European Journal of Operational Research 284(3):882-895 (14 pages) 01 Aug 2020

Using judgment to select and adjust forecasts from statistical models

De Baets, Shari* & Harvey, Nigel

Author details

Dr. Shari De Baets*

Tweekerkenstraat 2

9000 Ghent

Belgium

Faculty of Economics and Business Administration

Department of Business Informatics and Operations Management

Shari.debaets@ugent.be

(0032) 478 60 63 52

Professor Nigel Harvey

Department of Experimental Psychology,

University College London

Gower Street

London WC1E 6BT

United Kingdom

n.harvey@ucl.ac.uk

(44) 207 679 5387

* Corresponding author

Using judgment to select and adjust forecasts from statistical models

Abstract

Forecasting support systems allow users to choose different statistical forecasting methods. But how well do they make this choice? We examine this in two experiments. In the first one (N = 191), people selected the model that they judged to perform the best. Their choice outperformed forecasts made by averaging the model outputs and improved with a larger difference in quality between models and a lower level of noise in the data series. In a second experiment (N = 161), participants were asked to make a forecast and were then offered advice in the form of a model forecast. They could then re-adjust their forecast. Final forecasts were more influenced by models that made better forecasts. As forecasters gained experience, they followed input from high-quality models more readily. Thus, both experiments show that forecasters have ability to use and learn from visual records of past performance to select and adjust model-based forecasts appropriately.

Keywords: forecasting, judgmental selection, judgmental adjustment, forecast support systems

1. Introduction

Forecasting is important in many areas of human activity. Our research is framed in the context of demand forecasting in supply chain management, an area that has been the focus of much forecasting research in recent years. However, it is reasonable to expect that our findings will be relevant to forecasters in other domains in which predictions are made in a similar way.

In demand forecasting, predictions for future sales are based on records of previous sales and on information about future events that may affect sales levels. In the past, these sales forecasts were often made using unaided judgment. However, recent surveys (e.g., Fildes & Goodwin, 2007; Fildes & Petropoulos, 2015) have shown that nowadays between a sixth and a quarter of organizations use this approach, approximately a quarter use a purely statistical method, and just over a third make judgmental adjustments to statistical forecasts. Judgmental adjustments to statistical forecasts are typically made to allow for effects of exogenous events (e.g., planned promotions) that have effects not included in the statistical model. Research has identified the conditions under which these judgmental adjustments do and do not improve forecast accuracy (e.g., Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009).

Forecasting Support Systems (FSS) allow users to apply many different types of forecasting methods (algorithms) to make the required predictions. Typically, an FSS provides a graphical display of historical data and allows for a selection of different forecast methods, results of which are then overlaid on the data display. These different forecast methods typically produce different forecasts from

the same data series. However, no approach is universally superior to all others. Forecasting competitions have shown that which approach is best depends on the characteristics of the data series (e.g., Makridakis & Hibon, 2000). Thus, whether forecasters use an approach that is purely statistical, or one in which judgment is used to adjust statistical forecasts, they must decide which statistical model to use. Thus, in practice, judgment pervades the performance of forecasting tasks (Leitner & Wildburger, 2011; Perera, Hurley, Fahimnia, & Reisi, 2019).

Hence, we address two sets of questions. The first set is concerned with forecasters' ability to distinguish good statistical forecasting models from poor ones. Do their choices outperform a simple average of the available forecasting models? What factors influence these choices? Is it easier to discriminate between two models when the difference in quality between them is greater (i.e., when signal strength is higher) and when noise in the data series is lower? The second set of questions is concerned with judgmental adjustment of initial judgmental forecasts. Is the importance that people place on the advice they receive from a statistical model greater when the model is of higher quality? In other words, do they take more account of advice from better models? Is the degree to which they take account of advice affected by the feedback they gain from experience with different models?

We report two behavioural experiments. The first focusses on the first set of questions and therefore deals with judgmental selection of statistical models. We investigate how well people can select models by viewing records of past forecasts and outcomes and examine factors on which their ability to perform this task depends. The second experiment is concerned with forecasts made by combining judgment and the output of a statistical algorithm. We vary the quality of the statistical forecasts and examine whether judges responsible for producing the final forecasts are more influenced by forecasts produced by better statistical models. Also, if they are given feedback about their performance and are sensitive to it, the greater influence of good advice should increase as they gain experience with the task. We examine whether this is so.

The first experiment uses a choice paradigm, thereby making the judgmental selection of the best forecast and, thus, the detection of its value as source of advice, *explicit.* In the second experiment, recognizing the value of the forecast is measured by the degree of adjustment towards the forecast advice, thus making the detection of the value of the advice *implicit.*

We first discuss the state of the art in the literature on judgmental selection of forecasts. Next, we report an experiment to test the first set of questions posited above. This is followed by a brief discussion section and introduction to the second experiment. We conclude the paper with a discussion of the results of both experiments, limitations to our study and ideas for future research.

2. Literature review

Forecasting is an essential activity for companies who want to remain competitive in today's world. In Fildes & Petropoulos's (2015) survey, over half of their respondents indicated that they use some combination of statistical and judgmental methods in their forecasting. Given this pervasiveness of

judgment in forecasting practice, operational researchers have recognized that behavioural insights are needed and this has led to the development of behavioural operational research (Bendoly, Croson, Goncalves, & Schultz, 2010). This new area has received input from various disciplines. For example, some of the systematic deviations from normative decision making identified by cognitive psychologists have important implications for forecasting performance in practice. Some of these effects, such as trend damping (Bolger & Harvey, 1993; Lawrence & Makridakis, 1989) may arise from use of the heuristics (e.g., anchoring-and-adjustment) identified by Tversky and Kahneman (1974) whereas others, such as optimism (Weinstein & Klein, 1996), may arise from motivated reasoning (Kunda, 1990).

Researchers have examined three broad approaches to reducing these forecasting errors: a) selection of a presentation format for the data series that enhances forecasting ability, b) provision of feedback that enables people to improve their forecasting ability by learning from experience, and c) provision of advice from experts or forecasting algorithms. We will discuss these approaches in turn.

In our experiments, participants responded to data series and forecasts from those data series that were presented to them in graphical format. This type of presentation has long been use in studies of judgment in forecasting tasks (Willemain, 1989, 1991) and practitioners increasingly use data visualization not just in forecasting but also in other supply chain management tasks (Bendoly, 2016). There is good reason for this. People are better at making judgments about data that are graphically displayed than about data provided in numerical form. For example, Harvey and Bolger (1996) found that trends were harder to discern when data series were presented in a tabular than in a graphical format and that, as a result, forecast error was lower when data could be visualised. Furthermore, the type of graphical format (line graphs versus point graphs) influences accuracy of judgmental forecasts (Theocharis, Smith, & Harvey, 2019).

In our second experiment, we examine the effect of providing forecasters with outcome and performance feedback on their point forecasts and on point forecasts produced by forecasting algorithms. Outcome feedback provides information about the actual value of the variable being forecast or the optimal forecast that should have been made. Performance feedback gives information about the error in the forecast that has been made. A third type of information that can be given to forecasters is often erroneously termed "task properties feedback": it provides people with advance information about their task (e.g., statistical features of the data series to be forecast) and so should actually be called feedforward or guidance (Björkman, 1972). Should we expect outcome and performance feedback to improve accuracy? Reviewers are in agreement that the answer to this question is "no".

Goodwin and Wright (1993, p. 157) say that "Research suggests that outcome feedback has little value while task properties feedback can be effective". Consistent with this, Sanders (1997, pp. 136-137) argues that "Research studies have found that simple outcome feedback does not provide a significant contribution in helping subjects learn about judgmental tasks. … On the other hand, providing task properties feedback to subjects appears to improve the performance of judgmental tasks". Finally, Lawrence, Goodwin, O'Connor, and Önkal (2006, p. 507) do say that "Feedback has been

shown to improve the accuracy of point forecasts" but they are not specific about the type of feedback to which their statement refers. However, as the four papers that they cite as evidence in support of it all show effects of task properties feedback rather than outcome feedback, we assume that they are not claiming that outcome feedback has been shown to be effective. Thus the consensus is that outcome feedback (or the performance feedback implicit in its provision) does not improve accuracy in judgmental time series forecasting. However, explicit demonstrations of the ineffectiveness of outcome feedback have been published only for closely analogous judgment tasks (Harvey, 2011; Klayman, 1988).

In our second experiment, we investigate people's ability to use advice provided by a forecasting algorithm. There is a large literature on how well people are able to use advice (Bonaccio & Dalal, 2006). In general, people put too much weight on their own judgment and too little on information received from other sources (Harvey & Fischer, 1997; Yaniv, 2004) when they combine the two. That this is especially so when the other source is an algorithm (Gardner & Berry, 1995; Lim & O'Connor, 1995) implies that people are even more reluctant to take advice from an algorithm than from a human expert; in other words, they show algorithmic aversion. However, most research demonstrating this phenomenon has focussed on people's preference for *selecting* a human over an algorithm as a means of achieving some decision goal rather than on how those sources of information are weighted when *combined* to reach that goal. For example, Meehl (1954) showed that algorithmic methods outperform clinical judgment in diagnosis, prognosis and treatment decisions but that doctors prefer to use their judgment. Recent meta-analyses by Grove et al. (2000) and White (2006) included many studies carried out since Meehl's (1954) original one and found that his conclusions remain valid. Despite this, clinicians still prefer to use judgment (Grove, 2005; Keeffe et al., 2005). This phenomenon is now known as algorithm aversion.

Dietvorst, Simmons, and Massey (2015) were interested what causes it. They carried out a series of experiments in which people placed bets either (1) on their own performance or on that of a model or, (2) on the performance of someone else or on that of a model. Their profits at the end of the experiment were entirely dependent on the accuracy of the forecast source they chose throughout the session. In all their experiments, the model significantly outperformed judgments of the participants themselves and judgments of the other person. Despite this, people placed their bets much more often on their own judgment (or on the other person's judgment) than on the recommendations of the superior algorithm. They did this even when the superior performance of the algorithm was directly contrasted with the inferior performance of themselves or the other human. This effect was consistent across different tasks and incentive structures.

Dietvorst et al. (2015) also measured people's trust in the decisions based on algorithms and on those based on judgment. They found that confidence in algorithms dropped much more after seeing an error than their confidence in judgment did. It appears that people expect algorithms to be perfect: in Madhavan and Wiegmann's (2007) terms, they are subject to a 'perfection schema'. In contrast, they

expect humans to be imperfect. Hence, when an algorithm makes an error, people experience this as an unexpected event and, as a result, it has a greater negative effect on the reputation of the algorithm than it would have on the reputation of a human advisor making the same error. Prahl and Van Swol (2017) work supports this: they found that people's use of advice dropped significantly more after they received bad advice from an algorithm than bad advice from a human judge. Findings consistent with these conclusions have also been reported in studies of the effects of automation in other domains. For example, humans exaggerate the errors made by automated machines, even when such errors occur only very rarely (Dzindolet, Pierce, Beck, & Dawe, 2002; Hoff & Bashir, 2015).

Within the forecasting literature, preference for a human judge over an algorithm as a source of advice has been demonstrated by Önkal, Goodwin, Thomson, Gönul, and Pollock (2009). Their participants made an initial forecast from a graph of past sales data. They were then given advice before having the opportunity to revise their forecast. One group of participants were told that this advice came from an expert forecaster and a second group were informed that it had been produced by computer using a statistical forecasting algorithm. The advice itself was exactly the same: only its apparent source differed. Participants were much less accepting of the advice if they thought that it came from an algorithm. Those in both groups adjusted their forecasts towards the advice but did much more so when they had been told that the advice was produced by a human source.

According to the Hovland-Yale model of persuasion and attitude change (Hovland, Janis, & Kelly, 1953), factors that lead people to change their opinions can be classified into those related to the source of the new information, the nature of the message carrying the new information, the characteristics of the receivers of that information, and the context or channel through which it is transmitted. Advice Response Theory (Feng & MacGeorge, 2010) is based on this approach. It stresses the importance of message-related (e.g., politeness), advisor-related (e.g., expertise) and receiver-related (e.g., current mood) characteristics of the advice. Though it was developed to account for differences in effectiveness of advice given by people to other people, it can also be applied to cases in which humans receive advice from algorithms. For example, in Önkal et al.'s (2011) study, messages for both groups were the same but characteristics of the advisor were not. Perceived competence of advisors is likely to have differed. Sniezek and Van Swol (2001) have shown that these factors affect trust in and credibility of advice, which, in turn, predict the degree to which advice is utilized by those receiving it.

Forecasters assess the competence of human and algorithmic advisors differently. As we have seen, they expect algorithmic advisors to produce consistently good advice and, when they do not, the reputational consequences are more severe than they are for human advisors. Why is this? Forecasters receiving advice from an algorithm are likely to rely only on its performance whereas those receiving advice from a person can also take account of other factors, such as intentions, integrity, emotions, and nonverbal cues that are irrelevant when it comes to algorithmic advice. Hence, when an algorithm makes an error, it is perception of its competence that drops. In contrast to the situation in which a human advisor makes an error, there are no other characteristics that can be used to explain the failure as an

exceptional event rather than reflective of underlying competence (Gefen, Karahanna, & Straub, 2003; Paravastu, Gefen, & Creason, 2014).

As our aim here was to identify causal relationships, an experimental approach was required. This was because it provided us with control over the independent variables that we hypothesised would influence the dependent ones that we measured (Shadish, Cook, & Campbell, 2002). Hence, we carried out two behavioural experiments. In the first one, we examined forecasters' ability to distinguish good algorithmic-based forecasting advice from poor algorithmic-based forecasting advice (in the form of good versus poor statistical models). This experiment used a choice paradigm, making the detection of the quality of the forecast *explicit.* Later, in our second experiment, we studied whether forecasters' adjustment behaviour is influenced by the perceived accuracy of the statistical model, making the detection of the quality of the forecast *implicit*.

3. Experiment 1: Judgmental selection of statistical forecasts

Research on judgmental selection of statistical forecasts is limited in quantity. Lawrence, Goodwin, and Fildes (2002), focussing on the design of forecast support systems, divided participants into two groups: those who had total control over the FSS, including how it was displayed and selection of the model, and those who could not modify anything but were presented with an optimal model selection by the FSS. All participants could subsequently accept forecasts as specified by the FSS or else make adjustments to modify them. Participants in the first group made these adjustments significantly less often and were thus more accepting of the final model. However, this acceptance of the selected model forecast came at a price: participants were not very adept at selecting a good forecasting method and, consequently, their forecast accuracy was significantly worse than that of participants in the second (no-modification) condition.

Recently, Petropoulos, Kourentzes, Nikolopoulos, and Siemsen (2018; see also Petropoulos, 2019) compared three ways of selecting forecasting models; a) algorithmic selection, b) judgmental selection, c) selection based on judges' detection of the absence or presence of trends and seasonality in the data (termed 'model-build'). Selection was made from four exponential smoothing models that could capture different patterns in the data (level, trend, seasonality, trend plus seasonality). Furthermore, the data series were selected so that each of the four models was best for some of the series. Thus, there was a strong correspondence between the characteristics of the models that could be selected and the features of the data series for which the selection was made.

Petropoulos et al. (2018) found that judgmental selection was worse at selecting the best model than either the algorithmic selection or the model-build approach (which were no different from one another). However, judgmental selection was better than algorithmic selection at avoiding the 'worst' models. As a result, overall accuracy of judgmental selection was better than that of algorithmic selection. Although the tasks used are not directly comparable, this finding throws a much more positive light on judgmental selection than that of Lawrence et al. (2002). Was there some aspect of Petropoulos

et al.'s (2018) task that could account for this? Harvey (2019) highlighted the strong correspondence between the features of the models from which selection was made and those of the data series for which the selection was made. He argued that: "If the models had been more heterogeneous (e.g., exponential smoothing, ARIMA, frequency domain approaches, etc.) and the data series more varied (e.g., containing trends, autoregressive and moving average components, fractal patterns with different Hurst exponents, etc.), selecting the best model may have been much more difficult, and the accuracy may have been reduced below that of algorithmic selection".

Although these two studies do not provide unequivocal guidance about the method of model selection that should be adopted, it is likely that, in practice, judgmental selection is much more common than algorithmic selection. Judgment and forecasting are, in essence, inseparable (Perera et al., 2019) and, as we have seen, there is a great deal of evidence from research in many domains, including forecasting, that people prefer to use their judgment, even in situations in which objective measures show that it performs less well than algorithmic approaches. Further research into the quality of judgmental selection of forecasts is therefore important.

In this experiment, forecasters were required to identify the better of two models by comparing the two series of forecasts that those models had made in the past with the series of outcomes that those models had been forecasting. They were asked to select the model forecast that performed best, i.e., was closest to the real data. This task should be easier when signal strength (i.e., the difference in quality between the two models) is greater (Stanislaw & Todorov, 1999) and when the data series are subject to a lower level of noise (Han, Wang, Petropoulos, & Wang, 2018; O'Connor, Remus, & Griggs, 1993). In contrast to Petropoulos et al. (2018), different models were not more effective for different types of series: the task was not to match the model type to the series type that it was most effective in forecasting. Instead, performance of the three models was ranked in the same way for all series. We expected that people would find it more difficult to distinguish the first-ranked (i.e., good) model from the second-ranked (i.e., intermediate) model than to distinguish the first-ranked model from the third-ranked (i.e., poor) model and that their performance in both these discriminations would be worse when there was more noise in the data series. The objective benchmark with which participants' judgment accuracy was compared was that produced by the average of the two models from which they had made their selection. More formally:

Hypothesis 1 (H_1): The percentage of correctly identified (PCI) statistical forecasting models will be greater when data series contain low noise series than when they contain high noise.

Hypothesis 2 (H_2): The PCI will be greater when the best performing model is compared with the poor model than when it is compared with the model of intermediate quality.

Hypothesis 3 (H_3): Forecasting using the statistical model selected by participants leads to a lower mean absolute error (MAE) than the simple average of the forecasts produced by the two models from which participants made their selection.

3.1. Method

Each participant saw graphs of time series on their own individual computer screen and made their responses by clicking a mouse. Dependent variables were PCI and MAE. Independent variables were noise level in the data series (high versus low) and choice of statistical advice (between good and poor quality versus between good and intermediate quality). Both were varied between-participants in a factorial fashion. To ensure some generality to our findings, series autocorrelation was also varied.

3.1.1. Participants

One hundred and ninety-one participants took part in the study, of which 139 were female. Their mean age was 20.10 (SD = 4.10). Participation was mandatory as part of a course at University College London. Participants had received introductory courses on statistics and not participated in similar experiments by the researchers. The top performers (the three students who identified the best performing model most often) in every condition were awarded a £5.00 cash prize. Incentives such as these are often used by researchers to mimic real-life rewards tied to performance (Perera et al., 2019).

3.1.2. Stimulus materials

Sixty 50-point sales series were simulated with a mean value of 300. Half of the series had a standard deviation (SD) of 21 (low noise series, 7% of the mean) and the other half had an SD of 36 (high noise series, 12% of the mean). For each noise type, 10 series were independent (AR = 0.0), 10 series had positive autocorrelation (AR = 0.8) and 10 series had negative autocorrelation (AR = -0.8).

Three forecasting models were calculated for every series. These were the naïve forecast ($F_t = X_{t-1}$, where the last actual value in the series X_{t-1} served as the forecast F_t); an exponential smoothing forecast ($F_t = \alpha X_{t-1} + (1 - \alpha)F_{t-1}$, where the smoothing constant, α, was determined by minimizing the sum of the squared forecast error); and an autoregressive forecast ($F_t = aX_{t-1}$, where the first-order autocorrelation, a, that was employed to generate the data series, was used to make forecasts). Note that, over the long-term, the autoregressive forecast could not be outperformed and so, for convenience, we refer to it here as the 'ideal' forecast. (While business practitioners may have access to more than three models, we limited our choices to three levels of performance based on the simulated historical data. While a naïve forecast may be of value in certain circumstances, it was a poor choice when compared to the historic data that was provided to participants. Similarly, while an autoregressive forecast may not be ideal in practice, it was deemed so for this experiment as it was the underlying signal of our historic sales series.)

The data series were presented by a blue line graph labelled as 'sales'. Two out of three forecasting methods were selected, depending on the condition, and presented using a yellow line and a green line overlaid on the sales series indicating the statistical forecast history from week 2 to 50 and the forecast for week 51. These forecasts were labelled 'model 1' and 'model 2' (see Figure 1).

Models and series appeared in different colors to render them easily discriminable. The experiment was coded in PHP and run online.

3.1.3. Design

There were four conditions: noise level in the data series (high versus low) was crossed with type of choice of statistical advice (good/poor quality versus good/intermediate quality). Each participant was randomly assigned to one of these four experimental conditions, resulting in the sample sizes in each condition displayed in Table 1. The 30 data series within the two high-noise and two low-noise groups were presented in a different random order for each participant.

Table 1. Representation of the four experimental conditions with their sample sizes.

	Good (AR) vs. Poor (Naïve) model	Good (AR) vs intermediate (ES) model
Low Noise (7% of the mean)	N = 42	N = 49
High Noise (12% of the mean)	N = 53	N = 47

We report participants' MAE scores relative to the time series signal produced by the generating algorithm (i.e., excluding the contribution from noise term). This is equivalent to the ideal forecast: it produces the minimum error in the long-term. Hence, the MAE score for the autoregressive forecast was 0 (SD = 0). For comparison, the MAE score for the exponential smoothing forecast was 7.67 (*SD* = 6.69) and for the naïve forecast, it was 19.46 (*SD* = 19.55). These three errors are all significantly different from one another at $p < .001$.

3.1.4. Procedure

Participants were instructed to study the graphs they were given and to choose the most accurate forecasting model, i.e., the model with the lowest error. They initially saw the graph with only the sales series in form a line. When they were ready, they clicked on a button labelled 'Show forecasting models'. Both models then appeared overlaid on the sales series. A choice box appeared on the right side of the graph. Participants used this to indicate the type of forecast that they deemed to be most accurate. Once they had done this, the sales series was updated with the value for week 51. This enabled them to evaluate their choice post-hoc (but they could no longer change their choice). They then clicked a box to move on to the next graph. After they had responded to all 30 graphs, they provided demographic details (age and gender).

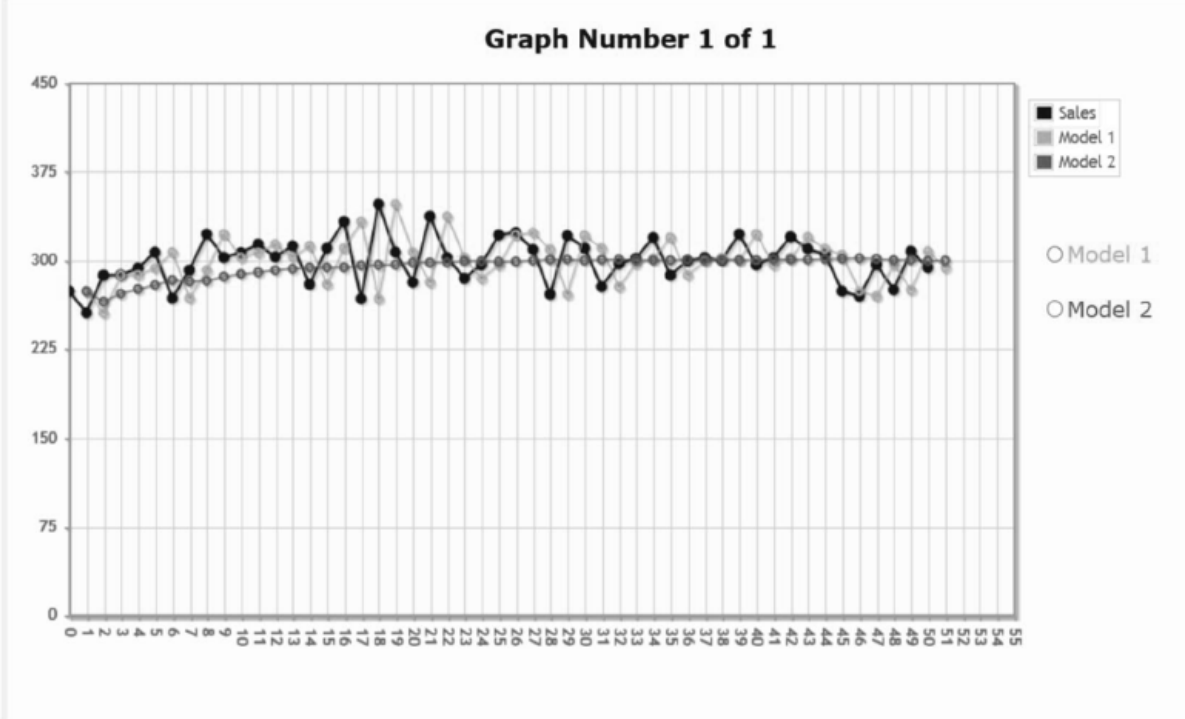

Figure 1. Display of an experimental trial.

3.2. Results

Below we report results of analyses of PCI and MAE. We also compare the MAE produced by participants' selection of one of the statistical forecasting models with the MAE obtained by taking the average of the forecasts produced by the selected and non-selected statistical models.

3.2.1. Percentage correctly identified

Overall, mean PCI was 60.31 (*SD* = 14.95), significantly higher than the 50% expected by chance (t (190) = 9.53, $p < .001$). This was true for each of the four conditions.

A two-way factorial between-participants analysis of variance (ANOVA) using noise level (high versus low) and choice of forecast (good/poor versus good/intermediate) as factors revealed only a main effect of choice of forecast (F (1, 187) = 4.35; p = 0.038), such that the PCI for a good model contrasted with a poor model (PCI = 62.53, *SD* = 17.26) was higher than the PCI for a good model contrasted with an intermediate model (PCI = 58.11, *SD* = 11.93). There was no main effect of noise (F (1, 187) = .83; p = 0.364), nor an interaction effect between the two variables (F (1, 187) = .32; p = 0.570).

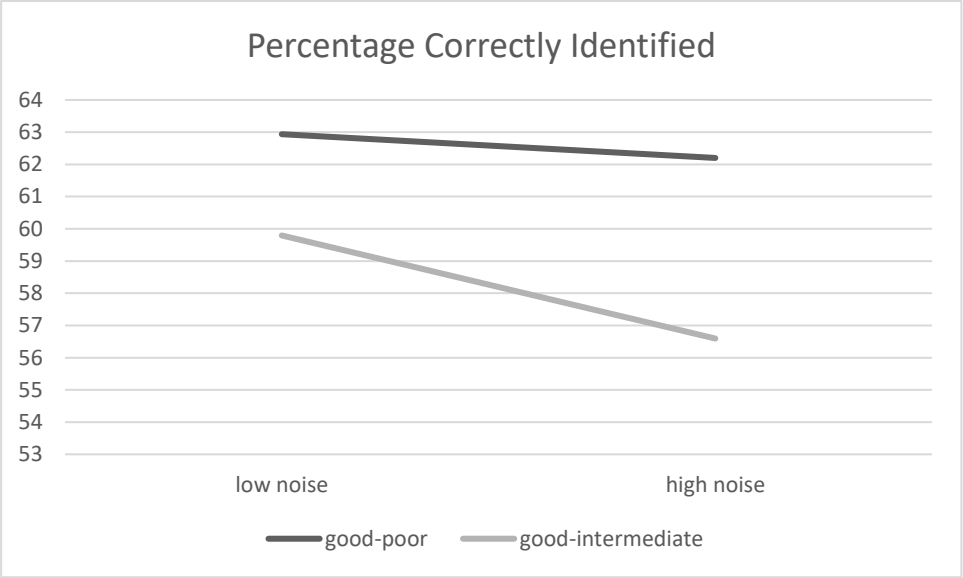

Figure 2. Mean PCI scores in each of the four conditions of the experiment formed by crossing noise condition (low/high) with type of comparison (good versus poor forecasting algorithm versus good versus intermediate quality forecasting algorithm).

3.2.2 Mean absolute error

A two-way factorial between-participants ANOVA using the same factors as before showed a main effect of noise level (F (1, 187) = 25.55; p < .001), such that high noise series had a higher MAE (*M* = 6.08, *SD* = 4.53) than low noise series (*M* = 3.49, *SD* = 2.46). There was also a main effect of choice of forecast (*F* (1, 187) = 56.09, *p* < .001), such that the good model contrasted with the poor model led to a higher MAE (*M* = 6.71, *SD* = 4.67) than the good model contrasted with an intermediate-quality model (*M* = 3.00, *SD* = 3.90). Additionally, there was a marginally significant interaction effect between the two variables (*F* (1, 187) = 4.07, p = .045). Simple effect analysis showed that the statistical difference between low and high noise series was more marked in series where the good model was contrasted with the poor model (*F* (1, 187) = 24.72, *p* < .001) than where the good model was contrasted with the intermediate model (*F* (1, 187) = 4.67, *p* = .032).

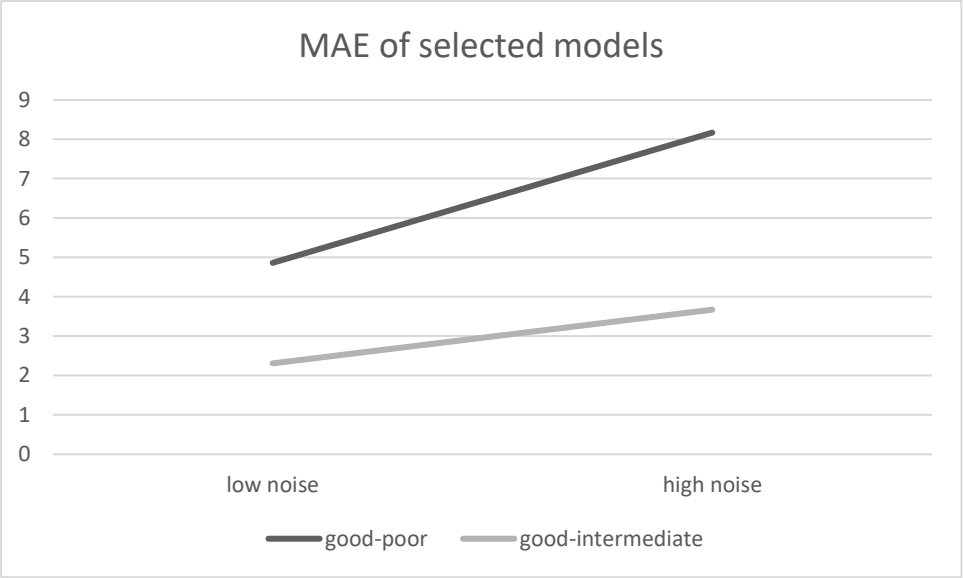

Figure 3. Mean MAE scores in each of the four conditions of the experiment formed by crossing noise condition (low/high) with type of comparison (good versus poor forecasting algorithm versus good versus intermediate quality forecasting algorithm).

3.2.3. Comparison of MAE with that obtained by simple statistical averaging of both models

To test hypothesis H_3, we examined whether the MAE that participants produced by selecting a forecasting approach was different from the MAE obtained by taking the simple average of the results of the two approaches. Results of this analysis are displayed in Table 2. In all conditions, irrespective of noise level and the quality of the forecasts that were compared, people's choices outperformed the simple average. This is consistent with H_3.

.

Noise level	Contrasted forecast qualities	Participant MAE	MAE of averaged models	Df	t	p
Noise: low	Good vs. poor	4.87 (*SD* = 2.89)	7.83 (*SD* = 7.69)	41	-6.66	<.001
	Good vs. intermediate	2.32 (*SD* = 1.06)	3.50 (*SD* = 3.12)	48	-7.89	<.001
Noise: high	Good vs. poor	8.17 (*SD* = 5.29)	12.47 (*SD* = 11.88)	52	-5.93	<.001
	Good vs. intermediate	3.73 (*SD* = 1.36)	4.60 (*SD* = 3.98)	46	-4.40	<.001

Table 2. One sample t-tests for the comparison of model averaging and participant's choice

3.3 Discussion

Forecasters were required to identify the better of two models by comparing the two series of forecasts that those models had made in the past with the series of outcomes that those models had been forecasting. Forecasters were adept at distinguishing the better statistical model, both when people were selecting between good and poor statistical models and when they were selecting between good and intermediate-quality models. This effect was not dependent on the noise level, a finding inconsistent with H_1. However, noise played a significant role in the level of accuracy resulting from the choice of a forecast.

Results were consistent with hypothesis H_2: forecasters were better at identifying the better statistical model when comparing forecasts produced by the good model with those produced by the poor one than when comparing forecasts produced by the good model with those produced by a model of intermediate quality. Their ability to distinguish models was higher when signals were more discriminable.

The analysis also showed that MAE was higher when people were choosing between the good and poor models than when they were choosing between the good and intermediate-quality forecasting models. This implies that forecasting errors produced by incorrect choices in the former case led to a much higher level of MAE than those produced by incorrect choices in the latter case. This means that, although there were fewer errors when people selected between good and poor forecasts, those that did occur were large enough to ensure that the overall MAE score was larger than when they chose between good and intermediate quality forecasts. People were better able to distinguish a good model from a poor model, but a wrong choice was 'punished' more severely in this situation. This finding is remarkably consistent with one reported by Petropoulos et al. (2018). They showed that, though judgmental selection was worse at selecting the best model than algorithmic selection, it was better than algorithmic selection at avoiding the worst models and that, as a result, overall accuracy of judgmental selection was better than that of algorithmic selection.

Finally, results were consistent with hypothesis H_3: the forecasts produced by the model selected by participants outperformed the simple average of the forecasts from which participants made their selection. This result appears to conflict with findings reported by Petropoulos et al. (2018). They compared a) judgmental selection with b) the average of four models and c) the average of the best two of those four models. Mean absolute percentage error (MAPE) scores for forecasts made using these three approaches were 23%, 22%, and 23%, respectively. Thus, in their study, judgmental selection did not outperform the simple average of forecasts from which participants made their selection. There are various possible reasons for the difference between their results and ours. One is that we measured error using MAE whereas they used MAPE. (If we look at the size of the mean percentage error (MPE) scores that they report, the pattern obtained for the above three approaches was more similar to the one that we found: 1.5%, 2.9% and 4.7%.) Another is that the variance in the quality of the forecasts produced by

their models was probably less than that produced by ours and so there would have been less to gain by averaging them.

In summary, the experiment showed that people can make reasonably good *explicit* discriminations to determine which of two models produces the better forecasts and that their ability to do so depends on the factors that detection theory implies are important in this type of task (Stanislaw & Todorov, 1999). In our second experiment, we do not focus on *explicit* discrimination. Instead, we ask whether better forecast models have a greater influence on final forecasts when people combine model-based forecasts with their judgment. In other words, does their behavior indicate that they make good *implicit* assessments of the quality of model-based forecasts?

4. Experiment 2: The influence off statistical models on judgment

In this experiment, we asked whether people who make forecasts by combining statistical models with their judgment make more use of advice from statistical models that produce better forecasts. Participants first made a purely judgmental forecast from the data series. They were then provided with statistical advice from an ideal, intermediate or poor algorithm and made another forecast from the same data series in the presence of this advice. While they did not explicitly state how good they judged the model to be (in contrast to Experiment 1), the size of the shift in the forecast in response to the advice from the statistical model (Harvey and Fischer, 1997) provides a measure of the weight given to the advice and the reduction in forecast error produces an estimate of its beneficial effect. The degree to which people weight information that they receive from different sources is not something to which they have conscious access: the weights that people judge that they use often fail to match the weights that statistical analyses show they actually use (e.g., Harries, Evans, & Dennis, 2000). Whereas Experiment 1 showed that people can explicitly select the better forecasting model, here we ask whether they implicitly put greater weight on better models. Kahneman (2011) and others have argued that implicit and explicit tasks are performed by different cognitive processing systems: if he is correct, there is no reason to suppose that a task (e.g., discriminating between the quality of different forecasting models) performed well by one of these systems will be performed well by the other.

The research on algorithm aversion discussed above implies that people will be reluctant to take the algorithmic advice and, as a result, will not draw as much benefit from it as they could do. They will study the performance of the model in the previous time periods and observe the errors that it makes. This will lead to an undervaluation of the statistical model as predicted by algorithm aversion research (Dietvorst et al., 2015). However, if they can distinguish good from poor statistical advice, it is still the case that they should be more influenced by better advice and that such advice should have a greater beneficial effect. Hence:

Hypothesis 4 (H_4): accuracy of final forecasts will be higher when better quality forecast advice is provided.

Hypothesis 5 (H_5): the shift of in participants' initial judgments towards advice will be greater when the advice is of higher quality.

After their final forecast for each series, participants were informed of a) the forecast error in their final judgment, b) the forecast error in the algorithmic advice they received, and c) whether they could have improved their accuracy by completely following the algorithmic advice. By providing this feedback, we aimed to determine whether forecasters learn across the session to shift their forecasts closer to the good advice that would improve them.

Madhavan & Wiegmann's (2007) argument that people apply 'perfection schema' to algorithms implies that they will continue to trust an algorithm when it does not make errors that could be avoided. In other words, if the forecast errors appear to be produced by random error in the data series rather than by a failure to pick up pattern components in the series, forecasters will increasingly come to assume that the 'perfection schema' is appropriate for the forecast model and, hence, increase their trust in it. This suggests that, given the appropriate feedback, they will use forecasts produced by such a model more when they combine them with their judgmental forecasts. In contrast, if feedback indicates that a forecasting model produces clear errors (e.g., by failing to pick up obvious patterns in the data series), then, as Dietvorst et al (2015) suggest, forecasters will rapidly lose trust in the forecasting model and show evidence of algorithm aversion. They will use the model's forecasts less even though they may still be superior to those produced by judgment alone.

Hypothesis 6 (H_6): only participants in the ideal forecast condition will learn over trials to move towards the advice from the algorithm.

4.1 Method

Each participant saw graphs of time series on their own individual computer screen and made their responses using a mouse click. Dependent variables were the advice-induced shift in the forecast and MAE. Independent variables were noise level in the data series (high versus low) and quality of statistical advice (advice was either of ideal, intermediate or poor quality). Noise level was varied within participants and advice quality was varied between participants. To ensure some generality to our findings, series autocorrelation was again varied.

4.1.1. Participants

One hundred and sixty-one participants took part in the study, of which 79 were female. Their mean age was 25.75 (SD = 7.47). Participants were solicited via Prolific Academic and paid a fixed fee of £1.50 for their participation. In addition, the best performers in each condition were awarded a £ 5.00 Amazon voucher. Participants had not participated previously in similar experiments by the researchers.

4.1.2. Stimulus materials

The series used in this experiment were sampled from the series simulated for Experiment 1. Ten series were selected from each series type (AR 0, AR .8, AR -.8), with half of these being low noise series and the other half being high noise series. For these 30 series, the three forecasting models (naïve forecast, exponential smoothing and AR forecast) described above were used to produce advice.

The graphical presentation of the data series and the statistical forecast series were akin to those used in the first experiment. A graph depicting sales data from week 1 to week 50 was shown at the start of each trial. Once the participant had made an initial judgmental forecast, a series showing forecasts from the statistical model were overlaid on the graph from week 2 to week 51. The data series was labelled 'Sales' and the forecast model was labelled 'Model' (see Figure 4).

4.1.3. Design

Each participant was randomly assigned to one of the three experimental conditions, resulting in the following sample sizes in each condition: good advice, n = 55; intermediate-quality advice, n = 57; poor advice, n = 49. The 30 data series were presented in a different random order for each participant.

4.1.4. Procedure

Participants first saw screens that introduced the experiment, requested their participation in it, and provided them with instructions. On each trial, they initially saw a graph that displayed only the data series. They then made their initial forecast for week 51 by clicking on the graph at the desired sales height. After that, the forecasting model was overlaid on the data graph. They were then given the option of adjusting their initial forecast using the advice provided by the forecasting model. Once they were content with their final forecast, they clicked on 'View Feedback'. This uploaded a display of the expected value of the series for week 51 (i.e., the value produced by the generating equation without the contribution from the error term). They also saw a panel on the right-hand side of the graph with information about their performance, the model's performance, and a comparison between them (Figure 4). They then clicked on a button requesting the next graph. After the thirty graphs were completed, they provided demographic details (age and gender) and were thanked for their participation.

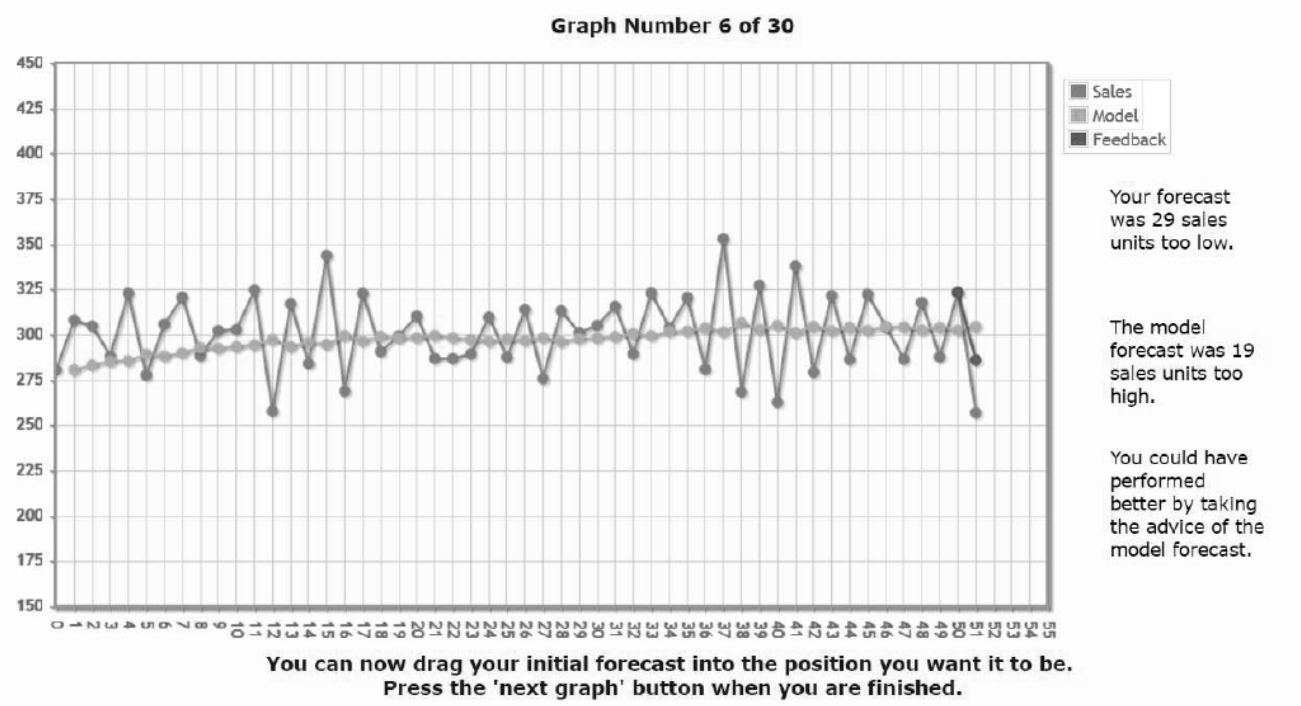

Figure 4. Experimental display after feedback had been requested

4.2. Results

First, we report on the error of the final judgment made by the participants after they received the 'advice' of the model forecasts. Second, we report analyses of how much people changed their initial forecast after they were shown the forecasting model's advice. We use the Shift measure of Harvey and Fischer (1997):

$$SHIFT = 100 * \frac{(Adjusted\ forecast - Initial\ forecast)}{(External\ forecast - Initial\ forecast)}$$

Here the initial forecast is the participant's judgmental forecast before receiving advice, the adjusted forecast is the participant's final forecast after receiving advice, and the external forecast is the forecast provided to participants as advice by the forecasting model. A shift value of 0 indicates that the participant retained their initial judgmental forecast. A shift value of 100 shows that the participant adjusted their initial forecast until it had the same value as the advice given by the model. A shift value of 50 means that the adjusted forecast was halfway between the initial judgmental forecast and the model forecast. Values over 100 indicate 'over-shifting': participants move beyond the model advice – i.e.,

their estimate was no longer in the interval between their initial judgmental forecast and the model forecast[1].

Finally, we report analyses of changes in the level of the MAE across the session in the three different experimental groups.

4.2.1. Mean Absolute Error

As before, MAE scores were calculated relative to the time series signal produced by the generating algorithm (i.e., excluding the contribution from noise term). Against this benchmark, the model forecasts that were used to give advice in the three conditions had a MAE of 20.12 (*SD* = 18.44) for the naïve forecast, a MAE of 8.92 (*SD* = 7.69) for the exponential smoothing forecast, and a MAE of 0 (*SD* = 0) for the autoregressive forecast. These errors are all significantly different from each other at $p < .001$.

Figure 5 shows MAE scores for high and low noise data series when participants were provided with good, intermediate-quality, and poor advice from statistical models. A two-way factorial mixed ANOVA with noise level as a within-participants factor and advice quality as a between-participants factor showed a main effect of advice quality ($F(2, 158) = 52.88$, $p < .001$): mean absolute error in the good advice condition ($M = 7.44$, $SD = 6.46$) was significantly lower than the values in both the intermediate-quality advice condition ($M = 15.63$, $SD = 4.32$; $F(1, 94)$ [2] $= -7.86$, $p < .001$) and the poor advice condition ($M = 16.01$, $SD = 3.21$; $F(1, 81)$ [3] $= -8.71$, $p < .001$). The latter two were not significantly different from one another ($F(1, 104) = .52$, $p = .606$). This provides a partial confirmation of H_4. A main effect of noise level was also obtained ($F(1, 158) = 119.72$, $p < .001$): low noise in data series ($M = 10.96$, $SD = 5.80$) led to significantly lower error than high noise ($M = 14.94$, $SD = 7.50$). This replicates a finding obtained in the first experiment: series with low noise lead to significantly lower error than series with higher noise. There was no interaction effect between condition and the noise level ($F(2, 158) = 2.60$, $p = .078$).

[1] Fifteen participants had a mean shift of 0 across all trials, evenly divided across the three conditions. This could be due to a lack of motivation or a distrust of the advice given. Removal of these participants did not affect the results. Six participants had a mean shift > 100. The reason behind this is unclear. Removal of these participants did not affect the results either.

[2] Levene's test for Equality of Variances was found to be statistically significant, resulting in an adjustment of the degrees of freedom

[3] Levene's test for Equality of Variances was found to be statistically significant, resulting in an adjustment of the degrees of freedom

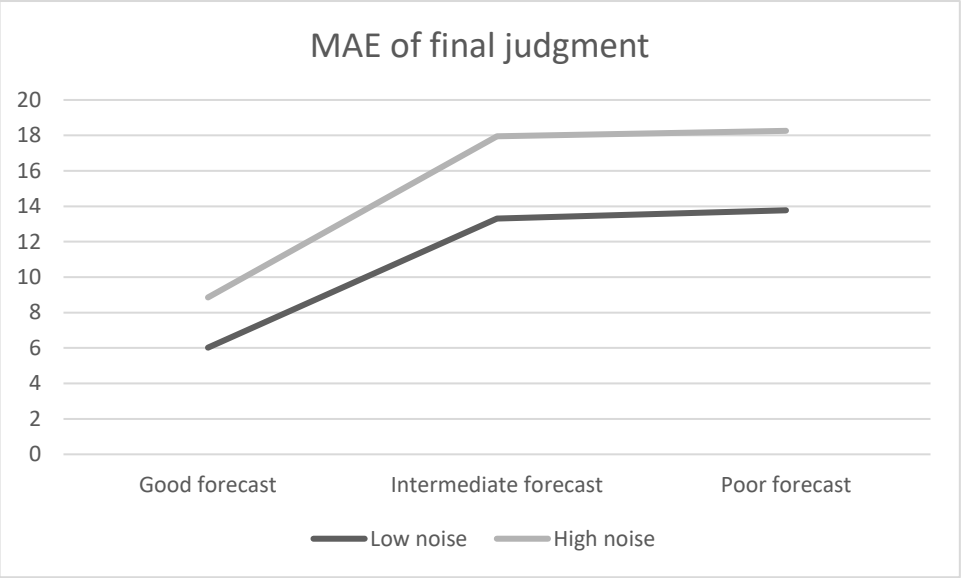

Figure 5. MAE scores for final judgments after seeing statistical forecast advice of different qualities for series with each noise level.

To determine whether the error of the model predicts the error of the final judgment of the participant, the error in the model forecast was calculated on each of the 30 trials. We then used multilevel linear modelling (Gellman & Hill, 2007) to analyse the relationship between trial number, model error and error of the participant's final judgment. As fixed effects, we entered trial number and model MAE into the model. As random effects, we had intercepts for participants. Maximum Likelihood tests were used to obtain p values. Table 3 shows the results of this analysis for the three conditions.

The effect of the model error on the participant's error could not be calculated for the good forecast condition because the model error was always zero and so could not serve as a predictor. However, in both the intermediate forecast condition and in the poor forecast condition, the model error was a significant predictor of the final judgment error ($p < .001$). In both cases, this relationship is positive: higher model error led to higher error in the final judgment. Thus, within the broader categories of 'intermediate' and 'poor' quality, participants were sensitive to the error size and varying quality of the model.

Table 3. Multilevel linear models of the relationship between trial number, model error and error of the participant's final judgment.

		Estimate	SE	df	t	Sig.
Good forecast	Intercept	10.64	.60	1635	17.73	< .001
	Trial	-.20	.03	1635	-6.01	< .001
	Model MAE	NA	NA	NA	NA	NA
	Intercept	12.67	.81	1665	15.67	< .001

Intermediate	Trial	.09	.04	1665	2.42	.016
forecast	Model MAE	.19	.04	1665	4.21	< .001
Poor forecast	Intercept	11.54	.89	1391	12.98	< .001
	Trial	-.01	.04	1391	-.32	.749
	Model MAE	.24	.02	1391	11.63	< .001

4.2.2 Shift scores

Figure 6 and Table 4 show that Shift scores for high and low noise data series when participants were provided with good, intermediate-quality, and poor advice from statistical models. A two-way ANOVA with noise level as a within-participants factor and advice quality as a between-participants factor revealed a main effect of advice quality (F (2,158) = 16.21, $p < .001$). Tukey's post hoc analysis showed that the shift produced by good advice (M = 52.93, SD = 31.08) was significantly greater than both that produced by the intermediate-quality advice (M = 27.92, SD = 32.11) and that produced by poor advice (M = 20.67, SD = 26.40). Shift scores for these latter two conditions were not significantly different from one another. These results provide partial support for H_5. There was no main effect of noise (F (1, 158) = .41, p = .522), nor an interaction effect (F (2, 158) = 1.32, p = .271).

When the noise level was taken into account, for low noise series, people's shift behaviour was not significantly different (t (104) = -.65, p = .258) for the intermediate model (*Shift* = 27.98, SD = 38.50) compared to the poor model (*Shift* = 23.26, SD = 35.35); in high noise series, however, the shift behaviour diverges: a one tailed t-test (t (102.50) = -1.95, p = .027) shows that the intermediate model led to a greater shift (*Shift* = 27.86, SD = 29.32) than the poor forecasting model (*Shift* = 18.09, SD = 22.25) under high noise conditions.

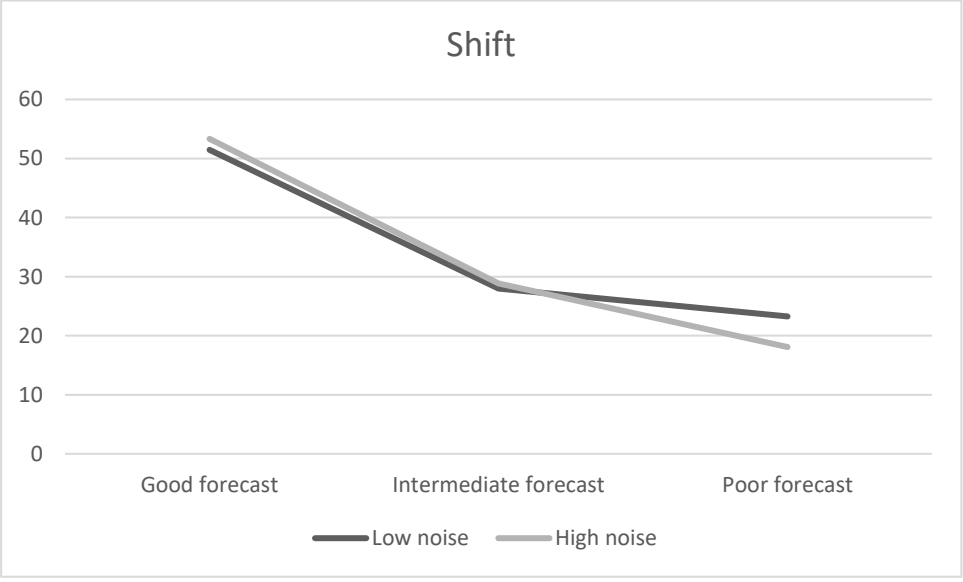

Figure 6. Mean shift scores for each level of noise, and forecast quality.

Table 4. Model MAE, Participant MAE and Shift per condition.

	Good model (AR)	**Intermediate model (Exponential Smoothing)**	**Poor model (Naive forecast)**
MAE of the model	0 (*SD* = 0)	8.92 (*SD* = 7.69)	20.12 (*SD* = 18.44)
MAE of final judgment	7.44 (*SD* = 6.46)	15.63 (*SD* = 4.32)	16.01 (*SD* = 3.21)
Shift	52.39 (*SD* = 31.08)	27.92 (*SD* = 32.11)	20.68 (*SD* = 26.4)

4.2.3 Changes in MAE over the session

The multilevel linear modelling (Table 3) also showed an effect of trial number on final forecast error, showing a change in final judgment error as trials progressed. However, the nature of this effect depended on the quality of the advice they were given. In the good forecast condition, the relationship between trial and final judgment MAE was significant and negative: as the experiment progressed, the participants' MAE scores decreased, a finding consistent with H_6. However, error in the last 10 trials (MAEtrials21-30 = 6.21, *SD* = 11.00), though significantly smaller than that in the first 10 trials (MAEtrials1-10 = 9.97, *SD* = 13.55; t (1045.33) = 5.04, $p < .001$), was still not equal to that of the forecast advice (*MAE* = 0, *SD* = 0) (see Figure 7). This indicates that participants did not follow the algorithm completely even by the end of the experiment.

In the intermediate forecast condition, the relationship between trial and final judgment MAE was significant but positive: accuracy became somewhat worse over the session. In the poor forecast condition, there was no significant relationship between MAE and trial number.

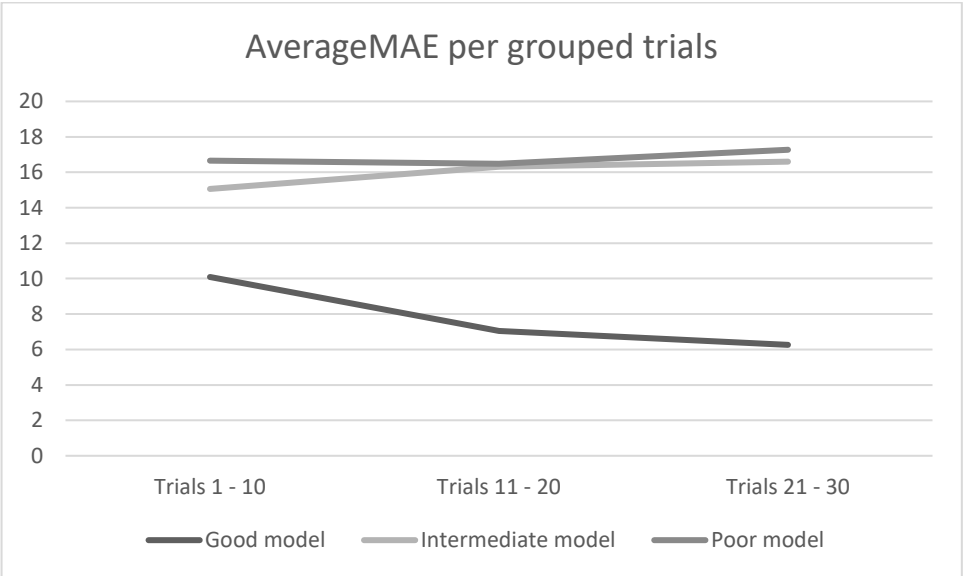

Figure 7. Average MAE per group of ten trials.

4.3 Discussion

We hypothesized that participants would shift more with improved forecast quality. Looking at the mean shifts per condition, we see that the good forecasting model did indeed lead to the highest shift. Thus the experiment provided some evidence that people can discriminate between the quality of different forecasting models implicitly as well as explicitly (Experiment 1). However, the difference in shift between the intermediate model and poor model was not significant. The latter finding is surprising, as the model error of the naïve forecast was significantly higher (more than twice as large) than the model error of the intermediate model.

Participants learned to take more account of forecasts produced by a good model: one interpretation of this is that this model was endowed with the 'perfection schema' posited by Madhaven and Wiegman (2007). In contrast, final forecasts made by participants exposed to forecasts from the intermediate-quality statistical model became somewhat worse over the session. Because the model produced forecasts that clearly failed to pick up some of the pattern elements in the data series (e.g., see Figure 4), forecasters lost confidence in the model and displayed algorithm aversion (Dietvorst et al, 2015). Interestingly, final forecasts of participants advised by the poor model showed no evidence of any change in the way the model's forecasts were used. We suspect that this model was treated more charitably than the intermediate one because it was clear to participants that it was able to pick up the pattern elements in the data series, albeit at a delay of one period.

Although participants learned to improve their forecasts over the session in the good advice group, their performance levelled off: they did not learn to follow the algorithm completely (Figure 7). This may have been because the ideal algorithm still made errors because it could not predict the contribution of the random error term in the generating equation. However, research on advice does indicate that people are conservative in changing their initial assessments: they fail to take enough account of perfect advice (Gardner & Berry, 1995) or good advice (Harvey & Fischer, 1997).

5. General discussion

In forecast support systems, forecasters often have the option of allowing the system to select a forecasting model appropriate to the type of data and context or choosing a suitable model themselves from a drop-down menu (Lawrence et al., 2002). Here we were interested in whether forecasters who decide on the latter approach can distinguish between forecasting models that vary in quality given graphical records of their past performance.

In our first experiment, forecasters were shown records produced by two different models and explicitly asked to select the better one. We found that people were better able to select between good and poor models than between good and intermediate quality models. Despite this, forecast error was higher on average when people selected between good and poor models. This was because incorrect selections had a much more detrimental effect on forecast accuracy than they did when people selected

between good and intermediate quality models. This finding is similar to one reported by Petropoulos et al. (2018).

There is a controversy about whether it is it better to combine or select between sources of advice (Fifić & Gigerenzer, 2014; Soll & Larrick, 2009). We found that mean forecast error associated with selecting between forecast models was lower than that produced by taking the simple average of the forecasts produced by the two models. In contrast, we know that it is difficult to use judgment to combine forecasts in a way that outperforms the simple average of the forecasts (Fischer & Harvey, 1999). It appears, therefore, that, at least when there only two sources of advice, selecting between them is better than combining them.

In our second experiment, we provided further evidence that forecasters can use graphical records of the past performance to distinguish between forecasting models varying in accuracy. In this case, however, they were not asked to make an explicit selection between alternative models simultaneously presented. Instead, each forecaster made an initial judgmental forecast, saw past records of the performance of a single type of model, and then made a final forecast that took account of the advice provided by the model.

We found that people shifted their forecasts more in response to advice from good models than in response to advice from intermediate quality or poor models. Evidence that they shifted their forecasts more in response to advice from an intermediate quality model than in response to advice from a poor model was less strong: the difference attained significance (on a one-tailed test) only when the data series contained high levels of noise. Given these findings, it is not surprising that forecasts based on advice from the good model were more accurate than forecasts based on advice from intermediate quality and poor models but that there was no difference in accuracy of forecasts based on advice from intermediate quality and poor models.

The average shift in the good forecast condition was only 52.39. This means that participants adjusted their initial forecast, on average, only halfway towards the advised forecast. This is despite the that fact that the good forecast was ideal in the sense that it was based on a principle that produced forecasts that, in the long run, could not be improved upon. Furthermore, this was so even though participants received feedback after every graph showing that, overall, the model outperformed their own judgment. These findings are consistent with those reported in the literature on advice-taking: people are reluctant to seek advice and, even when they obtain it, they are reluctant to act on it (e.g., Gardner & Berry, 1995). Furthermore, people are conservative (Phillips & Edwards, 1966): in advice-taking, they are reluctant to change their minds as much as they should do according to objective criteria (Harvey & Fischer, 1997; Yaniv, 2004). Conversely, people shifted their forecasts quite a lot (20%) even in response to very poor advice (Figure 6). This too is an established finding in the advice-taking literature (e.g., Harvey & Fischer, 1997): it is as if people are reluctant to reject help that is offered to them even when it is of little use. Additionally, an anchoring bias may be in operation here. This would

have led people to place too much value on their initial judgment and to adjust too little towards the advice provided to them (Tversky & Kahneman, 1974).

In the second experiment, we investigated not only whether the performance differed between the three groups receiving advice that differed in quality but also whether errors in the model forecasts *within* two of those groups predicted errors in the participants' final forecasts. Multilevel linear modelling showed that they did. Given that we know that forecasters shifted their forecasts in response to the model forecasts in these two groups but only a small amount (< 30%), this means that taking account of advice even to a relatively small extent has significant effects on outcomes.

Throughout the second experiment, forecasters were given feedback about their performance, the performance of the model, and whether their own performance would have benefitted from taking more account of the advice from the model. We anticipated that feedback would help forecasters take account of the advice from the model and, hence, improve the accuracy of their final forecasts. However, feedback does not always have the expected beneficial effects (Harvey, 2011). In fact, we found that people learned across trials but only when the model produced ideal forecasts.

This is consistent with research on algorithm aversion: when models clearly make errors, those using them tend to lose confidence in them (Dietvorst et al, 2015). Here, the intermediate quality model produced forecasts that were clearly incorrect: for example, this exponential smoothing model often fails to pick up patterns that are clearly apparent in the data series (e.g., Figure 4). As a result, there would be no incentive for forecasters to increase the influence of these models on their forecasting beyond the minimum that we observed (i.e., a shift of < 30%). In contrast, the good model always picks up the pattern in the data and errors in its forecasts arise from unpredictable factors. If we assume that forecasters can appreciate these aspects of the model's forecasts, then they will not see deviations between forecasts and data points as errors in forecasting. They will retain confidence in the model and endeavour to rely on it more and more: as a result, they process feedback and doing so improves their performance.

5.1 Limitations

Though we endeavoured to ensure some generality to our findings by varying noise levels and the patterns in the data series, other pattern elements (e.g., trends of various types) could have been introduced as well. It is unlikely that this would radically affect our conclusions. More salient pattern components and a greater number of pattern components are likely to render the task of discriminating between good forecasting models and poor ones easier rather than harder.

We examined only three forecasting models. We chose them because they produced different levels of forecasting accuracy for the data series that we employed. In other words, we selected them because of their performance rather than because they represented specific approaches to forecasting. Nevertheless, exponential smoothing and naïve forecasting are some of the commonest approaches to

forecasting used by practitioners (Weller & Crone, 2012). Hence, though we used few approaches to forecasting, those that we did use are reasonably representative of practice.

Our participants were not practitioner forecasters. However, with pure time-series forecasting, there is little evidence of any difference in accuracy between experienced and non-experienced forecasters (e.g., Lawrence, Edmundson, & O'Connor, 1985; Petropoulos et al., 2018). More generally, the literature contains evidence of inverse expertise effects (e.g., Önkal & Muradoğlu, 1994; Yates, McDaniel, & Brown, 1991). Also, as one might expect, more experienced forecasters tend to take less notice of advice (Harvey & Fischer, 1997). Thus, overall, there is little reason to suppose that practitioners would have performed better when performing the type of tasks studied here.

5.2 Implications and future research

Our findings imply that forecasters can distinguish between forecasting models that vary in quality: they can identify good models and they are more influenced by these models. However, their criteria for discriminating models in terms of quality may not be the same as those of statisticians. We saw that they paid less attention to advice from the exponential smoothing model over the session but did not change how much they used the forecasts from the less accurate naïve forecasting model (Figure 7). We suspect that this was because the exponential smoothing model showed little evidence of picking up pattern elements in the data series whereas the naïve forecasting model did. When evidence of accuracy and evidence of ability to identify pattern components conflict, people may find it difficult to decide which model to prefer. If forecasters using an FSS reject algorithmic selection of the forecasting model to use on the data and insist on choosing the model themselves, their choice may be improved by giving them information about the ranking of available models in terms of the forecast accuracy they produce for the type of data series under consideration. Provision of this type of information is guidance, sometimes called feedforward (Björkman, 1972) and often incorrectly termed 'task properties feedback'. It has proved to be an effective approach to improving accuracy in judgment tasks (Balzer, Doherty, & O'Connor, 1989). Thus, our primary recommendation for the design of effective FSS would be to provide additional information on the proposed models with regards to their accuracy. The amount of information that should be provided is an interesting avenue for future research.

Alternatively, increasing the acceptability of algorithmic selection of the forecasting model would increase forecast accuracy. We know this because Petropoulos et al. (2018) showed that, though people were able to avoid the worst models, they were outperformed by an automatic selection of the best algorithm. If we could find a way of lessening algorithm aversion, forecasters would be more likely to be satisfied with automatic selection of a forecast model. Recently, Dietvorst, Simmons, and Massey (2016) discovered that people are more likely to accept algorithmic forecasts if they were given the opportunity of adjusting those forecasts. Although the adjustment tended to impair the forecast, results were still better than those produced by judgment alone and forecasters were more satisified with them than they were with those produced by an unmodifiable algorithm. This work suggests that one approach

would be to use an algorithm to produce a small set of two or three forecasting models. Then forecasters would be allowed to make the final choice between them. It would be interesting to see if this approach resulted in more accurate forecasts than those produced when forecasters use their judgment alone to select the forecast model. Thus, our second recommendation for the design of FSS is to allow for the creation of multiple forecasts, leading to an ultimate choice by the forecasters themselves. Future research could be designed to investigate the ideal number of produced forecasts, in order to optimize efficiency and accuracy.

Regarding the feedback in Experiment 2, we chose to use outcome feedback, as this is the closest approximation to reality. It was effective in increasing the influence of the formal model only when that model was optimal and, even then, the effect was only partial. In fact, it is unusual for outcome or performance feedback to have any effect in judgment tasks (Harvey, 2011). Lawrence et al. (2006, p 507) point out that one reason for its general lack of effectiveness "is probably because the most recent outcome contains noise and hence it is difficult for the forecaster to distinguish the error arising from a systematic deficiency in their judgement from error caused by random factors". Hence, it is possible that feedback provision would be more effective if it were given in summary form over a number of forecasts rather than after every forecast because this would reduce the degree to which feedback is influenced by error. The effectiveness of summary feedback has been adequately researched only within the context of skilled behaviour. For example, Schmidt, Young, Swinnen, and Shapiro (1989) varied the frequency of provision of summary feedback. They found that, during acquisition (i.e., during learning when feedback was provided), giving summary feedback more frequently (e.g., after every five rather than after every 15 attempts at the task) was more effective. However, during retention trials (i.e., after feedback had been withdrawn), they obtained the opposite effect and found that this increased over time.

To explain such findings, we should recognise that feedback has an effect both as an incentive and as a facilitator of learning (Annett, 1969). People are incentivized to put more effort into a task when they know that they will find out how well they have performed: hence, the more often they find this out, the more they work at the task. This explains the effect that Schmidt et al. (1989) obtained during acquisition. However, learning is greater when people receive information that provides a longer term summary, possibly because of the reduction in the influence of random error (Lawrence et al, 2006). This effect can be observed only when people have to make use of their learning after feedback (and its incentivising effects) have been withdrawn. This explains the effect that Schmidt et al. (1989) obtained during retention. Hence, our third recommendation for the design of FSS is to provide summary feedback and to tailor the frequency of such feedback to how the system will be used. If users are given a separate training period before they operate it, less frequent summary feedback during that training would be appropriate. If they are not, more frequent summary feedback could produce better results. Further research is needed to determine whether the findings from the skilled behaviour literature carry over to use of FSS in forecasting and, if they do, to assess the most appropriate frequency of summary feedback.

6. Acknowledgements

This work was supported by a personal grant from the FWO Research Foundation Flanders to the first author.

7. Reference list

Annett, J. A. (1969). *Feedback and human behaviour*. Harmondsworth, UK: Penguin Books.

Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). The effects of cognitive feedback on performance. *Psychological Bulletin, 106,* 410-433.

Bendoly, E. (2016). Fit, bias, and enacted sensemaking in data visualization: frameworks for continuous development in operations and supply chain management analytics. *Journal of Business Logistics, 37*(1), 6-17.

Bendoly, E., Croson, R., Goncalves, P., & Schultz, K. (2010). Bodies of knowledge for research in Behavioral Operations. *Production and operations management, 19*(4), 434-452.

Björkman, M. (1972). Feedforward and feedback as determiners of knowledge and policy: Notes on a neglected issue. *Scandinavian Journal of Psychology, 13*, 152-158.

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior & Human Decision Processes, 101*(2), 127-151.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114-126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155-1170.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79-94.

Feng, B., & MacGeorge, E. L. (2010). The influences of message and source factors on advice outcomes *Communication Research, 37*(4), 553-575.

Fifić, M., & Gigerenzer, G. (2014). Are two interviewers better than one? *Journal of Business Research, 67*, 1771-1779.

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

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*(1), 3 - 23.

Fildes, R., & Petropoulos, F. (2015). Improving Forecast Quality in Practice. *Foresight: The International Journal of Applied Forecasting, Winter,* 5 - 12.

Fischer, I., & Harvey, N. (1999). Combining forecasts: what information do judges need to outperform the simple average*? International Journal of Forecasting, 15*, 227 - 246.

Gardner, P. H., & Berry, D. C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9*(7), S55-S79.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: an integrated model. *MIS Quarterly, 27*(1), 51-90.

Gellman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

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.

Grove, W. (2005). Clinical versus statistical prediction: The contribution of Paul E. Meehl. *Journal of Clinical Psychology, 61*(10), 1233-1243.

Grove, W. M., Zald, D. H., Hallberg, A. M., Lebow, B., Snitz, E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19-30.

Han, W., Wang, X., Petropoulos, F., & Wang, J. (2018). Brain imaging and forecasting: Insights from judgmental model selection. *Omega, 87*(1), 1-9.

Harries, C., Evans, J. S. B., & Dennis, I. (2000). Measuring doctors' self-insight into their treatment decisions. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 14*(5), 455-477.

Harvey, N. (2011). Learning judgment and decision making from feedback. In M. K. Dhami, A. Schlottmann, & M. R. Waldmann (Eds.), *Judgment and decision making as a skill: Learning, development, and evolution* (pp. 406-464). Cambridge, UK: Cambridge University Press.

Harvey, N. (2019). Commentary: Algorithmic aversion and judgmental wisdom. *Foresight: The International Journal of Applied Forecasting, 54*, 13-14.

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

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

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434.

Hovland, C. I., Janis, I. L., & Kelly, H. H. (1953). *Communication and persuasion; psychological studies of opinion change.* New Haven, CT, US: Yale University Press.

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

Keeffe, B., Subramanian, U., Tierney, W. M., Udris, E., Willemans, J., McDonell, M., & Fihn, S. (2005). Provider response to computer-based care suggestions for chronic heart failure. *Medical Care, 43*(5), 461 - 465.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. *Advances in psychology, 54*, 115-162.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480 - 498.

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., Goodwin, P., & Fildes, R. (2002). Influence of user participation on DSS use and decision accuracy. *Omega: the International Journal of Management Science, 30,* 381-392.

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

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

Leitner, J., & Wildburger, U. L. (2011). Experiments on forecasting behavior with several sources of information: A review of the literature. *European Journal of Operational Research, 213*(3), 459 - 469.

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

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting, 16,* 451-467.

Meehl, P. E. (1954). *Clinical vs statistical prediction*. Brattleboro, VT: Echo Point Books and Media.

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

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

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

Paravastu, N., Gefen, D., & Creason, S. B. (2014). Human trust in other humans, automation, robots, and cognitive agents: Neural correlates and design implications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58*(1), 340-344.

Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research, 274*(2), 574-600.

Petropoulos, F. (2019). Judgmental model selection. *Foresight: The International Journal of Applied Forecasting, 54,* Summer 2019, 4-10.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management, 60,* 34-46.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology: General, 72*(3), 346-354.

Prahl, A., & Van Swol, L. M. (2017). Towards an understanding of algorithm aversion: Why do decision-makers discount advice from automation. *Journal of Forecasting, 36*(6), 691-702.

Sanders, N. R. (1997). The impact of task properties feedback on time series judgmental forecasting tasks. *Omega, International Journal of Management Science, 25*(2), 135 - 144.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(2), 352-359.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton, Mifflin and Company.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior & Human Decision Processes, 84*(2), 288-307.

Soll, J. B., & Larrick, J. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 780-805.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavioral Research Methods, Instruments and Computers, 31,* 137-149.

Theocharis, Z., Smith, L. A., & Harvey, N. (2019). The influence of graphical format on judgmental forecasting accuracy: Lines versus points. *Futures & Foresight Science, 1*(1), e7.

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

Weinstein, N. D., & Klein, W. M. (1996). Unrealistic optimism: Present and future. *Journal of Social and Clinical Psychology, 15*(1), 1-8.

Weller, M., & Crone, S. (2012). Supply chain forecasting: Best practices & benchmarking study. *Lancaster Centre for Forecasting Technical report.*

White, M. J. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34,* 341-382.

Willemain, T. R. (1989). Graphical adjustment of statistical forecasts. *International Journal of Forecasting, 5,* 179 - 185.

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

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science, 13,* 75 - 78.

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.