When does more mean worse? Accuracy of judgmental forecasting is nonlinearly related to length of data series

Zoe Theocharis and Nigel Harvey

University College London

Nigel Harvey
Department of Experimental Psychology
University College London
Gower Street
London WC1E 6BT
UK
Tel: +44 207 679 5387
Fax: +44 207 436 4276
Email: n.harvey@ucl.ac.uk

Running head: Accuracy of judgmental forecasting

Abstract

When people make forecasts from series of data, how does their accuracy depend on the length of the series? Previous research has produced highly conflicting findings: some work shows accuracy increases with more data; other research shows that it decreases. In two experiments, we found an inverted U-shaped relation between forecast error and series length for various series containing different patterns and noise levels: error decreased as the length of the series increased from five through 20 to 40 items but also decreased as the series length decreased from five through two to one item. We argue that, with short series, people use a simple heuristic approach to forecasting (e.g., the naïve forecast). With longer series, they extract patterns from the series and extrapolate from them to produce their forecasts. Use of heuristics is poorer but extraction of patterns is better when there are more items in the series. For series of intermediate length, neither type of strategy operates well, thereby producing the inverted U-shaped relation that we observed. Implications for unaided judgmental forecasting and for forecasting based on a combination of judgmental and statistical methods are discussed.

Keywords: Judgmental forecasting; Time series; Heuristics; Pattern extraction

1. Introduction

Demand forecasting plays an essential role in supply chain management. Although development of formal methods continues apace, surveys have shown that this type of forecasting still frequently relies on judgment [10, 11, 36, 37, 44, 45, 49, 55]. Nowadays, this reliance is most often partial: judgment is combined in some way with statistical forecasting. However, the most recent surveys have revealed that between 16% and 26% of respondents still use unaided judgment to make their forecasts [10, 11, 55]. Hence, it remains vital that we document the factors that affect the quality of unaided judgmental forecasting.

One of the most important issues still to be resolved in the literature on unaided judgmental forecasting concerns the effects of the amount of historical data used as a basis for forecasts. Although hard data are not available, it is likely that the length of series available to forecasters varies a great deal across firms and across SKUs within firms. There are various possible reasons for this. Consultants may encourage forecasters to use only recent data because of concerns about the relevance of earlier records. Retaining and retrieving data for forecasting purposes may be expensive (though perhaps less expensive than in the past). Also, the importance of retaining those data may not be fully appreciated, particularly when the ownership of a firm changes: at such a time, historical data may be lost. Finally, when products are relatively new, historical data series will be short.

In what follows, we report studies designed to address this issue. We ask how the accuracy of unaided judgmental forecasts depends on the length of the data series available as a basis for those forecasts.

1.1 Conflicts in existing literature

When forecasts are produced by formal statistical means, "increasing the amount of data will generally increase the accuracy of forecasts", though rate of improvement declines as series lengthen [35]. 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 were merely to extract the naïve forecast (i.e. to use the last data point as the forecast for the next one).

As far as we can determine, there are just three studies that address this issue directly. In the first one, Wagenaar and Timmers [52] 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 [52] 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 we expect with formal approaches to forecasting.

In the second study, Lawrence and O'Connor [30] 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 [30] 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 [30] and reviewers of these results [14, 53] have suggested possible explanations. One is that that people suffer from cognitive overload when they are presented with too much data[1]. Research on effects of information load has indeed shown that an increase in information helps decision-making processes initially but that, after a certain point, any additional information has a detrimental effect, reducing the quality of decisions [2, 22, 40]. As a result, a U-shaped relationship between amount of available information and judgment error has been observed [4, 18, 54].

The third study was carried out by Andersson, Gärling, Hedesström and Biel [1]. 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). The results of this study appear to contradict those of the other two. Unlike them, they are consistent with what we would expect 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 [1] study used series of independent data points with or without a linear trend. In Wagenaar and Timmers [52] study, series had non-linear trends and, in Lawrence and O'Connor's [30] study, points were not independent: in other words, series were more complex than in Andersson et al's [1] 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 [52] study (3, 5 and 7), high in Lawrence and O'Connor's [30] study (20 and 40) but between these two extremes in Andersson et al's [1] experiments (5, 10, and 15).

1.2. Hypotheses

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 error [c.f. 4, 18, 54] because of the effects of information load. Hence, we test the hypotheses that the relation between forecast accuracy and series length varies with series complexity (H_1). Furthermore, because of the effects of information load discussed above, we expect there to be a U-shaped relation between forecast error and series length (H_2).

2. Study 1

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, we manipulated the length of the time series and the complexity of the pattern in the data series[3].

2.1 Method

2.1.1. Participants

One hundred and fifty students (52 men, 98 women) from University College London 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. They had not attended a course on forecasting.

2.1.2. 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 cyclical 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 1 with optimal forecasts.

Linear trended series were generated from the equation: $X_t = 5t + \varepsilon_t$. 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.

Cyclical series were constructed by using the equation: $X_t = 70\cos(100t + 20) + 170 + \varepsilon_t$, 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 cyclical cycle towards its peak. The wavelength of each cycle was 15.9 periods and this was

represented in the graphical display as a sequence of 12 points (Figure 1b). There were 3.33 wavelengths in the screen. Each wavelength's width corresponded to a 30% of the screen width.

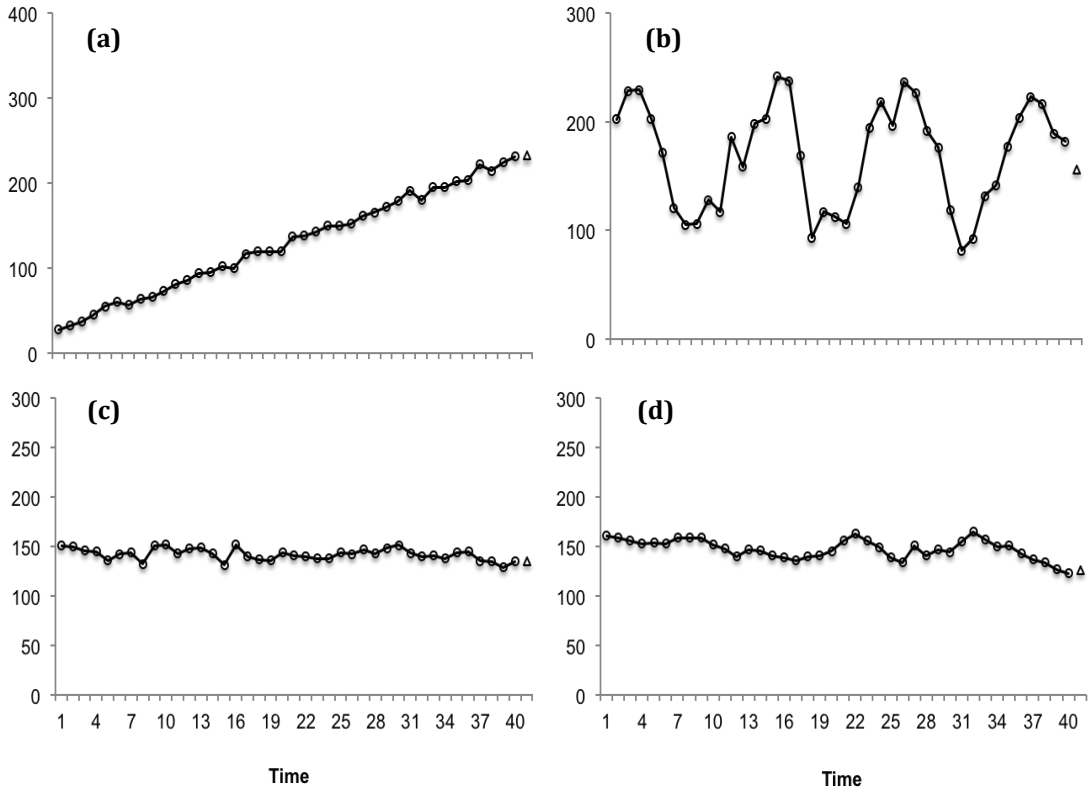

Figure 1. Examples of the four types of series comprising 40 data points (seen by participants) followed by the optimal forecast (not seen by participants) are shown for the four series types: a) linearly trended, b) cyclically trended, c) fractal, and d) linear autoregressive.

The autocorrelated series were produced 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 ($\alpha = 0.9$), 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 we used the multiple time-scale fluctuation approach [27]. 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 [12].

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 [3, 29]. Hence, the vertical axes of the graphs used to present the series were unlabelled. However, a numerical scale for the vertical axis was provided as shown in Figure 1.

2.1.3. Design

Participants were randomly assigned to one of five groups, each of which corresponded 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, 20, five, two, or one data point(s) 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. In other words, each participant made exactly four forecasts. 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.

2.1.4. Procedure

Participants performed the task one at a time on a computer. No other participants were present in the room but the experimenter (ZT) was available to answer questions. They read a short introduction to the study that provided them with instructions for their task. They were told that they would see points on graphs that depicted values of a

variable over time but given no further information about the time series that they would see. They were asked to forecast the next point in each series as accurately as possible. They were told that the accuracy of each of their forecasts would be measured by its absolute error and that the five participants with the lowest mean absolute error scores across all four of their forecasts would receive £5.00. After receiving their instructions, they entered their demographic details (age, sex).

The experiment then 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. A forecast was made on this line by moving the mouse. The chosen vertical position for the forecast was signified by a blue dot that 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. The experiment took approximately 10 minutes to complete.

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 cyclical series of 20 data points is shown in Figure 2. In this figure, we have also depicted the vertical bar on which participants made their forecasts.

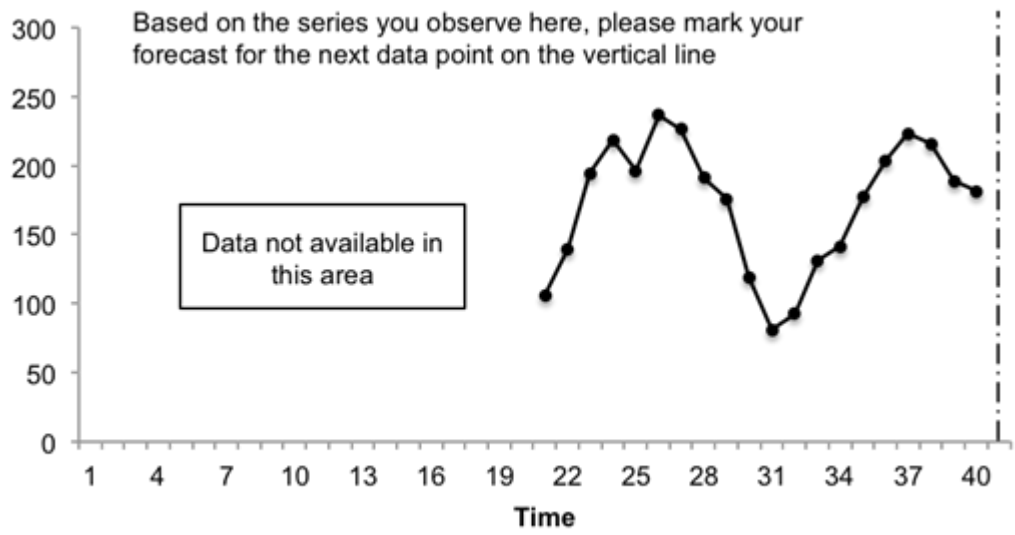

Figure 2. Example of the task with 20 data points of a cyclically trended series and showing the vertical bar on which participants made their forecast for the immediate (one step ahead) forecast horizon.

2.2. Results

For six participants, all four of their forecasts were at least 3 inter-quartile ranges from the median of each group. It was clear that they had not understood the task at all or were not taking it seriously. Hence, they were removed and replaced by new participants drawn from the same pool. This resulted in a total of 150 participants, thirty in each length condition.

To assess H_1 and H_2, absolute errors were calculated and compared across the five length conditions. The base line against which these errors were measured was the optimal forecast produced by the equation that generated the series but without the random noise component.

Graphs of MAE against series length (Figure 3) show an inverted U-shape function for all series' types. To examine the significance of these effects, we 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.

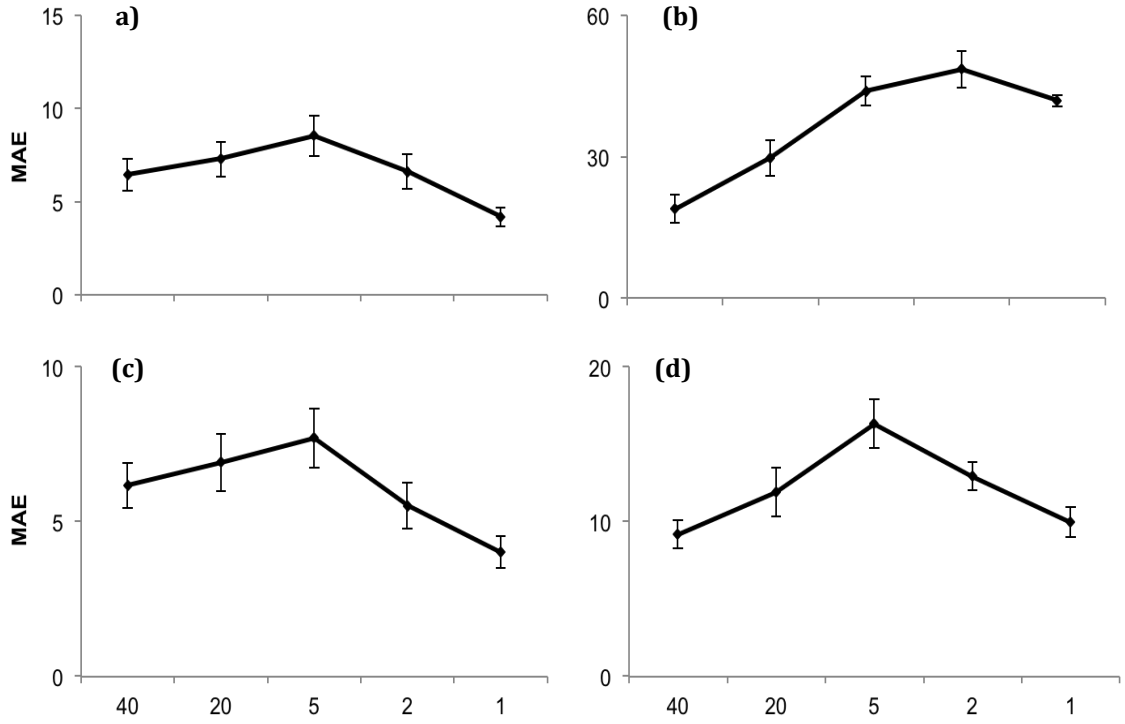

Figure 3. Graphs of mean absolute error (together with standard error bars) against series length for the four different types of series: a) linearly trended, b) cyclically trended, c) fractal, and d) linear autoregressive.

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 again 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 used to compare participants' predictions among the ten different pairs of lengths. Two-

tailed tests ($p < 0.05$) 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 cyclical 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 ($F(4, 71.39) = 4.14$; $p < 0.025$). 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 ($p < 0.025$); in all other cases, no significant differences occurred.

For all series types, these analyses are consistent with our second hypothesis (H_2) that the relation between forecast accuracy and series length is non-linear: for each of the four types of time series, the contrasts analysis showed the quadratic component to be significant. The analyses also show that the very short series length (L = 1) produced higher forecast accuracy than the medium length (L = 5). As the same inverted U-shaped relation between forecast accuracy and series length was obtained for all series types, there was no evidence for H_1.

2.3. Discussion

For all series types, forecast error was related to series length via an *inverted* U-shaped function rather than via the U-shaped function predicted by the effects of information load. Thus 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 cyclically trended series), and then decreased again for L = 1 and L = 2 (L = 1 for cyclically trended series). This finding appears robust in that it holds for series containing a variety of different patterns

These results are consistent with those of Andersson et al [1]. 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 [52]: 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 [52] and that it increases with longer series [1] 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 error to series length.

We can make some tentative inferences about the cognitive processes underlying forecasting performance. We base our interpretation of our results on dual processing models of cognition [5, 6, 24, 47]. In Kahneman's [24] realisation of this approach, System 1 carries out intuitive processing that employs heuristics and that produces adequate results with little cognitive effort whereas System 2 carries out deliberative processing that produces much better results with much more cognitive effort.

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 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 extrapolating from patterns extracted from the series. This is likely to involve deeper deliberative processing.

When no pattern information can be extracted from series because they are too short, forecasters have no alternative but to use heuristic approaches. We know that heuristic processing can be impaired by provision of more information: this is known as the less-is-more effect and has been well-documented [e.g.,]. Thus, if heuristic methods are used to forecast from very short series (L = 1), we would expect them to perform less well as series lengthen somewhat (L = 2). As series lengthen further, there is increasing likelihood that pattern information can be used for forecasting. For fairly short series (e.g., L = 5), it is unlikely that any deliberative pattern extraction performs much better than heuristic approaches. However, further lengthening of the series (L = 20; L = 40)

allows pattern extraction to become increasingly effective. As a result of the switch from heuristic to deliberative processing as series length increases, the observed inverted U-shaped curve relating forecast error to series length is obtained.

Moritz, Siemsen and Kremer [38] have also argued that dual process theories are applicable to judgmental forecasting from time series. Participants were presented with series comprising 30 periods (Studies 1 and 3) or 76 periods (Study 2). These series lengths are within the range that we would expect to be processed deliberatively by pattern extraction. Moritz et al showed that forecasters who scored highly on a cognitive reflection test that measures components of deliberative processing produced forecasts with lower MAE scores. Thus their results support our contention that forecasts made from longer series (> 20 periods) are made by deliberative processing.

Overall, these results indicate that the conflict between Wagenaar and Timmers [49] findings and Andersson et al's [1] findings arose because they examined series covering different ranges of lengths (H_2) rather than because they examined series of different levels of complexity (H_1).

Our results are not consistent with those of Lawrence and O'Connor [30]. However, their experiment differed from that of Wagenaar and Timmers [52] and from our own in a number of ways. For example, they calculated the accuracy of forecasts by averaging over four horizons whereas we 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. We examine this possibility in the next study.

3. Study 2

The first study was able to reconcile the apparently conflicting results of Andersson et al [1] and Wagenaar and Timmers [52]: the former compared longer series drawn from that part of the inverted U-shaped curve where error decreased with increasing 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 [30] results remain anomalous: they used longer series but found that error decreased with decreasing length.

We mentioned above that, in contrast to the other studies, Lawrence and O'Connor [30] averaged error scores across four horizons. It is possible that, had they reported data only for the most immediate (first) horizon, their results would have been similar to those of Andersson et al [1]. 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 us to ask whether the inverted U-shaped curve relating error to series length that we found for the immediate forecast horizon is maintained or changed (e.g., reversed) for later forecast horizons. For example, one possibility is that the minimum accuracy in the U-shaped curve is shifted to the left for more distant horizons: a minimum accuracy at series lengths of 30-40 rather than 5-10 would allow us to reconcile the results obtained by Lawrence and O'Connor [30] with all the other findings.

Thus, our second experiment is similar to the first one, except that participants made forecasts for the third rather than for the first forecast horizon. Consequently, for each series, there was a larger gap between the last data point and the point for which a forecast was required.

3.1. Method

3.1.1. Participants

One hundred and fifty participants (81 men, 69 women) were recruited from Amazon's Mechanical Turk online pool. Their mean age was 33 years. They were paid 0.5 $ for their participation.

3.1.2. 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.

3.1.3. 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 shows an example of the task screen in this experiment.

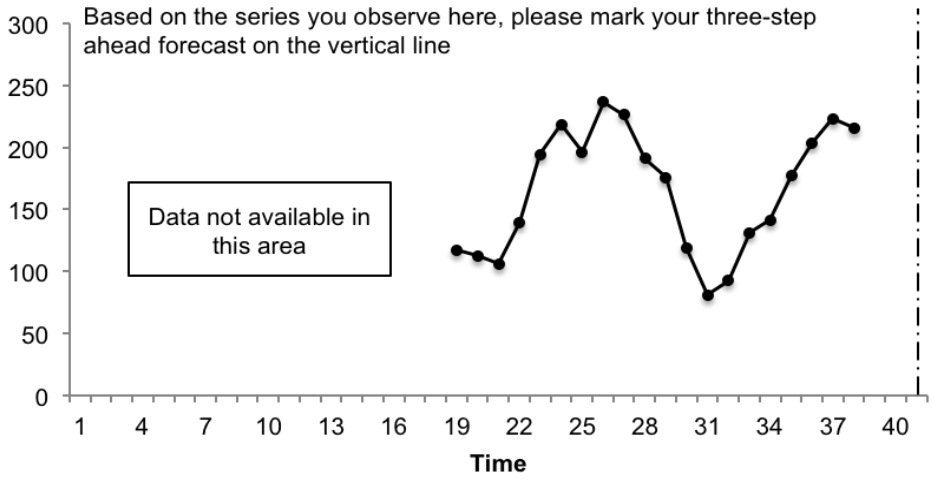

Figure 4. Example of the task with 20 data points of a cyclically trended series and showing the vertical bar on which participants made their forecast for the more distant (three steps ahead) forecast horizon.

3.2. 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.

3.2.1. Effects of series length on accuracy

Graphs of MAE against series length are shown in Figure 5 for each of the four series types.

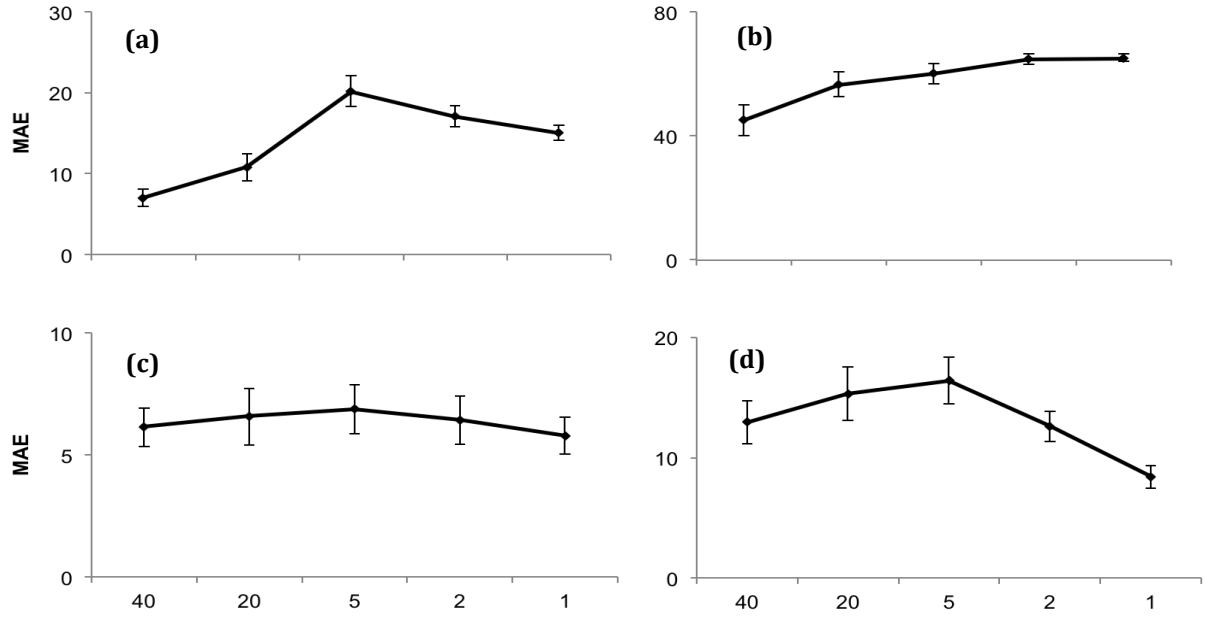

Figure 5. Graphs of mean absolute error (together with standard error bars) against series length for the four different types of series: a) linearly trended, b) cyclically trended, c) fractal, and d) linear autoregressive.

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 ($p < 0.001$). The relation between forecast error and series length again described an inverted U-shaped curve with a minimum 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 cyclical series, there was a main effect of length across groups ($F(4, 68.48) = 4.80$; $p < 0.01$). 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 ($F(4, 70.44) = 5.21$; $p < .001$). Polynomial contrasts showed 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.

3.3. Discussion

To reconcile Lawrence and O'Connor's [30] results with our earlier findings and with those reported by Andersson et al [1] and Wagenaar and Timmers [52], the relation between forecast accuracy and series length would have had to have been radically different from how it appeared in Experiment 1. Accuracy would have had to have been higher for L = 20 than for L = 40.

This is not what we found. As Figure 5 shows, results were very similar to those in Experiment 1 (Figure 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 cyclically 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 1: instead they maintained the same high value. Otherwise, results were as before.

In summary, though error levels tended to be considerably higher here than they were in Experiment 1 (particularly for linearly and cyclically trended series), the way that MAE depended on series length was very similar in the two experiments. 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 us to reconcile our results from Experiment 1, Andersson et al [1] results and Wagenaar and Timmers [52] results with the findings reported by Lawrence and O'Connor [30].

4. General discussion

Wagenaar and Timmers [52] and Lawrence and O'Connor [30] found that, in contrast with forecasts produced by formal methods, judgmental forecasts were more accurate when made from shorter series. Effects of information overload could account for this effect [30, 53]. Such effects have been observed in other tasks [2, 22, 40].

However, Andersson et al's [1] finding that judgmental forecasts improved as length of data series increased from five to 10 or 15 items is not consistent with the effects of information load. Neither are our results. We found forecast error to be related to series length via an inverted U-shaped function rather than via the U-shaped function predicted by the effects of information load: accuracy dropped as series length increased to five items and then increased again as series length rose further. Thus, although the relation that we obtained was non-linear (H_2), it was the opposite to that predicted by the effects of information load.

The inverted U-shaped function between forecast error and number of items that we found can be explained in terms of a shift from heuristic to systematic processing as the number of items in the data series increases. This account is also consistent with the findings of Andersson et al [1] and Wagenaar and Timmers [52].

With very short series, forecasters cannot extract patterns from the data with any degree of confidence. To produce a forecast, they have no choice but to fall back upon a heuristic that does not depend on ability to extract patterns from the data. One such heuristic is the naïve forecast. Studies in many domains have shown that the naïve forecast is not out-performed by forecasts produced by much more complex formal methods [15, 46]. It is the only approach that can be used with only a single data point. However, when there are two data points available, forecasters can extrapolate from the 'local trend' [17]: for example, if the last point was 16 and the one before it was 10, they could forecast that the next point would be 22.

With long series, forecasters have sufficient data available to use deliberative processing to extract patterns from the data with some degree of confidence. We assume that they then use these patterns, in conjunction with their real-world knowledge [16, 43] to produce their forecasts. Moritz et al's [38] findings that ability to forecast from these long series is correlated with scores on a test that measures aspects of deliberative processing supports this view.

As with the account of the U-shaped relation between judgment error and amount of information, we assume that the quality of systematic processing decreases with less information but that the quality of heuristic processing increases with less information[25, 48] . Thus, as the length of the series decreases, a point is reached where the quality of forecasts produced by pattern extraction is worse than that

produced by heuristic processing. Near this point, forecasters shift to heuristic processing. Once they have done so, their accuracy increases as the length of the series decreases further. The point at which the shift occurs may vary somewhat for different types of series. It may not even be a sudden switch in mode of processing. For series lengths at which both types of processing perform poorly (because they are too long for good heuristic processing and too short for good systematic processing), forecasts obtained by both approaches may be extracted and then integrated.

Finally, we should mention that dual process theories have been criticized [26, 28]. This is primarily because the two systems are characterized not just by a dichotomy between intuitive and deliberative processing but also by a number of other dichotomies, such as emotional/logical, automatic/controlled, exemplar-based/rule-based, and so on. Dual system theory requires that these dichotomies are aligned: for example System 1 processing is not only intuitive but also emotional, automatic, and exemplar-based. Such alignment seems to stretch credibility: for example, it is unlikely that all intuitive processing is emotional. Dual process theorists have responded to these critiques [7]. From our perspective, this debate is not of crucial importance. We have made a distinction between intuitive and deliberative processing and, for explanatory purposes, we have embedded this distinction within Kahneman's [24] dual system model. However, our distinction does not depend on our accepting that model. It can be treated in a stand-alone fashion. Seen in this way, it is not susceptible to the above criticisms of dual system theory.

4.1 Results in the context of other areas of statistical judgment

We have focussed on the use of judgment to make forecasts from time series. However, this type of task is part of a larger research domain: statistical judgment. This is broadly

concerned with the quality of people's judgment when they make estimates of means, variances, autocorrelations, correlations, frequencies, and other statistical features of data. When provided with more data, formal procedures produce estimates that are less variable and less biased. Hence, if people operate as intuitive statisticians [41] or as 'naïve intuitive statisticians' [9, 23], they should also make better statistical judgments when provided with more data.

Studies in which people have been presented with data and required to make judgments about their means, variances or distributional properties have shown the expected effects of sample size in some cases [19, 20] but not in others [21, 32, 33, 34, 50]. As a result, Pollard [42, p 15] concluded that: "On the basis of these often conflicting results, there is insufficient support for the Peterson and Beach idea that descriptive tasks can be viewed as tasks on which subjects make inferences that are properly influenced by sample size".

Our results also indicate that the effect of sample size on judgment approximates what would be expected from a 'naïve intuitive statistician' in some cases (with relatively large samples) but not in others (with relatively small samples). Our suggestion that this occurs because people use different modes of processing in the two cases may help to reconcile some of the contradictions in the literature on other types of statistical judgment. Recently, Tong and Feiler [51] extended the notion of a naïve intuitive statistician developed by Fiedler [9] to enable it to account for a number of phenomena that are observed in judgmental forecasting from time series. However, their model is based on people using a single processing system and it is not clear how it would provide a basis for explaining the inverted U-shaped relation between forecast error and series length that we obtained.

4.2. Limitations

First, we failed in our attempt to reconcile Lawrence and O'Connor's [30] findings with our own and with those reported by Andersson et al [1]. Like us, they 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 we obtained the opposite result. There are some procedural differences that may help to explain these divergent findings. In Study 2, we excluded one of these procedural differences as the source of the conflicting results. However, others remain. For example, we used a variety of series types, including those with high levels of autocorrelation, whereas they employed autoregressive moving average series.

Second, it would be useful to examine a wider range of series lengths. For most series, we obtained a peak error with a series length of five periods. However, it is possible that error would have been even higher with a series length of 10 or 15 periods. While this would still produce a U-shaped function between accuracy and series length and remain consistent with our account of that relation, it could have implications for practice (discussed below).

Third, our participants were not practitioner forecasters. However, previous work indicates that level of expertise does not increase accuracy in tasks involving judgmental forecasting from time series alone [31]. In fact, inverse expertise effects have been reported [39, 57]. In any case, we would expect the U-shaped relation to be maintained even when overall level of accuracy varies: the quality of System 1 and System 2 processing may vary with expertise but the factors affecting both types of processing should remain constant.

4.3. Implications for practice

There are occasions when practitioners do need to make forecasts from very short series. As we mentioned above, products might be relatively new or data may be missing. In fact, Goodwin and Fildes [13] carried out a survey of company forecasting behaviour and found that, even when formal methods were used, statistical models were often fitted to very short data series (e.g., six points). As a result, they performed very poorly and the managers were therefore highly inclined to use their judgment to produce final forecasts.

If around 20 items are available from which to make forecasts, 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.) Increasing series length in this range allows better use of System 2 processing.

If around five items are available and logistics or costs prevent series length from being increased to 20 or more, then shortening the series to, say, one item is likely to improve accuracy of judgmental forecasts for most series types and not impair it for others. Decreasing the series length in this range should facilitate System 1 processing and improve forecast accuracy but this is an approach that appears counter-intuitive and unlikely to be implemented in practice. One alternative might be to provide rolling averages of the most recent periods to reduce the series length without discarding data.

Our results also have implications for those forecasters who take the average of forecasts produced by judgmental and statistical methods. As we have seen, in the most recent survey [13], this approach was adopted by almost one in five respondents. Results from the current study imply that the degree to which judgment is weighted in that average should depend on the length of series on which forecasts are based.

Whereas accuracy of statistical methods increases as length of series increases, accuracy of judgmental forecasting initially decreases and later increases as length of series increase. This suggests that, for series of intermediate length, the contribution of judgmental forecasts to the overall average should be de-emphasized (or, possibly, ignored altogether). This is clearly an area for further research.

5. Conclusion

Accuracy of judgmental forecasts first decreases and then increases as the length of time series increases. We attribute this effect to a switch in processing mode. Patterns cannot be reliably extracted from very short series and so forecasters use simple heuristic processing to make their forecasts. For longer series, they are able to use more systematic pattern extraction processes to produce forecasts. Heuristic processing is impaired by more data [25, 48], whereas systematic processing is improved by it. For series of intermediate lengths, neither approach performs well. Hence we observe the U-shaped relation between forecast accuracy and series length. This relation has implications for practice: series of intermediate length should be lengthened or shortened to increase accuracy of judgmental forecasts; when the average of a judgmental forecast and a formal statistical forecast is used as the final forecast, the weight given to the former should be reduced when series are of intermediate length.

Footnotes

1. An alternative explanation is that people are more likely to think that the patterns in series will change when those series extend over a longer time period. As a result, they are more likely to forecast away from points produced by simple extrapolation of the existing patterns in the series when the series has already extended over a longer period of time. Lawrence and O'Connor [30] liken this to the 'gamblers' fallacy', where runs or trends are expected to reverse [56]. However, without elaboration, this explanation cannot account for Wagenaar and Timmers' [52] 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.

2. The reason that performance deteriorates with more information (rather than merely failing to improve) is that System 1 processing is automatic: once activated, it cannot be inhibited. Hence people cannot ignore the additional information supplied to them even though this information impairs their performance.

3. While complexity can be a subjective construct, more complex series are harder for people to process [54] and their description requires more information [8]. On these grounds, the linearly trended series can be characterised as the simplest and the non-linear fractal series as the most complex. The other two types of series fall between these extremes.

References

[1] 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.

[2] Chewning, E. G., & Harrell, A. M. (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, *15*, 527–542.

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

[4] Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of the literature from organization science, accounting, MIS, and related disciplines. *The Information Society*, *20*, 325–344.

[5] Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems*. Journal of Personality and Social Psychology, 62,* 328-339.

[6] Evans, J. St. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences, 7,* 454-459.

[7] Evans, J. St. B. T. & Stanovich, K. E. (2013). Dual process theories of higher cognition: Advancing the debate, *Perspectives on Psychological Science, 8,* 223-241.

[8] Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630-633.

[9] Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*, 659-676.

[10] Fildes, R., & Goodwin, P. (2007). Good and bad judgment in forecasting: Lessons from four companies. *Foresight: The International Journal of Applied Forecasting, Fall 2007,* 5-10.

[11] Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting, Winter 2015,* 5-12.

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

[13] Goodwin, P. & Fildes, R. (2007). Forecasting in supply chain companies: Should you trust your judgment? *OR Insight, 24,* 159-167.

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

[15] Green, K. C. & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research, 68,* 1678-1685.

[16] 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.

[17] Hohle, S.M., & Teigen, K. H. (2016). Forecasting forecasts: The trend effect. *Judgment and Decision Making, 10,* 416-428.

[18] Hwang, M. I., & Lin, J. W. (1999). Information dimension, information overload and decision quality. *Journal of Information Science*, *25*, 213–218.

[19] Irwin, F.W. & Smith, W.A.S. (1956). Further tests of theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 52*, 345-348. (Erratum: *Journal of Experimental Psychology, 53*, 152)

[20] Irwin, F.W. & Smith, W.A.S. (1957). Value, cost, and information as determiners of decision. *Journal of Experimental Psychology, 54,* 229-232.

[21] Irwin, F.W., Smith, W.A.S. & Mayfield, J.F. (1956). Tests of two theories of decision in an "expanded judgment" situation. *Journal of Experimental Psychology, 51,* 261-268.

[22] Jacoby, J., Speller, D. E., & Berning, C. K. (1974). Brand choice behavior as a function of information load: Replication and extension. *Journal of Consumer Research*, *1*, 33–42.

[23] Juslin, P., Winman, A. & Hannson, P. (2007). The *naïve* intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*, 678-703.

[24] Kahneman, D. (2013). *Thinking, fast and slow.* London: Penguin Books Ltd.

[25] Katsikopoulos, K. V. (2010). The less is more effect: Predictions and tests. *Judgment and Decision Making,* 5, 244-257.

[26] Keren. G. & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4,* 533-550.

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

[28] Kruglanski, A. W. & Gigerenzer, G. (2011). Intutitive and deliberative judgments are based on common principles. *Psychological Review, 118*, 97-109.

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

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

[31] Lawrence, M.J., R.H. Edmundson and M. O'Connor, 1985. An examination of the accuracy of judgmental extrapolation of time series, *International Journal of Forecasting, 1,* 25-36.

[32] Lindskog, M. (2015). Where did that come from? – Identifying the source of a sample. *Quarterly Journal of Experimental Psychology, 68*, 499-522.

[33] Lovie, P. (1978). Teaching intuitive statistics II: Aiding the estimation of standard deviation. *International Journal of Mathematical Education in Science and Technology, 9,* 213-219.

[34]Lovie, P. & Lovie, A.D. (1976). Teaching intuitive statistics I: Estimating means and variances. *International Journal of Mathematical Education in Science and Technology, 7,* 29-39.

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

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

[37] Mentzer, J.T. and Kahn, K. B. (1995). Forecasting technique familiarity, satisfaction, usage, and application. *Journal of Forecasting, 14,* 465-476.

[38] Moritz, B., Siemsen, S. & Kremer, M. (2014). Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management, 23*, 1146-1160.

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

[40] Paquette, L., & Kida, T. (1988). The effect of decision strategy and task complexity on decision performance. *Organizational Behavior and Human Decision Processes*, *41*, 128–142.

[41] Peterson, C.R. & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29-46.

[42] Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychological Research and Reviews, 3*, 5-18.

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

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

[45] Sanders, N. R. & Manrodt, K. B. (2003). Forecasting software in practice: Use, satisfaction, and performance. *Interfaces, 33,* 90-93.

[46] Sherden, W.A. (1998). *The Fortune Sellers: The Big Business of Buying and Selling Predictions*. New York: Wiley.

[47] Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3-22.

[48] Smithson, M. (2010). When less is more in the recognition heuristic. *Judgment and Decision Making, 5,* 230-243.

[49] Sparkes, J.R. and McHugh A.K. (1984). Awareness and use of forecasting techniques in British industry. *Journal of Forecasting, 3,* 37-42.

[50] Spencer, J. (1963). A further study of estimating averages. *Ergonomics, 6,* 255-265.

[51] Tong, J. & Feiler, D. (2016). A behavioural model of forecasting: Naïve statistics on mental samples. Management Science, 63, 6609-6627.

[52] Wagenaar, W. A. & Timmers, H. (1978). Extrapolation of exponential time-series is not enhanced by having more data points. *Perception & Psychophysics, 24,* 182-184.

[53] Webby, R. and O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting, 12,* 91-118.

[54] Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M. & Harvey, N. (2018). Task complexity moderates the influence of descriptions in decisions from experience. *Cognition, 170,* 209-227.

[55] Weller, M., & Crone, S. F. (2012). Supply chain forecasting: Best practices & benchmarking study. Lancaster University. http://www.lancaster.ac.uk/ lums/forecasting/material/

[56] Xu, J. & Harvey, N. (2014). Carry on winning: The gamblers' fallacy creates hot hand effects in online gambling. *Cognition, 131,* 173-180.

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