Eds.) Dhami M. K., Schlottmann A., Waldmann M., editors. Cambridge: Cambridge University Press.

Harvey, N., & Reimers, S. (2012). Bars, lines, and points: the effect of graph format on judgmental forecasting. *In the Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, 2011, IIF.

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

Harvey, N. (In Press). Anchoring and adjustment: A Bayesian heuristic? In W. Brun, G. Keren, G. Kirkebøen, & H. Montgomery (Eds.), *Perspectives on Thinking, Judging, and Decision Making*. Oslo: Universitetsforlaget.

Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M., & Mass, O. F. (1992). The application of signal detection theory to weather forecasting behaviour. *Monthly Weather Review*, *120*, 863–883.

Hertwig, R., Barron, G., Weber, E., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539.

Hogarth, R. M., (1981). beyond discrete biases: functional and dysfunctional aspects of judgmental heuristic. *Psychological Bulletin, 90*(2), 197-217.

Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Science, 27*, 115-138.

Hogarth, R. M., Portell, M., & Cuxart, A. (2007). What risks do people perceive in everyday life? A perspective gained from the experience sampling method (ESM). *Risk Analysis, 27,* 1427–1439.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrica, 68*, 165-176.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110,* 291-304.

Hurst, H. E. (1951). Long-term storage capacities of reservoirs. *Transactions of the American Society of Civil Engineering, 116*, 776-808.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

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

Keren, G. (1983). Cultural differences in the misperception of exponential growth. *Perception & Psychophysics, 34,* 289-293.

Kiani, R., Hanks, T.D., & Shadlen, M.N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is

dictated by the environment. *The Journal of Neuroscience, 28(12),* 3017-3029.

Koutsoyiannis. D. (2000). A generalized mathematical framework for stochastic simulation and forecast of hvdrologic time series. *Water Resources Research, 36*(6), 1519-1534.

Koutsoyiannis, D. (2002). The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences, 47,* 573-595.

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

Lawrence, M., Edmundson, R., & O'Connor, M. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting, 1*, 25-35.

Lawrence, M., Edmundson, R., & O'Connor, M. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science, 32*, 1521-1532.

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

Lawrence, M., & O'Connor, M. (1993). Scale, variability and the calibration of judgmental prediction intervals. *Organizational Behavior and Human Decision Processes, 56*, 441-458.

Lawrence, M., & O'Connor, M. (1995). The anchoring and adjustment heuristic in time series forecasting. *Journal of Forecasting, 14*, 443-451.

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

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

Lawrence, M., & O'Connor, M. (2005). Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting, 21,* 3–14. doi:10.1016/j.ijforecast.2004.02.003

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

Legge, G. E., Gu, Y., & Luebker A., (1989). Efficiency of graphical perception. *Perception & Psychophysics, 46*, 365-374.

Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information - A review of the literature. *European Journal of Operational Research, 213,* 459–469.

Lemos, M. C., Finan, T., Fox, R., Nelson, D., & Tucker, J. (2002). The use of seasonal climate forecasting in policymaking: lessons from Northeast Brazil. *Climatic Change, 55*, 479–507.

Lewandowsky S. (2011). Popular consensus: climate change is set to continue. *Psychological Science, 22*(4), 460-463.

Lichtenstein, S., & Slovic, P. (1971). Reversal of preferences between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*, 46-55.

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

Lim, J., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems, 16,* 339-357.

Loewenstein, G., Read, D., & Baumeister, R. F. (Eds.). (2003). *Time and decision: Economic and psychological perspectives on intertemporal choice.* New York: Russell Sage Foundation.

Lopes, L., & Oden, G. (1987). Distinguishing between random and non-random events. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 13*, 392-400.

Lusk, C. M., & Hammond, K. R. (1991). Judgment in a dynamic task: microburst forecasting. *Journal of Behavioural Decision Making, 4*, 55-73.

MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In J. S. Armstrong (Eds.), *Principles of Forecasting.* Kluwer: Norwell, MA.

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

Madison, G. (2001). Variability in isochronous tapping: Higher order dependencies as a function of intertap interval. *Journal of Experimental Psychology: Human Perception and performance, 27*, 411-422.

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

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

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

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting, 16,* 451-476.

Mandelbrot, B. B., & Wallis, J. R. (1969). Computer experiments with fractional Gaussian noises. Part 1: Averages and variances. *Water Resources Research, 5*(1), 228-241.

Mandelbrot, B.B., & Wallis, J.R. (1969). Computer experiments with fractional Gaussian noises. Part 2: Rescaled ranges and spectra. *Water Resources Research, 5*(1), 242-259.

Mandelbrot, B.B., & Wallis, J.R. (1969). Computer experiments with fractional Gaussian noises. Part 3: Mathematical appendix. *Water Resources Research, 5*(1), 260-267.

Mandelbrot, B. B. (1971). A fast fractional Gaussian noise generator. *Water Resources Research, 7*(3), 543-553.

Mandelbrot, B. B. (1977). *The Fractal Geometry of Nature.* Freeman, New York, USA.

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

Mathews B. P., & Diamantopoulos A. (1989). Judgmental revision of sales forecasts: A longitudinal extension. *Journal of Forecasting, 8,* 129-140.

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

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

McDaniels, T., Mills, T., Gregory, R., & Ohlson, D. (2012). Using expert judgments to explore robust alternatives for forest management under climate change. *Risk Analysis, 32(12)*, 2098-2112.

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

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

Mentzer, J. T., & Kahn, K. (1997). State of sales forecasting systems in corporate America. *Journal of Business Forecasting, 16,* 6-13.

Mosteller, F., Siegel, A., Trapido, E., & Youtz, C. (1981). Eye fitting straight lines. *American Statistician, 35*(3), 150.

Muradoglu, G., & Önkal, D. (1994). An exploratory analysis of the portfolio managers' probabilistic forecasts of stock prices, *Journal of Forecasting, 13*, 565-578.

Nicholls N. (1999). Cognitive illusions, heuristics and climate prediction. *Bulletin of the American Meteorological Society, 80*(7)*,* 1385-1397.

Nikolopoulos K., & Fildes R. (2013). Adjusting supply-chain forecasts for short-term temperature estimates: a case study in a Brewing company. *Journal of Management Mathematics, 24*(1), 79-88.

Nuthall, P. L. (2001). Managerial ability a review of its basis and potential improvement using psychological concepts. *Agricultural Economics, 24*, 247-262.

O'Connor, M., & Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting, 8*, 141–155.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental 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., & Muradoglu, G. (1994). Evaluating probabilistic forecasts of stock prices in a developing stock market. *European Journal of Operational Research, 74*, 350–358.

Önkal, D., & Muradoglu, G. (1996). Effects of task format on probabilistic forecasting of stock prices. *International Journal of Forecasting, 12*, 9–24.

Önkal-Atay D. (1998). Financial forecasting with judgment. In: G. Wright, P. Goodwin (Eds.), *Forecasting with Judgment* (pp. 139–167). Chichester: Wiley.

Önkal, D., Yates, J. F., Şımga-Muğan, C., & Öztin, Ş (2003). Professional vs. amateur judgment accuracy: the case of foreign exchange rates. *Organizational Behavior and Human Decision Processes, 91*, 169–185.

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

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

Önkal, D., Lawrence, M., & Sayım K. Z. (2011). Influence of differentiated roles on group forecasting accuracy. *International Journal of Forecasting, 27*, 50-68.

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

Ong, S. E., & Chew, T. I. (1996). Singapore residential market: an expert judgemental forecast incorporating the analytical hierarchy process. *Journal of Property Valuation and Investment, 14*(1), 50-66.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Payne, J. W. (2005). It is whether you win or lose: the importance of the overall probabilities of winning or losing in risky choice. *Journal of Risk Uncertainty*, *30*, 5–19.

Pollock, A. C., Macaulay, A., Onkal-Atay, D., & Wilkie-Thomson, M. E. (1999). Evaluating predictive performance of judgmental extrapolations from simulated currency series. *European Journal of Operational Research*, *114*, 281-293.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53*(3)*,* 195–237.

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

Remus, W., O'Connor, M., & Griggs, K. (1995). Does reliable information improve the accuracy of judgmental forecasts? *International Journal of Forecasting*, 11, 285-293.

Remus, W., & Kottemann, J. (1987). Semi-structured recurring decisions: an experimental study of decision-making models and some suggestions for DSS. *Management Information Systems Quarterly, 11*(2), 233-244.

Remus, W., & Kottemann, J. (1995). Anchor-and-adjustment behavior in a dynamic decision environment, *Decision Support Systems, 15*, 63-74.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, *66*, 22–30.

Remus, W., O'Connor, M., & Griggs, K. (1998). The impact of incentives on the accuracy of subjects in judgmental forecasting experiments. *International Journal of Forecasting, 14* (4), 515–522.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15,* 353-375.

Rowe, G., & Wright G. (2001). Expert opinion in forecasting: The role of the Delphi technique. In: Armstrong J. S. (Eds). *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Sanders, N. R., (1992). Accuracy of judgmental forecasts: A comparison. *Omega: International Journal of Management Science*, *20*, 353-364.

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

Sanders, N. R. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega: The International Journal of Management, 25*(2), 135–144.

Sanders, N. R., & Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In: Armstrong J. S. (Eds). *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega: The International Journal of Management, 31*, 511–522.

Savio, N. D., & Nikolopoulos, K. (2009). Forecasting the effectiveness of policy implementation strategies. *International Journal of Public Administration, 33*, 88–97.

Sniezek, J. A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting, 5*, 171-178.

Sniezek, J. (1990). A comparison of techniques for judgmental forecasting by groups with common information, *Groups & Organizational Studies*, *14*(1), 5-19.

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

Speekenbrink, M., Twyman, M. A., & Harvey, N. (In Press). Change detection under autocorrelation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.),

Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Stephenson, D. B,, Pavan, V., & Bojariu, R. (2000). Is the North Atlantic Oscillation a random walk? *International Journal of Climatology, 20*, 1-18.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*, 881-911.

Song, H., Witt, S. F., & Zhang, X. (2008). Developing a web-based tourism demand forecasting system. *Tourism Economics, 14*(3), 445-468.

Straatemeier, T., Bertolini, L., & Brommelstroet, M. (2010). An experiential approach to research in planning. *Environment and Planning B: Planning and Design, 37*, 578–591.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: neural and computational mechanisms. *Frontiers in Neuroscience, 6*, 70.

Svenson, O. (1984). Cognitive processes in judging cumulative risk over different periods of time. *Organisational Behavior and Human Performance*, *33*, 22–41.

Syntetos, A. A., Boylan, J. E., & Disney, S. M., (2009). Forecasting for inventory planning: a 50-year review. *Journal of the Operational Research Society, 60*, 149–160.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., & Goodwin, P. (2009). The effects of integrating management judgment into intermittent demand forecasts. *International Journal of Production Economics, 118*(1), 72–81.

Thomson, M. E., Pollock, A. C., Henriksen, K. B., & Macaulay, A. (2004). The influence of the forecast horizon on the currency predictions of experts, novices and statistical models. *European Journal of Finance, 10*, 290-307.

Thomson, M.E., D. Önkal, A. Avcioglu, & P. Goodwin (2004). Aviation risk perception: A comparison between experts and novices. *Risk Analysis, 24*, 1585-1595.

Timmers, H., & Wagenaar, W. A. (1977). Inverse statistics and the misperception of exponential growth. *Perception & Psychophysics, 21*, 558-562.

Todd, P., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and brain sciences, 23*(5)*,* 727–741.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*, 440–463.

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the*

National Academy of Sciences of the United States of America, *109*(24), 9659-64. doi:10.1073/pnas.1119569109

Tsetsos, K., Usher, M., & McClelland, J. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in neuroscience, 5*. doi:10.3389/fnins.20- 11.00063.

Tsoularis, A., & Wallace, J. (2002). Analysis of logistic growth models. *Mathematical Biosciences, 179*, 21-55.

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

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

Usher, M., & McClelland, J. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550.

Van Orden, G. C., Holden, J. C., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*, 331-350.

Vere, D. T., & Griffiths, G R. (1995). Modifying quantitative forecasts of livestock production using expert judgements, an application to the Australian lamb industry. *Journal of Forecasting, 14*, 453-464.

Vogel, R. M., Tsai, Y., & Limbrunner, J. F. (1998). The regional persistence of annual stream flow in the United States. *Water Resources Research, 34(12)*, 3445-3459.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin, 77*, 65–72.

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

Wagenaar, W. A., & Timmers, H. (1978). Extrapolation of exponential time-series is not enhanced by having more data points. *Perception & Psychophysics, 24,* 182-184.

Wagenaar, W. A., & Timmers, H. (1979). The pond-and-duckweed problem: three experiments on the misperception of exponential growth. *Acta Psychologica, 43*, 239-251.

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience, 19*, 843–854.

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

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39,* 806-820.

Welch, E., Bretschneider, S., & Rohrbaugh, J. (1998). Accuracy of judgmental extrapolation of time series data. characteristics, causes, and remediation strategies for forecasting. *International Journal of Forecasting, 14*, 95–110.

Westheimer, G. (1991). Visual discrimination of fractal borders. *Proceedings of the Royal Society of London, 243*, 215-219.

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In M.P. Zanna (Eds.), *Advances in experimental social psychology*. San Diego, CA: Academic Press.

Wong, K., Huk, A., Shadlen, M., & Wang, X. (2007). Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience, 1.*

Woodman, G. F., Vogel, E. K., & Luck, S. J. (2001). Visual search remains efficient when visual working memory is full. *Psychological Science, 12*(3)*,* 219-224.

Yaniv, I., & Kleinberger E. (2000). Advice taking in decision-making: egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260–281.