The impact of judgment on statistical forecasts.

Table of contents

Table of contents ……………………………………………………………………………………………i

List of figures ……………………………………………………………………………………………vi

List of tables ……………………………………………………………………………………………vii

Abstract ……………………………………………………………………………………………………..

Acknowledgements ………………………………………………………………………………………

Dedication ………………………………………………………………………………………………….

Chapter 1. Introduction……………………………………………………………………………1

1.1 Background to research……………………………………………………………….1

1.2 Aim…………………………………………………………………………………………7

1.3 Objectives…………………………………………………………………………………7

1.4 Expected contribution………………………………………………………………7

1.4.1 Academic……………………………………………………………………7

1.4.2 Practical………………………………………………………………………8

1.5 Thesis structure……………………………………………………………………………9

Chapter 2. Literature review……………………………………………………………………13

2.1 Introduction……………………………………………………………………………13

2.2. Supply Chain Management…………………………………………………………..15

2.3 Supply Chain Forecasting……………………………………………………………17

2.3.1. Forecasting methods……………………………………………………19

2.3.2. Quantitative methods…………………………………………………………..19

2.3.3. Qualitative methods……………………………………………………………..23

2.4. The use of judgment in industry…………………………………………………………25

2.5. Spare parts in industry…………………………………………………………………….32

2.6. Combination of forecasts………………………………………………………………..35

2.7. Judgment and causal influences………………………………………………………..39

2.8. The importance of forecasting horizon……………………………………………..41

2.9. Types of knowledge………………………………………………………………………..44

2.10. Judgment and added opinion…………………………………………………………47

2.11. Forecasting and decision support systems……………………………………50

2.12. Feedback…………………………………………………………………………………………54

2.13. Experts and judgment……………………………………………………………………..57

2.14. Character of judgment adjustment……………………………………………………61

2.15. Judgmental bias………………………………………………………………………………..62

2.16. Judgment and spare parts inventory……………………………………………….65

2.17. Conclusion……………………………………………………………………………………….68

Chapter 3. Methodology ……………………………………………………………………………..72

3.1. Introduction………………………………………………………………………………………72

3.2. Philosophy…………………………………………………………………………………………..72

3.3. Approach……………………………………………………………………………………75

3.4. Research design………………………………………………………………………….77

3.5. Data collection and analysis method……………………………………………….83

3.6. Ethics, Reliability, Validity, Generalizability and Limitations………………..90

3.6.1. Ethics……………………………………………………………………………….90

3.6.2. Reliability…………………………………………………………………………91

3.6.3. Validity and Generalizability……………………………………………….92

3.6.4. Limitations………………………………………………………………………93

Chapter 4. Data findings and analysis……………………………………………………..95

4.1. Introduction……………………………………………………………………………….95

4.2. What effect did expertise have on the accuracy of forecasts?............97

4.3. How did accuracy vary between SKU groups?..................................105

4.4. Does forecast accuracy change with horizon?..................................120

4.5. What was the direction of the forecast adjustment?.........................121

4.6. What was the impact on accuracy from the size and direction on the judgmental adjustments?...129

4.7. Impact of adjustment on inventory……………………………………………….139

4.8. Conclusion………………………………………………………………………………….155

5. Discussion…………………………………………………………………………………………156

5.1. What effect did expertise have on the accuracy of forecasts?...........156

5.2. How do time series types effect judgmental adjustments?...............158

5.3. Does extending the horizon influence the judgmental adjustment in a positive or negative way?..166

5.4. How does judgmental adjustment size and direction effect accuracy of forecast? ……………………………………………………………………………………………168

5.4.1. Forecast adjustment direction………………………………………….168

5.4.2. Forecast adjustments made in relation to accuracy……………170

5.4.3. Final adjustments made in relation to the number of forecasts ………………………………………………………………………………………….172

5.4.4 Final adjustments made in relation to accuracy…………………..173

5.5. The impact of adjustment on inventory……………………………………….182

5.6. Conclusion…………………………………………………………………………………187

6. Limitations………………………………………………………………………………………….188

7. Further Research……………………………………………………………………………….190

7.1 Origin of statistical forecast……………………………………………………….190

7.2 Forecast techniques……………………………………………………………………190

7.3 Combination of forecasts…………………………………………………………...191

7.4 Spare parts……………………………………………………………………………….…192

7.5 Expertise / Causal influences / Opinion …………………………………………193

7.6 The use of FSS …………………………………………………………………………….194

7.7 Forecast Horizon…………………………………………………………………………195

7.8 Inventory……………………………………………………………………………………195

7.9 General discussion points………………………………………………………….……….196

8. Concluding remarks……………………………………………………………………….…197

9. References……………………………………………………………………………….……200

List of figures

Figure 2.1. Outline of literature review ……………………………………………………………13

Figure 2.2. Transforms Four Quadrants Using Product Portfolio Management Principles …………………………………………………………………………………………….….25

Figure 3.1. The research process onion ……………………………………………………………77

Figure 3.2. Experimental design table – Single participant single month task …….82

Figure 3.3. Experiment spreadsheet headings …………………………………………………84

Figure 3.4. Experiment spreadsheet headings (continued) ………………………………..86

Figure 3.5. Experiment spreadsheet headings (continued) ……………………………….…87

Figure 3.6. Experiment spreadsheet headings (continued) ………………………………..88

Figure 3.7. Mean Absolute Demand: Mean equation ………………………………………89

List of tables

Table 2.3. Survey results from Sanders and Manrodt (1994) showing the percentage of managers that reported using techniques for different time horizons. The number in parenthesis are those reported by Mentzer and Cox (1984) ……………27

Table 2.4. The table shows the percentage of managers reporting satisfaction with forecasting techniques Sanders and Manrodt (1994). The numbers in parenthesis show results by Mentzer and Coz (1984) ……………………………………………………………28

Table 2.5 Frequency of forecasting methods use. Values are numbers of firms. Data in parenthesis are percentages ……………………………………………………………………29

Table 2.6. Familiarity with forecasting techniques ……………………………………………43

Table 4.1. Forecast accuracy by participant across all groups ……………………………98

Table 4.2. Mark forecast accuracy by SKU group ……………………………………………99

Table 4.3. Dominic forecast accuracy by SKU group …………………………………………99

Table 4.4. Michelle forecast accuracy by SKU group ………………………………………100

Table 4.5. Dave forecast accuracy by SKU group ……………………………………………101

Table 4.6. Jane forecast accuracy by SKU group ……………………………………………101

Table 4.7. Professor forecast accuracy by SKU group ………………………………………102

Table 4.8. Statistic forecast accuracy by SKU group ………………………………………103

Table 4.9. SKU group error ranking ……………………………………………………………….103

Table 4.10. Comparison of average MAE/MD for participant and statistic ……….104

Table 4.11. Participants accuracy by SKU group (ranked by smallest MAE/MD) .106

Table 4.12. Accuracy by SKU group ranked ………………………………………………….107

Table 4.13. Coefficient of variation per SKU group ……………………………………….107

Table 4.14. Participant performance for High Frequency – Low Value …………….108

Table 4.15. Participant performance for High Frequency – High Value …………….109

Table 4.16. Participant performance for Low Frequency – High Value …………….110

Table 4.17. Participant performance for Lower Frequency – Lowest Value ……..110

Table 4.18. Participant performance for Very Low Frequency ……………………….111

Table 4.19. Participant performance for New Items ……………………………………..111

Table 4.20. Participant performance for Longer Lead-time ……………………………112

Table 4.21. Participant performance for Increasing Demand …………………………113

Table 4.22. Participant performance for Decreasing Demand ………………………..113

Table 4.23. Participants error ranking …………………………………………………………..114

Table 4.24. The impact of the final adjustment ……………………………………………..115

Table 4.25. Impact of final judgment on high frequency – low value ………………115

Table 4.26. Impact of final judgment on high frequency – high value ……………..116

Table 4.27. Impact of final judgment on lower frequency – high value …………….116

Table 4.28. Impact of final judgment on lower frequency – lowest value …………117

Table 4.29. Impact of final judgment on very low frequency ……………………………117

Table 4.30. Impact of final judgment on new items ………………………………………118

Table 4.31. Impact of final judgment on longer lead-time ……………………………..118

Table 4.32. Impact of final judgment on longer lead-time ………………………………119

Table 4.33. Impact of final judgment on decreasing demand …………………………119

Table 4.34. Forecast accuracy by horizon ……………………………………………………120

Table 4.35. Total adjustment direction …………………………………………………………121

Table 4.36. Direction of adjustments by SKU group ………………………………………122

Table 4.37. Adjustment totals by participant ………………………………………………..123

Table 4.38. Dominic forecast adjustment direction ………………………………………124

Table 4.39. Mark forecast adjustment direction ……………………………………………124

Table 4.40. Dave forecast adjustment direction ……………………………………………125

Table 4.41. Jane forecast adjustment direction ……………………………….………………125

Table 4.42. Professor forecast adjustment direction ………………………………….…126

Table 4.43. Michelle forecast adjustment direction ………………………………………127

Table 4.44. Total final adjustment direction …………………………………………………128

Table 4.45. Total % of SKU final adjustments by group …………………………………129

Table 4.46. Dominic % change to forecast and impact on accuracy …………………130

Table 4.47. Mark % change to forecast and impact on accuracy ………………………131

Table 4.48. Dave % change to forecast and impact on accuracy ………………………131

Table 4.49. Jane % change to forecast and impact on accuracy ……………………..…132

Table 4.50. Michelle % change to forecast and impact on accuracy …………………133

Table 4.51. Professor % change to forecast and impact on accuracy ………………133

Table 4.52. Aggregate % change to forecast and impact on accuracy ………………134

Table 4.53. Dominic Impact of final adjustment ……………………………………………135

Table 4.54. Mark Impact of final adjustment …………………………………………………136

Table 4.55. Dave Impact of final adjustment …………………………………………………136

Table 4.56. Jane Impact of final adjustment …………………………………………….……137

Table 4.57. Michelle impact of final adjustment ……………………………………………138

Table 4.58. Aggregated impact of final adjustment ………………………………….……138

Table 4.59. Stock values at the start and at the end of the experiment …………..140

Table 4.60. Dominic inventory impact to target stock by SKU group ………..………141

Table 4.61. Mark inventory impact to target stock by SKU group ……………………142

Table 4.62. Dave inventory impact to stock by SKU group ……………………….……143

Table 4.63. Jane inventory impact to stock by SKU group …………………………..…144

Table 4.64. Professor inventory impact to stock by SKU group ………………..……145

Table 4.65. Michelle inventory impact to stock by SKU group …………………..……146

Table 4.66 SKU group ranking by positive inventory impact versus target ……….147

Table 4.67. Pareto graph of SKU group inventory targets ………………..……………148

Table 4.68. Participant impact to target stock High frequency – Low Value …….149

Table 4.69. Participant impact to target stock High frequency – High value ……149

Table 4.70. Participant impact to target stock Lower frequency – Highest value

………150

Table 4.71. Participant impact to target stock Lower frequency – Lowest value 151

Table 4.72. Participant impact to target stock Very low frequency ………………….151

Table 4.73 Participant impact to target stock New Items ……………………………….152

Table 4.74. Participant impact to target stock Longer lead-time …………………….153

Table 4.75. Participant impact to target stock Increasing demand ……..…………153

Table 4.76. Participant impact to target stock Decreasing demand ………………..154

Table 4.77. Participant ranking by positive inventory impact versus target ….….154

Table 5.1. Mark Variance from statistic by SKU group …………………..………………161

Table 5.2. Dominic Variance from statistic by SKU group ……………………………….161

Table 5.3. Michelle Variance from statistic by SKU group ……………………………….162

Table 5.4. Jane Variance from statistic by SKU group ……………………………..……….163

Table 5.5. Dave Variance from statistic by SKU group ……………………………………..163

Table 5.6. Professor Variance from statistic by SKU group …………………………….164

Table 5.7. Average variance from statistic by SKU group ……………………………….165

Table 5.8. % adjustment range by participant (forecast adjustment) ………………171

Table 5.9. % adjustment range by participant (final adjustment) ………..…………176

Table 5.10. Average impact of final judgment by participant ………………………..178

Table 5.11. SKU groups plotted against contextual characteristics and effect on forecast ……………………………………………………………………………………………………179

Table 5.12. % direction of adjustment by SKU group ……………………………………..…181

Abstract

The thesis will aim to empirically assess how judgment impacts a statistical forecast in a spare parts supply chain.

The thesis will investigate what impact judgment has on forecast accuracy that is, does it improve the statistical forecast. If so, is there specific types of demand series, spare parts types or expertise which can affect the accuracy improvement. The results will be used to provide a matrix showing where judgment should or should not be applied to a statistical forecast with regards to accuracy improvement.

The size and direction of the judgmental adjustment will be scrutinised to explore where any correlation can be found to accuracy improvement.

The experiment will be for a 12 months longitudinal period using forecast experts who are working in a company and are forecasting the same spare parts on a day to day basis. The statistical forecast used will be the method that the company uses on a day to day basis.

In order to benchmark the performance of the experts a senior academic will also be forecasting the spare parts involved over the same period in order to show another comparison but with a more considered, complex statistical forecast rather than the relatively simple average based statistical forecast the company used.

Insights into further research, limitations of the experiment and a conclusion stating the impact to academic knowledge and possible practitioner usage will be discussed.

Acknowledgements

The idea and desire to start the PHD came from Prof. Aris Syntetos who was my initial supervisor – thanks for the inspiration and your involvement in the experiment itself. The person who hauled me to the finishing line was Dr Yiannis Polychronakis – thanks for your motivation and structure which was well received and required!

My family must be thanked for their support and the many evenings without my help when I was in the library. Leaving Sajida to deal with Luqa, my son and Frida, my daughter, on her own (they were both much younger when I started the PHD!).

In order to run the experiment, I needed the participation for 12 months of the forecasting team dealing with spare parts at the company where the experiment was held. My sincere thanks to Dave, Dominic, Jane and Michelle for your patience in waiting for the outcome.

Dedication

I dedicate this thesis to my children Luqa and Frida as an inspiration to follow their dreams and to complete the tasks they start on, whatever they may be.

Also, to all the forecaster's striving to make incremental improvements to their objectives – keep going it is worth it!

1. Introduction

1.1. Background to research

Forecasting is an area that has been extensively researched by academics and is vital in industry for practitioners. For many functions within companies it is important to have a perspective on the future, to make informed decisions on issues such as sales, inventory management, capacity planning and finance.

This research is undertaken within a large electronic manufacturing company headquarters in a forecasting department with responsibility for forecasting and inventory control for a pan European stockholding of spare parts.

Specifically, forecasting of spare parts provides a very challenging environment as reported by Bacchetti and Saccani (2012). The number of stock keeping units (SKU's) and the varied time series they present alongside the variance in cost, impact of failure rates and lifecycle, means that there are many factors that need to be considered (Syntetos *et al,* 2009). These factors can induce the forecaster to judgmentally adjust forecasts provided.

The methods which are used to provide a forecast vary from pure judgment (involving no quantitative procedure), devising an excel spreadsheet, to reliance on software packages (both stand alone and within Enterprise Resource Systems (ERP)) that can use very complicated algorithms. Forecasts are also obtained by using a combination of judgment and quantitative methods (Boulaksil and Franses, 2009).

The utilisation of judgment is prolific within industry as reported by Sanders and Manrodt, (1994) and it is therefore important to understand what implications this has and how it effects forecast accuracy and inventory targets.

The use of judgment within US corporations (Sanders and Mandrodt, 1994) highlighted how managers relied heavily on judgmental forecasting methods and although awareness of quantitative forecasting had improved, the level of usage had not increased. Similarly, Canadian firms were reported to use judgmental methods predominantly when compared to quantitative, causal and newer methods (Klassen and Flores, 2001).

It has been shown in the literature that where statistical methods and judgment are integrated the accuracy of the forecast can be improved (Armstrong and Collopy, 1998). Here, judgment was perceived to be more effective as an input to the statistical method not as a revision to the forecast. The positive effect of the use of judgment when used in conjunction with a statistical method was also investigated within a laboratory experiment where several combination methods were compared, however, here the most accurate results were from a 'correct not combine' method (Goodwin, 2000). More recently, investigation into whether the characteristics of the time series and the nature of the judgmental adjustments are important to forecast accuracy has also shown conditional positive effects and shown improvement to stock control performance (Syntetos et al, 2008).

Where specific causal influences exist that are random, such as price reductions or product promotions, judgmental forecasters should have an advantage as they can consider these influences when compared to any statistical model. The occurrence of this type of fluctuation is increased by short product lifecycles and ever more aggressive marketing in the current environment. Webby and O'Connor (1996) found, in their investigation into judgmental and statistical time series forecasting, that the major contribution of judgmental approaches lies in the ability to integrate the non-time series information into the forecasts. Additionally, although causal influence can be reflected judgmentally, the results are not optimal and are not significantly better than purely statistical forecasts (Lim and O'Connor, 1996). The inefficient use of causal information in conjunction with a statistical forecast was shown to be improved by regular explanations of the statistical forecast (Goodwin and Fildes, 1999). Furthermore, the use of a hybrid model that integrated judgment and statistical methods was shown to improve forecast accuracy whilst still allowing added value from adjustments (Trapero et al, 2013).

The size and direction of adjustments have been investigated to examine how they affect the accuracy of forecasts. It has been shown that the performance on ascending trends outperformed descending trends (Thomson et al, 2013). Here, where attention was drawn to the direction and strength of the trend via additional questions, performance improved. An investigation (Fildes et al, 2009) into companies that adjusted an initial computerised forecast primarily to consider exceptional circumstances showed that whilst overall judgmental adjustments did improve accuracy there were some noticeable differences. Larger adjustments were seen to improve accuracy where smaller ones were seen to reduce it. Positive adjustments (that is adjustments upwards) were less likely to improve accuracy than negative ones, they were also often in the wrong direction which pointed to an optimism bias.

Improvements to judgmental forecast accuracy have been seen when forecast horizons have been increased and the period furthest away in the horizon is forecast first (Theocharis and Harvey, 2016). This process, sometimes referred to as 'end anchoring', was seen to reduce the noise in forecasts (Harvey, 1995). Harvey also noted that forecasting for a period further away required more resources and took longer to produce.

Some have argued that two types of judgmental forecasts can be distinguished: judgment based on contextual knowledge and technical knowledge. Sanders and Ritzman, (1995) looked at how this knowledge combined with statistical forecasts improve accuracy. Contextual knowledge is defined as information gained through experience of the job or with the data series and products which are forecasted. Technical knowledge on the other hand is defined as information gained from education on forecasting models and data analysis. Sanders and Ritzman concluded that contextual knowledge was more likely to improve forecast accuracy in combination forecasts and that the more variable the data series, the more contextual knowledge was required. If the variability was low, then less importance should be given to the contextual knowledge. They also stated that in very variable series that contextual knowledge alone could give the most accurate forecast.

The benefit of addition opinions is discussed by Yaniv (2004) in the context of judgment and decision making. Yaniv noted that according to empirical results incorporating even a few opinions was beneficial for the accuracy of a forecast whilst reducing the bias. Decision makers however tended to discount dissenting advice and accept that which concurred with their own opinion. A more recent study looking at group forecasting accuracy found that when undistorted forecasts are given, they contribute positively to the forecast accuracy with the results showing a strong tendency to favour optimistic forecasts (Onkal et al, 2011).

The integration of the statistical forecast with judgment within a forecasting support system (FSS) and the decision support system (DSS) that helped the forecaster decide when and when not to adjust was often carried out poorly with impacts to accuracy (Fildes et al, 2006). This study concluded that FSS systems should be acceptable to users, easy to use, offer a flexible range of methods, viable for commercial marketing and would support an appropriate use of judgment and statistical methods. A DSS would be improved if there was faster and regular testing of the different systems employed, and which was fed back to the forecaster. The advice which is given to the forecaster was seen to be accepted where it was longer and more detailed than when it was short and brief (Gonul et al, 2006).

The impact of feedback regarding time series forecasting accuracy was considered to be positive particularly where there are high levels of noise (Sanders, 1997). Feedback using a rolling training approach and reporting accuracy metrics for each period rather than a summary over several periods indicated an improvement in accuracy for forecasters with technical knowledge (Petropoulos et al, 2017) underlining the positive effect of feedback.

To improve judgment accuracy advice can be sought. The influence of advice on judgment accuracy and how forecasters responded has been reported to show that individuals place a greater weight on their own opinion than their advisors; the more knowledgeable the individual, the increased likelihood that the advice was likely to be discounted; furthermore, the weight of the advice decreased the greater the distance from the initial opinion increased and that the use of the advice did improve accuracy (Yaniv, 2004).

In addition, Franses (2004) reported that the opinion of experts aligned with computing power and improved data, would improve forecast accuracy, in a survey of editorial boards when noting that improvements in forecasting over 30 years had been modest. The impact of experts and whether they improve judgments overall was further investigated (Fildes and Goodwin, 2007); in this study, although there was a positive effect, the experts appeared to weight their own opinion too strongly when adjusting the forecast. This effect was also discussed by Franses and Legestree (2010); they too concluded that experts weighted their own opinion too highly and the forecast accuracy would be improved if the effect would be reduced.

When judgmental adjustments are made, forecasters may introduce bias into the statistical forecasts. Eroglu and Croxton (2010), considered three types of possible bias: optimism bias, anchoring bias and overreaction bias. They also explored individual differences: personality, motivation and work locus of control and found that decisions were driven by experience of current position, contextual knowledge, locus of control and challenge-seeking. The results showed that the level of bias depended on the forecaster’s personality and motivational orientation.

There have been many empirical studies into judgmental adjustments using surveys, field retrospective studies, laboratory experiments (some with rewards) undertaken by managers, graduate or undergraduate students but never in a real word environment by managers who are doing the forecast as well as the experiment.

1.2. Aim

To empirically assess the impact of judgmental forecasting in spare parts supply chains and to also understand the effect on inventory management because of the forecasts placed.

1.3. Objectives

Examine the current academic research to address issues in a thorough literature review.

Explore what impact expertise has on the judgmental adjustment accuracy.

By using time series characteristics to group SKU's, explore what effect this has on how statistical forecasts are judgmentally adjusted. Also investigating the impact of any final adjustments made post the initial forecasts.

To investigate whether extending the forecast horizon influences forecast accuracy in a positive or negative way.

To understand how different judgmental adjustment sizes and directions effect forecast accuracy and what inventory implications this has.

To produce a matrix indicating when judgment should be applied when considered against spare part times-series types. Where does adding judgment make the most improvement?

1.4. Expected contribution

1.4.1. Academic

The research will be over a 12-month period studying the judgmental accuracy of the judgmental forecast. This was highlighted as a required addition to the current literature by Nikolopoulos et al, (2005).

Syntetos et al, (2008) called for more investigation on the implications of other variables, specifically inventory after judgmental adjustment which will be addressed in this research.

There have been many laboratory experiments related to judgmental adjustments but few in a real-life setting. This was highlighted by Franses (2013) as an area for further research.

A comparison of how judgmental adjustments compare when done over short and long horizons was an area where more understanding was required as reported by Theocharis and Harvey (2016).

1.4.2. Practical

Insights into how judgmental adjustments should be incorporated with a statistical forecast.

Are judgmental adjustments more accurate with specific types of time series?

What are the benefits of expertise with respect to judgmental adjustment?

What are the implications for inventory management of judgmental adjustments for different time series and different levels of expertise?

Methodology

The research uses a quantitative strategy and a positivist philosophy. The experiment will measure the impact of alterations to statistical forecasts and the effect this has on forecast accuracy and inventory. There will be 6 participants who will forecast 90 SKU’s monthly over a 12-month period. The experiment is situated in a real-world environment using inventory planners in real time.

1.5. Thesis structure

- Introduction
    - An explanation of why the subject of this report is important.
    - A review of who has worked on this subject highlighting areas of interest to this research.
    - A description of the gaps in the literature and how this research will add insight.
    - What this research will aim to do.
- Literature review
    - Address the academic issues in the literature
- Methodology
    - What approach the research will take?
    - Why this approach was chosen.
- Data findings and analysis
    - How the analysis was conducted.
    - How the observations were collated.
    - What the analysis suggests.
- Conclusion and further research recommendations
    - What research aims were achieved?

- What areas of further research are necessary?
- How does the conclusion effect practice?

Humans intervene often in statistically derived results. Within the realm of spare parts management this then has impacts on forecasting accuracy and inventory control. This PHD will utilise a unique opportunity to provide insight into the area using managers in the real word with live data for 12 months.

Spare parts demand is typically variable. Whilst there are typically some SKU's that may be relatively constant in demand which can often correlate with the number of products sold making forecasts easier to produce there are usually a body of SKU's (often the majority) that are sporadic, intermittent or extremely lumpy in their demand characteristics which make forecasts very difficult and induce managers to use judgment. This is the case across many differing industries where spare parts are forecast not just electronics.

It is this very nature of the demand series that induces a judgmental intervention. Either the forecaster has no statistical forecast, or the statistical forecasts provided are obviously not optimal. For example, many statistical forecasts within Enterprise Resource Systems (in this case SAP R3) do not have suitable algorithms to deal with intermittent demand (where a demand series has zero demand periods) or where there are "one shot" demands (high quantities of demand amongst otherwise very low quantities). When an average type method (moving average, weighted average or exponential average) is being used to derive the statistical forecast the intermittence or the "one shot" characteristics can have large negative implications for the statistical forecast produced. When the forecaster examines the figure, they see the incongruity and subsequently adjust the statistical forecast accordingly as it is so clearly not what they expected (in their judgment).

From a theoretical viewpoint, the amount of insight provided by researchers has shed light on many aspects of judgmental forecasting and its effect on a forecast. However, one of the gaps in the literature is that all the studies and experiments (to my knowledge) have not been "live" where forecasters are performing their day to day functions and where their forecasts are real – world producing purchase orders and the studies have not been longitudinal (observed over a long period, in this case 12 months) allowing a deeper insight into the impact of judgment across a broad range of SKU's. The time duration of the experiment is also important as the instances of outlier occurrences and the subsequent forecast impacts can be tracked and reported over the full experimental period for a more meaningful understanding of the impact of judgment to a statistical forecast and the impact on inventory control.

The broader application of this research is to try and indicate where (from a forecast accuracy and inventory control perspective) judgmental adjustments can be most useful. The time series types chosen are not specific to spare parts although as explained earlier slower moving and more intermittent series are probably more prevalent in spare parts demand series. If judgmental forecasting can be targeted where the outcomes are optimal then the resource and time required can be reduced. Many small and medium businesses (SMB's) do not have an abundance of resources and larger companies would benefit by being able to target resource more efficiently.

In conclusion the PHD aims to utilise the opportunity of access to the participants (which included the PhD candidate Mark) in the real world (highlighted by Bunn and Wright, 1991) doing exactly their normal processes and examining their judgmental adjustments from a forecast accuracy and inventory performance perspective to show insight into the forecast accuracy of judgmental adjustments to statistical forecasts.

2 Literature review

2.1. Introduction

In the review of the literature I will critically evaluate previous research in the subject area of the PHD. The literature will help identify theories and ideas some of which will be tested using the data generated by the PHD experiment.

The review is set out as per the diagram below:

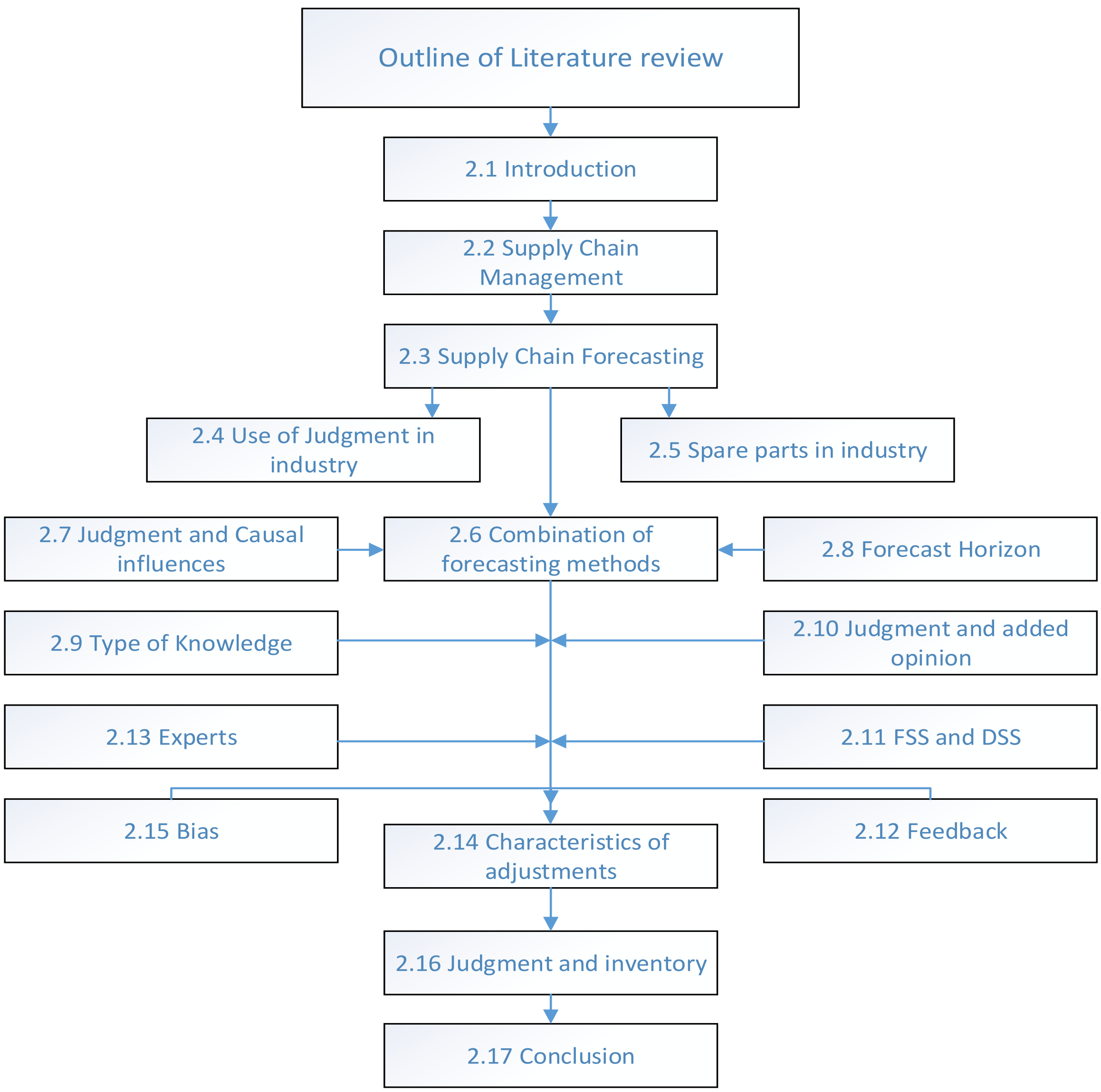


Figure 2.1. Outline of literature review

The chapter starts by reviewing supply chain management in general which overarches the topic of the PHD. Forecasting is also reviewed broadly looking at most aspects of the methods used within the area. To highlight how widespread judgment as a method is used, I have reviewed studies that look across many companies and industry sectors.

As the PHD is focused on an experiment in a spare parts environment the role of spare parts in industry and its monetary and supply chain importance is reviewed.

Forecasts can be produced using many methods and these methods can be combined to produce a combined forecast sometimes using judgment. From this point where a forecast is produced it is interesting to investigate what influences may have been involved in the production (causal) and to understand what other variables could have an influence on the forecast (horizon).

Once the forecast has been produced then there is the question whether it should be adjusted? The following section in the chapter review what can influence the character of an adjustment if it were to take place?

Whether an adjustment is made or not there will be an effect on the inventory level for the article forecast. In this case stock keeping units (SKU'S).

The review will end with a conclusion summarising the chapter and linking forward to the further chapters of the PHD.

2.2. Supply chain management

Logistics can be described as a plan to allow a product and information flow through a business. Supply chain management (SCM) is a wider concept than logistics. Christopher (1992) explains that it builds upon the framework of logistics and tries to co-ordinate between the processes of other areas i.e. the organization itself, customers and suppliers. An example of an aim of SCM could be to reduce the level of buffer stock that exists in a chain of organizations through the sharing of information regarding stock levels and demand. Christopher (1992) defines SCM as “the management of upstream and downstream relationships with suppliers and customers to deliver superior customer value at less cost to the supply chain as a whole”.

Companies should aim to maximize their competitive advantage to differentiate itself from other competitors and maximize profit by operating at a lower cost. To do this an organization needs to look at itself in a holistic way. The many different activities that an organization performs such as procurement, design, manufacturing, distribution, marketing, and after sales service are all linked and are part of the supply chain management task (Porter, 1985). Spare parts are found throughout the supply chain from purchasing to manufacture through to after sales service. This can also be said of customer service which is also multi-dimensional.

Customer services can be separated into three headings: pre-transaction elements, transaction elements and post transaction elements (LaLonde and Zinszer, 1976).

The pre-transaction elements relate to a company’s policies e.g. how good is a company’s customer service policy? Is it understood within the organization and is it easily communicated outside the organization too? What is the adequacy of service processes for example, can a complaint be escalated to management easily and without error? Another tenant of pre-transaction service can also be whether there is flexibility within the system, for example can a deliver be sent by air instead of road?

The availability, fill rate and order cycle time are all transactional elements. Also, status information may be important. Does the customer receive an e-mail or track and trace information regarding delivery?

Post transactional elements include spares availability. Does the company have a high in stock level of spare parts? If so, what is the service level offered and are there ranges of service level agreements that a customer can choose from. If the spare part is to be part of a repair how long does the service engineer take to arrive? Many companies offer warranties on their products so how long is the warranty and it is included in the retail price or as an extra?

The availability of products is an important part in the supply chain for any manufacturing company. A study by Corsten and Gruen (2004) investigated the cost incurred when a stock out occurs. They found that over a quarter of people would buy a different brand and that nearly 40% would look elsewhere for a similar product. Thus, highlighting the importance of forecasting the correct stock requirements to meet demand.

Setting customer service priorities means that there will some sort of classification of stock keeping units (SKU's) with some receiving more resource than others. The

Pareto rule, or 80/20 rule can be used to provide a cost-effective service strategy. This rule reflects that 20% of the input created 80% of the result (in this case SKU's).

This was highlighted in a case study of an electronics manufacturer (Syntetos et al., 2008) where the introduction of a Pareto classification (in this case value per SKU) resulted in significant improvements in inventory and availability.

This area is important to the PHD experiment as the case study is within a Pan European Spare parts center which has target hit ratios based on a Pareto like ABC structures that aims to keep the level of service at an optimal level. This optimal level is a tradeoff between the investment into inventory and the level of service provided to customers. This classification is not peculiar to spare parts and its usage is ubiquitous across products, consumables and accessories also.

2.3. Supply Chain Forecasting

In answering the question 'why forecast?' Makridakis et al., (1983) defined it as the need to determine when an event would occur, or a need arise, so appropriate action could be taken.

The production of a forecast can be important as it implies, we can change variables in the present to alter the future or to be prepared for an event.

Forecasting is essential for business planning and has been an area of academic research for many years. For example, Yule (1923) discussed periodicities in disturbed series using, amongst other methods, harmonic curves and regression equations. Similarly, we can see that the use of judgment was reported by Dalkey and Helmer (1962) in their report into the number of A-bombs required to successfully destroy specific locations in case of war.

Makridakis et al., (1983) suggested that management practice was the area where greatest gains from forecasting research would result. It was thought that management knowledge rather than improvements in methodologies would be where the greater benefits would be realized. This was reiterated by Armstrong (1988) who called for more research on implementation of forecast techniques and posed the question of how to more effectively gain acceptance from management of favorable forecasts.

McCarthy et al., (2006) commented on the way business environment has changed due to advancement and adoption of information technology, globalization and new business practices such as E-Commerce. Their study looked at how forecasting management practices had changed over 20 years compared to previous findings as reported by Mentzer and Cox (1984) and Mentzer and Kahn (1995). By conducting a web-based survey, they explored trends in management of forecasts, familiarity, satisfaction, usage and accuracy amongst a variety of industries. Their findings were that for all the technical advancements in both methodology and technology, accuracy had not improved, methodological understanding had not improved and overall satisfaction with techniques and management showed no improvement either. Amongst their reasons for lack of improvements were:

- Poor understanding of techniques (black box forecasting)
- Lack of satisfaction meaning people resorting to manual data calculations (spreadsheets)

- Little accountability for forecasting accuracy
- No recognized forecasting function (disjointed approach).

They concluded that companies should commit resources to a cross functional forecast process with people trained in both quantitative and qualitative forecasting techniques. These issues were also discussed in an overview of empirical studies on forecasting practices by Winklhofer and Diamantopolous (1996) with similar findings but 10 years earlier.

The comments by McCarthy et al., (2006) will be considered in the results from the experiment.

2.3.1. Forecasting methods

The PHD will look at the forecasting of spare parts so the discussion is framed with a view to dealing with these SKU's specifically.

Forecasting techniques fall into two broad categories:

- Quantitative forecasting. This approach relies on time series data alone or is based on the relationship between the data and other variables (causal).
- Judgmental forecasting. Which is the subjective assessment of an individual or group of individuals. Which can also be scientific in its structure.

2.3.2. Quantitative methods

Considering quantitative methods using time series, the techniques look for underlining patterns such as trend and seasonality in the historical data and try to project them into the future. There are numerous time series methods some of which are used for specific patterns:

- Naïve
- Moving average
- Weighted average methods
    - Exponential smoothing
    - Holts and Winters methods
- Decomposition
- ARIMA

Dealing with the methods in order of the list above the Naïve is the most basic method as it assumes the future will replicate the past. As we know the marketplace is a dynamic one so there will be series where this is not the best method.

Weighted moving average techniques are often referred to as smoothing models since they level out random fluctuations. Perhaps the most commonly used method is exponential smoothing. This technique allocates more importance to recent periods about producing a forecast. Winters (1960) method is an example of numerous techniques that use this general starting point.

Decomposition methods by splitting the forecasting task into components and them combining these smaller (less demanding) forecasts. Edmundson (1990) found that by obtaining separate estimates for random components, seasonality and trend when extrapolating a times series and then combining them the forecast accuracy was improved when compared to the holistic forecasts. Armstrong and

Collopy (1993) showed more accurate forecasts were produced by asking judges to focus on specific factors that were influencing a time series when selecting and weighting a statistical forecast. A similar study where the impact of numerous special events effected a time series was shown to produce more accurate forecasts when each event was separately forecast then combined rather than forecast all together (Webby, O'Connor and Edmundson, 2005).

ARIMA is the most advanced technique in the list and is a combination of both time series and regression models. It works by identifying an association between variables at different time periods using correlation coefficients. This model is the least used in practice due to the relative complexity of the technique and the lack of practitioners who are prepared to use it.

The strengths of time series methods can be summarised as (Chase, 2013):

- They are suited if forecasts are required for many products
- If the time series is relatively stable, they work well
- Most of the techniques are quite simple to understand
- If the forecast horizon is short, they perform well (up to 3 periods in the future)

The weaknesses of time series methods can be summarised as (Chase, 2013):

- The amount of data required can be large
- Large fluctuations in data create large errors
- The forecast horizon needs to be short
- Smoothing factors can take time to optimize

The other quantitative method is causal. The idea here is that future forecasts are related to changes in one or more variables (price, promotions or advertising for example). This relation is quantified allowing forecasts to be produced.

Therefore, once the nature of that association is quantified, it can be used to forecast sales. The most widely used causal methods are simple or multiple regression and ARIMAX.

Simple regression models the interrelationship between two data sets. The causal relationship allows one variable to be used to predict another (for example price and demand). Multiple regression uses two or more independent data series.

The ARIMAX models are simply extensions of ARIMA models. Explanatory variables are added to the model to explain any variance. For example, by adding advertising, price and sales promotions the existing model can be enhanced by explaining away unexplained variance. The introduction of more and more variables does not always mean the model will produce more accurate forecasts although it will be better at explaining history.

The strengths of causal methods can be summarised as (Chase 2013):

- They are available in software packages
- They can provide medium term forecasts better than time series models
- "What if" analysis can be tested. Meaning a variable can be changed to see what the possible implications are for the future forecasts.

The weaknesses of causal methods can be summarised as (Chase, 2013):

- The variables need to have a consistent relationship

- Often practitioners see these forecasts as ‘black box’ techniques
- The level of understanding required is much greater

2.3.3. Qualitative forecasting

Judgmental or qualitative forecasting in its purest form involves no manipulation of data. The forecaster uses their knowledge of the product history to produce the forecast. The use of judgment can also be in combination with a quantitative method or can be used to adjust the final forecast.

The use of judgment is particularly useful for new product releases and where there is a known factor which will cause a large fluctuation in the time series such as a price promotion or a marketing push for a specific product.

In relation to spare parts an example of where judgment would outperform or help a quantitative forecast would be where there are known faults in a product and demand could be predicted to rise considerably.

Some of the widely used judgmental forecasting techniques include independent judgment, combination of judgment with a quantitative forecast, juries of opinion and expert opinion. These will be explained in greater detail further in this chapter.

The strengths of judgmental methods are:

- They are low cost.
- The forecast can be produced quickly.
- Large changes to demand can be anticipated with prior knowledge.

- Management understand the factors that influence the forecast.

The weaknesses of judgmental methods are:

- They are subjective and therefore biased.
- The forecaster may not have all the necessary information.
- Are not suited to large numbers of products.

Chase (2013) argues that there are segments of a company's products that will have different forecast-ability, a different forecasting technique should be applied, depending on which segment a specific product is located in. This is because it is known that some techniques work better for differing time series.

It is argued that often the wrong forecasting technique is used for different products rather than applying an appropriate one depending on where the product resides in a segment resulting in poor forecasts (Lawrence et al., 2006).

Chase argues that companies can segment their products to get the most accuracy across their product portfolio. This is shown in the table below:

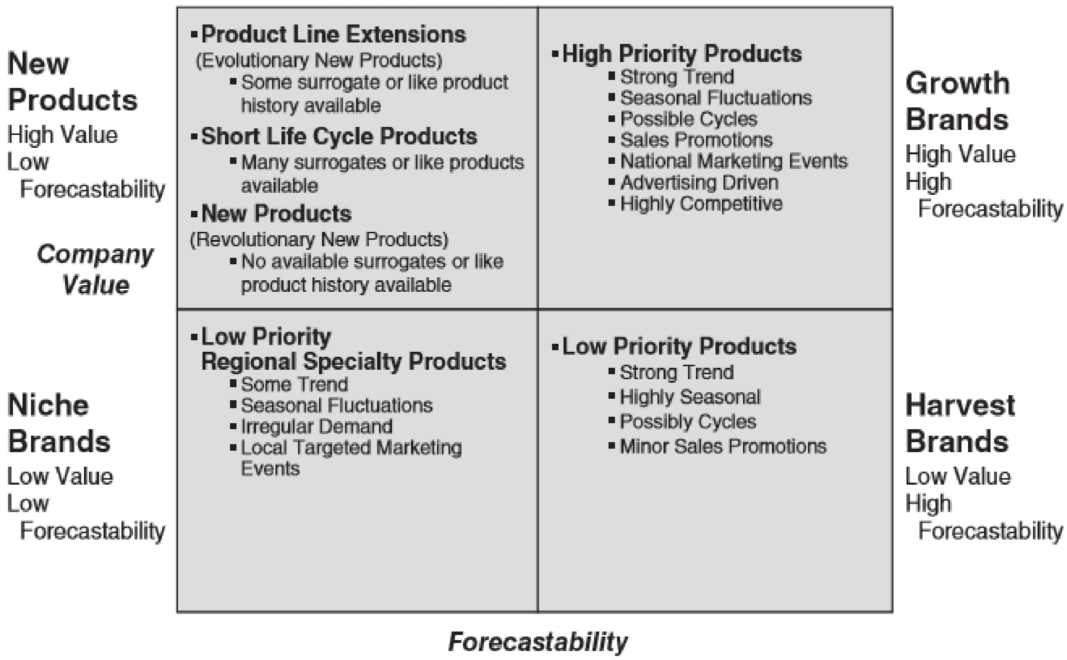


Figure 2.2. Transforms Four Quadrants Using Product Portfolio Management Principles

Using the table, a company should locate the product and then apply the best technique to give the best possible forecast accuracy. The term 'judgment' is replaced by domain knowledge and it is recommended that this is used in combination with quantitative techniques where possible.

2.4. The use of Judgment in industry

There have been numerous studies examining the use of judgment in industry compared to quantitative techniques.

By surveying forecasting practices in 500 US corporations, Sanders and Mandrodt (1994) explored the reasons why judgmental methods were used so heavily in comparison to quantitative procedures. The investigation found that although managers were more familiar with quantitative methods than in the past, the level of usage had not increased. They found that the main issues were a lack of data and support within the organization. Even when quantitative methods were used, they were regularly judgmentally adjusted. This research supported previous surveys where judgment was also found to be preferred to quantitative methods (Dalrymple 1975, 1987; Mentzer and Cox 1984; Wheelwright and Clarke 1976).

The main areas of interest in the survey were: whether forecasting practices were changing due to the availability of software; to understand the role of judgment and why judgment was so prominent; examine forecasting problems and the requirements of practitioners.

Sanders and Mandrodt (1994) saw that practitioners were much more familiar with quantitative methods than previously reported by Mentzer and Cox (75% compared to 61%) and unfamiliarity of such methods was also reduced in comparison (15% and 27%). They argued that more difficult techniques such as Box Jenkins and Decomposition still showed a low level of familiarity which was on a par with the Mentzer and Cox findings.

When the results of the survey were analyzed to show the percentage of methods used the following table showed that Judgment was used more than any other technique for short and medium-term forecasting and that regression was used for longer forecasting requirements.

Forecasting Techniques	Forecast Period: Immediate <1 month	Short 1 month–<6 months	Medium 6 months–1 year	Long >1 year
Judgmental				
Manager's opinion	27.9 (—)	39.8 (—)	37.1 (—)	9.3 (—)
Jury of executive opinion	17.5 (—)	28.9 (37)	40.1 (42)	26.2 (38)
Sales force composite	28.6 (—)	17.5 (37)	33.1 (36)	8.7 (8)
Quantitative				
Moving average	17.7 (—)	33.5 (24)	28.3 (22)	8.7 (5)
Straight line projection	7.6 (—)	13.2 (13)	12.5 (16)	8.2 (10)
Naive	16.0 (—)	18.5 (—)	13.8 (—)	0 (—)
Exponential smoothing	12.9 (—)	19.6 (24)	16.8 (17)	4.2 (6)
Regression	13.4 (—)	25.1 (14)	26.4 (36)	16.5 (28)
Simulation	3.4 (—)	7.8 (4)	11.2 (9)	8.3 (10)
Classical decomposition	0 (—)	6.8 (9)	11.9 (13)	9.3 (5)
Box-Jenkins	2.4 (—)	2.4 (5)	4.9 (6)	3.4 (2)

Table 2.3. Survey results from Sanders and Manrodt (1994) showing the percentage of managers that reported using techniques for different time horizons. The number in parenthesis are those reported by Mentzer and Cox (1984).

When Sanders and Manrodt (1994) considered the satisfaction, managers reported with different techniques, they reported that the response showed a high level of satisfaction with judgmental methods. This is illustrated in the table below:

		Percentage Response		
Techniques	n	Satisfied	Neutral	Dissatisfied
Judgmental				
Manager's opinion	81	64.6 (—)	14.6 (—)	20.8 (—)
Jury of executive opinion	78	72.0 (54)	13.7 (24)	14.3 (22)
Sales force composite	77	47.7 (43)	38.4 (25)	13.9 (32)
Quantitative				
Moving average	79	44.6 (58)	31.2 (21)	24.2 (21)
Straight line projection	64	26.9 (32)	31.8 (31)	41.3 (37)
Naive	64	18.9 (—)	27.4 (—)	53.7 (—)
Exponential smoothing	65	40.8 (60)	31.0 (19)	28.2 (21)
Regression	73	45.8 (67)	34.4 (19)	19.8 (14)
Simulation	49	42.5 (54)	37.7 (18)	19.8 (28)
Classical decomposition	45	39.4 (55)	34.3 (14)	26.3 (31)
Box-Jenkins	43	22.3 (30)	29.2 (13)	48.5 (57)

Table 2.4. the table shows the percentage of managers reporting satisfaction with forecasting techniques, Sanders and Manrodt (1994). The numbers in parenthesis show results by Mentzer and Coz (1984).

Sanders and Manrodt (1994) also asked what the relative frequency of use regarding judgmental methods was. Their findings indicated 57% for respondents reported that they always used judgment. It was also seen that where quantitative methods were used 45% of respondents always judgmentally adjusted the forecast. The reasons stated were that knowledge of the environment needed to be incorporated and special events or changing conditions.

When asked how researchers can help practitioners regarding forecasting, the highest percentage responses were easier techniques to use, better information regarding the techniques and more accurate techniques.

A survey of forecasting practice was carried out by Klassen and Flores (2001), with an aim to provide a better understanding of forecasting practices by utilizing a questionnaire. The results appeared to be like those of Sanders and Manrodt (1994).

Frequency of forecasting methods use. Values are numbers of firms. Data in parenthesis are percentages

Method	Never used	Currently in use	Previously used
Judgmental methods			
Naïve	22 (19)	82 (70)	6 (5)
Jury of executive opinion	49 (42)	45 (38)	5 (4)
Sales force composite	31 (27)	70 (60)	10 (9)
Counting methods			
Industrial survey	61 (52)	39 (33)	9 (8)
Intention to buy survey	62 (53)	45 (38)	5 (4)
Time series			
% rate of change	59 (50)	42 (36)	11 (9)
Unit rate of change	73 (62)	33 (28)	5 (4)
Moving averages	60 (51)	43 (37)	8 (7)
Single exponential smoothing	84 (72)	15 (13)	12 (10)
Holt	103 (88)	7 (6)	0 (0)
Winters	95 (81)	13 (11)	2 (2)
Box–Jenkins	103 (88)	5 (4)	4 (3)
Association or causal methods			
Simple regression analysis	77 (66)	25 (21)	8 (7)
Multiple regression analysis	84 (72)	21 (18)	6 (5)
Leading indicators	54 (46)	52 (44)	6 (5)
Econometric methods	83 (71)	26 (22)	3 (3)
Newer methods			
Chaos theory	109 (93)	1 (1)	2 (2)
Expert systems	90 (77)	18 (15)	1 (1)
Genetic algorithms	106 (91)	3 (3)	1 (1)
Neural networks	105 (90)	3 (3)	2 (2)

Table 2.5 Frequency of forecasting methods use. Values are numbers of firms. Data in parenthesis are percentages.

The table shows that judgmental methods are used in significantly more instances that all the other methods with the naïve and sales force composite methods being the most reported. In their conclusion, Klassen and Flores (2001) reported that judgmental procedures are used more frequently than any other method. Senior management also revises forecasts and believes that this improves accuracy (which, when investigated, showed the variation in improvement may negate any advantage gained). This was the case even when forecasts were required for longer horizons.

They also concluded that practitioners are not employing many of the technique's academics teach, specifically the use of sophisticated quantitative methods, combination of forecasts and the estimation of confidence intervals.

Further to their survey in 2001, Sanders and Manrodt (2003) designed a study by profiling differences between firms that used predominantly judgmental or quantitative methods to try to understand why this was the case.

Relying on survey data from 240 companies they statistically analyzed differences between these categories of users based on a range of organizational and forecasting issues. They discovered differences in error rates between the two groups with judgmental users being outperformed by users of quantitative methods. Companies using judgmental methods appeared to have lower access to data and used information and technology to a lesser degree.

They summarized their study by outlining five major points that highlighted differences between users of judgmental and quantitative methods:

- Companies who focused on judgment tended to work in areas where there was higher uncertainty. This gave more reliance to subjective information

- Subjective information was given greater importance in the generation of forecasts where judgment was used predominantly. There was also less access to quantifiable information

- Several of the judgment focused companies were not satisfied with the current forecasting software, adjust software output and do not integrate into other management systems
- The size, industry, or product positioning strategy does not have an impact on the prevalence of judgment focused companies
- There is a significantly greater number of errors at companies who use predominantly judgment

Sanders and Manrodt (2003) highlighted areas for potential research one of which could be a study where a company used primarily judgment to examine if this was after a quantitative forecast had been produced. Another area for future research called for was where forecasting resides in current organizations. Was there a specific forecasting department or did it reside in sales for example? What was the forecasters experience and organizational role and how these could potentially affect the forecasting methodology used?

A 25-year review of judgmental forecasting was undertaken by Lawrence et al., (2006). They commentated on the change in attitude of researchers to the role of judgment. Whilst judgment was previously often seen as an inaccurate technique (Hogarth and Makridakis, 1981) it is now seen as an important part of forecasting. Judgment can be shown to benefit forecast accuracy, but it can also be subject to bias.

The case of Nike was used as an example where due to a lack of management input huge inventory write offs were the result when the system accuracy was not adjusted (Worthen, 2003). In this case $400 million was invested into forecasting software which produced forecasts that proved not accurate. It was commented that in the United States many large corporations invested significantly into forecasting software and the results were poor. This experience of forecasting software therefore may explain that only 11% of the 240 US companies surveyed by Sanders and Manrodt (2003) used it.

Lawrence et al., (2006) regarded it as expected that wherever in sales forecasting there was an impact of price promotion or competitor activity then judgement would be built into the forecasts. They also noted that in macro-economic forecasts judgment was shown to be favorable (Fildes and Stekler, 2002).

The use of forecasting technique will be examined in the case study firstly as a reflection of the current method in the case study company as compared to the earlier studies reported and secondly to understand the reasons why the techniques found are used as oppose to the other options available.

2.5. Spare parts in industry

Koudal (2006) commented on Deloitte Research (2006) that estimated that Service operations were accounting for 25% of revenues of many of the world's largest manufacturing companies with combined total revenues of $1.5 trillion. Typically accounting for 40% of the annual procurement budget for maintenance and repair operations, spare parts also provide an opportunity for cost reduction and thus increased profits (Donnelly, 2013).

Syntetos et al., (2010) highlighted that after sales services were important to a range of companies generating 50% of total profit and this figure was similarly reported by Kim et al., (2007).

The inventory value of spare parts is also a high-risk area due to their characteristics (sporadic, intermittent, and thus difficult to forecast). The US Department of Defense reported that the value of its inventory was $95 billion in 2012 (GAO, 2012). This level had been static for some years (GAO, 2011). There was virtually no decrease despite a focus on reducing it. This level of high investment is mirrored in the UK where Morse (2012) reported a £35 billion inventory. There are huge savings to be made if the forecasting of spare parts could be improved across many sectors of industry.

Syntetos et al., (2008) state that forecasting at the stock keeping unit level (SKU) in order to support operations whilst managing inventory is a difficult task. There is research to suggest that practitioners do rely on judgmental forecasting methods. When quantitative methods are used to provide a statistical forecast they are jundgmentally adjusted (Sanders and Manrodt, 1994: Fildes and Goodwin, 2007).

The use of judment in forecasting when dealing with spares was commented on by Bacchetti and Saccani (2012). In their investigation of ten case studies, they found that whilst some companies did not carry out any forecasts using a reactive approach (forecasting after a demand occurs) only, others did forecast only using quantitative methods. There were others who either forecasted totally using judgment or judgmentally adjusted the statistical forecast provided. This highlighted that in practice there was a varied approach to spare parts forecasting and no real best practice with spare parts themselves not separated fom the other products in forecasting techniques. No firm in the study approached intermittent or lumpy demand in a different way. They highlighted the research gap in spare management practice in amongst other areas, the adoption of demand forecasting methods for spare parts. They called for more empirical research in the spare parts area through a vertical case study method.

A case study by Syntetos et al., (2009) regarding a manufacturing company looked at how spare parts were forecasted. In this case the parts were split into ABC using a simple Pareto method where the A items (the fastest moving and thus the most important to have in stock) were forecasted judgmentally. The number of items was relatively small and enabled the company to scrutinise each SKU on its own. The reasons for using judgmennt were that although time series techniques were good for spare parts (as the total SKU count was 3-4k), the most important items should be looked at judgmentally so any other impact could be included in the forecast. These were stated as fault notifications in product and also service agreements that required a replacement of certain parts that were scheduled across Europe.

It is clear from the literature that spare parts are intrinsically different to most products. The usually high number of spare parts means that judgment is precluded from being the main forecasting tool due to the time it would take and the resources available to companies. However, judgment is used for items of high importance which can be separated from the bulk of the SKU's to include issues outside the historical data. From the numbers reported by Deloitte (2006) we can see that the possible overstocking of spares is a real issue whilst companies are also very wary of not being able to support after sales due to the impact to profit (Syntetos et al., 2010).

The importance of spare parts within the case study company will be discussed in the PHD and the strategy for stocking and policies internally agreed will be used to reflect on the literature.

2.6. Combination of forecasts

All forecasts contain some sort of judgment. If a forecast is from a complex quantitative method, then that method was chosen by a human judgment. So, in a sense, all quantitative forecasts are combined forecasts. When a practitioner is faced with a choice of statistical models the judgment involved will have a direct effect on the accuracy of the statistical forecast. If an ERP system is used there are usually many options from which the forecast can be derived (different algorithms are available). Depending on the judgment of the user it may be a constant average that is chosen or a more complex algorithm either way this is a judgmental choice. In some systems there is the opportunity to choose the method which has the lowest errors based on the past periods. This would take the judgment part of the process away but can be dangerous when dealing with outliers. The existence of large errors can mean a less than optimal option is chosen. That is, if the outlier was ignored a different result would be delivered.

Collopy and Armstrong (1992) produced a rule-based method which incorporated judgment into it.

Their rule integrated different aspects of forecasting including:

- features of the data to establish weights for combining
- using heuristics to establish parameters for exponential smoothing
- using separate models for long and short-term forecasting
- damping the trend depending on conditions
- incorporating domain knowledge in the extrapolation

From these rules, we can see that there is not only an argument to combine statistical forecasts but also to combine them with judgment.

We do know from studies. (Klassen and Flores 2001; Sanders and Manrodt 1994) that the role of judgment is often employed at a significant level and sometimes to the quantity to be forecast itself. In many organizations, forecasts are either completely judgmental or are the judgmentally adjusted outcome of a statistical forecast.

Managers tend to feel the desire to alter statistical forecasts for several reasons. Statistical methods are considered too slow to react to step changes (Blattberg and Hoch, 1990). There can be instances where there is no previous data relating to the requirement of the SKU (Hughes, 2001); even where data does exist, it can include events that require the data to be smoothed manually such as price promotions or advertising campaigns and this can be either impossible or very time consuming (Edmundson et al., 1988). Judgment can be favored as employees either do not have the skills to understand the statistical forecast (Fildes and Hastings, 1994), or even if they do, behavioral factors may imply that they favor judgment anyway.

There are some companies that do use combination but not of a statistical method with judgment but of a combination of judgmental forecasts (O'Connor and

Lawrence, 2005). There are possible problems with this method where groups can be dominated by individuals which cause a lack of views being exchanged.

This led to the development of techniques such as the Delphi method (Rowe and Wright, 1996). There have been some discussions as to whether this technique where after many rounds of discussion a group of experts arrive at a consensus opinion needs to be more rigorous. Bolger and Wright (2011) discuss the Judge - Advisor system where they report people do not change their own opinions often enough. This is called *egocentric discounting* that is giving too much weight to their own opinion and not enough to the opinion of others. They suggest some form of incentive to increase people's participation and accountability.

Pure judgment when used on its own can have disadvantages to statistical methods. Inevitably people use strategies to simplify the forecasting task and these heuristics can bring judgment bias to the forecast (Bolger and Harvey, 1993). This will be covered in more detail later in the chapter.

Goodwin (2000) argues that the compelling reason for integration of judgment and statistical methods is because of the complementary strengths and weaknesses. Human judges can consider one off event's and are adaptable, although they are inconsistent and can only deal with a finite amount of data. They also bring a cognitive bias to the forecast. Statistical techniques however are rigid and consistent and can be applied to large amounts of data. Due to these characteristics, Goodwin argues it seems reasonable that an integration of the two would improve the forecast accuracy.

Two methods of integration are discussed, voluntary and mechanical. In voluntary integration, the forecaster is supplied with details of the statistical forecasts and then depending upon judgment decides how to use the information. This could be to ignore the statistical forecast completely, accept completely or use it to some extent to inform the judgmental forecast. This is usually in the form of adjustments to the statistical forecast but could be by altering previous forecasts after viewing the statistical forecast. Mechanical integration involves the application of a statistical technique to the judgmental forecast. A formula is used to combine the estimates of the statistical forecast with the judgmental forecast. This can be in the form of a simple formula such as an average of the statistical forecast with the judgmental one. Armstrong (2001) reported that this was superior to the individual forecasts themselves, working best when there was a negative correlation between the two forecasts.

Goodwin (2000) investigated the question of whether it is better to correct a judgmental forecast or to combine it (mechanical integration). It was concluded that where difficult to model, and non-time series information was available then the best role of statistical methods was to correct judgmental forecasts.

Goodwin (2000) also reported on voluntary integration. It was noted that when a forecaster was asked to request a change to a statistical forecast and subsequently codify the reason, the forecast accuracy improved. Goodwin (2000) argued that by making the change to a forecast as an option, it removed the belief from managers that they were expected to make changes. The fact that they could accept the statistical forecast could also be a reduction of effort from their part rather than explain a reason why the change should be made. Finally, Goodwin (2000) argued that the forecaster had been made to focus on the statistical forecast and consider more why a change was needed.

This is an important part of the PHD as the experiment will report what level of combination there is in the study and examine whether the accuracy of the forecast was improved or not.

2.7. Judgment and causal influences

Managers and experts often have useful knowledge about future situations. This domain knowledge can be important for forecasting. Many quantitative methods do not incorporate such knowledge and thus are often inaccurate.

Expectations from managers can be informative regarding the future trend of a time series. These causal forces can cause significant errors if a statistical forecast does not incorporate them. Knowledge of company decisions or strategy that will affect the demand series in the future is not available in historical time series.

A simple example of this would be where a company is phasing out a model and replacing it with a heavily marketed new one. The older product would create a contrary time series, one where the future trend would look like an upward one but where it would be downward in reality. If no causal information was used in the forecast, then the possible errors could be significant.

In their rule-based method for forecasting, Collopy and Armstrong (1992) included causal forces as they can play an important role in how a method is selected and weighted. They also showed how this causal information could be used to reduce errors. By using a naïve forecast for a large dataset from the M-competition

(Makridakis et al., 1982) when they encountered a contrary series, and this reduced both short and long-term forecast errors significantly.

There have been a few studies which have demonstrated that incorporating causal information in a final forecast is important.

Discontinuity of products has been used to emphasise the role of judgment as it can identify when a significant reduction in demand may occur (Kleinmuntz, 1990). Sanders and Ritzman (1992) showed that causal information improved the accuracy of final forecasts compared to the statistical and judgmental estimates. Mathews and Diamantopoulos (1986) showed that where managers had sufficient knowledge then they could reduce errors by adjusting exponential smoothing forecasts.

A study into the way people adjust statistical forecasts where causal information was available was undertaken by Lim and O'Connor (1996). Their results showed that the causal information was reasonably incorporated to cover for what a time series lacked and that the effectiveness of the causal adjustment was dependent on the causal information. Poor causal information did not result in significant improvement, but incorporation of good information outperformed the statistical models used. They commented that forecasters relied too much on the initial forecast in comparison to the optimal model and that over time people did not learn to be less conservative. The statistical forecast was also preferred to the forecast that included the causal information.

There are very few academic studies providing empirical evidence when looking at how people use causal information despite it being shown to be of importance to produce a forecast.

Harvey et al., (1994) explored the accuracy of judgmental extrapolation and causal forecasting. It was shown that forecasting based on the single data set alone was less accurate then when the causal information was included (also as a time series). Here, the two data sets concerned passengers and criminals on trains, and it was thought that in many cases the relationships between the data sets could be much more complex.

A further study, although not as an adjustment task, compared the accuracy of judgmental extrapolation and judgmental causal forecasting (Andreassen, 1991). The study demonstrated that using the naïve method for judgmental extrapolation of the last month's value was more accurate than the causal forecast, confirming Makridakis et al., (1982) that the naïve was a good forecast.

The effect of causal information will be considered in the PHD. The question of whether it was incorporated or not will be reported on.

2.8. The importance of forecasting horizon

There have been many studies that have looked at point forecasts (that is the next data point after the most recent one) but much fewer studies into judgmental forecasts for periods further into the future (horizons). The further the forecast is for into the future the more the uncertainty increases. Due to this, forecasts (both statistical and judgmental) tend to be worse the further the forecast horizon

(Lawrence et al., 1986). Judgmental forecasts show damp trending which causes errors to increase over forecast horizons.

Harvey and Reimers (2013) discussed that the trend damping can occur because a) people anchor to the last data point and make smaller adjustments to take a trend into account, b) they tend to adjust towards the average experienced within an experiment or c) the environment they work in naturally adapts them to dampen trends.

When looking at how forecasts for longer horizons are made Bolger and Harvey (1993) concluded in the analysis of their experiment that previous forecasts were used as mental anchors rather than the last data points. Theocharis and Harvey (2016) conclude that if this is the case any random noise added to forecasts would accumulate the further the forecast horizon. It then follows that any reduction of this noise would improve the accuracy, and the level of variability going forward, would be reduced.

A way of making this the case would be to ask the forecasters to forecast the furthest horizon first. Theocharis and Harvey (2016) conducted an experiment to test whether this was the case. By conducting two experiments; one requiring the forecast from the nearest horizon, and one in a different direction with the most distant horizon the first forecast (end-anchoring). This end anchoring appeared to improve the forecast accuracy, more so for horizons further away, and made the trajectory of the forecast closer to the optimal one. This simple change to the forecasting sequence was considered to be a cheaper and quicker method than other examples which were deemed to improve judgmental forecasting such as combination (Collopy and Armstrong, 1992, feedback (Goodwin and Fildes, 1999) and use of advisors (Lim and O'Connor 1995), the latter two to be discussed later in this chapter.

In a 20-year, longitudinal study of forecasting practices the use of forecasting techniques relating to forecast horizon was reported by McCarthy et al., (2006).

Technique	Short horizon ≤3 months			Mid horizon 4 months–2 years			Long horizon >2 years		
	M&C	M&K	PS	M&C	M&K	PS	M&C	M&K	PS
Qualitative									
Jury of executive opinion	1	5	na	1	2	1	1	1	1
Sales force composite	1	5	2	2	2	3	8	4	3
Customer expectations	3	3	1	5	8	4	4	7	na
Quantitative									
Moving average	4	1	6	6	6	7	10	9	na
Straight-line projection	8	3	6	8	9	8	6	10	6
Exponential smoothing	4	2	3	7	1	2	9	6	5
Regression	7	5	3	2	4	6	2	2	2
Trend-line analysis	6	8	3	4	5	5	3	3	4
Simulation	11	12	na	10	13	10	6	8	na
Life cycle analysis	12	12	6	12	10	11	5	5	na
Decomposition	9	8	na	9	7	8	10	10	na
Box–Jenkins time series	10	8	na	11	11	11	11	12	na
Expert systems	nm	12	na	nm	13	11	nm	11	6
Neural networks	nm	8	na	nm	12	11	nm	13	na

Notes: M&C, Mentzer and Cox (1984), sample size = 160; M&K, Mentzer and Kahn (1995), sample size = 186; PS, present study, sample size = 86; nm, not measured in the study; na, not applicable (no respondents indicated use of the technique for that time horizon).

Table 2.6. Familiarity with forecasting techniques

It showed that Qualitative methods (judgmental) were the most predominant for most time horizons. This could be down to the relative ease to understand the techniques and learn. The predominance of these judgmental techniques does not represent or reflect that they are the optimal ones (there are some studies that would argue otherwise), it is simply the case that the respondents in the survey were using them.

Analysing a database of pharmaceutical sales forecasts which had been adjusted by a range of managers, Franses and Legestree (2011) looked at whether the forecast horizon had an impact on what managers did and how good they were at adjusting model-based forecasts. Using regression, they showed that all horizons were judgmentally adjusted, the horizon that was most relevant to the manager showed overweighting, for all horizons the adjusted forecasts were less accurate than the statistical forecast with the horizons furthest away showing the least errors. When manager's adjustments were down weighted, the accuracy of the forecast improved. It was concluded that forecast horizons were indeed important to forecast accuracy and that managers needed more training to understand that they over weighted the forecasts which were relevant to them and over weighted generally.

The forecasts in the study will be over one to three months. It will be possible to check if the forecasts for the longer forecast horizon we more accurate than the shorter ones.

2.9. Types of Knowledge

There have been many studies into what type of knowledge is required by a forecaster to run a statistical method and to judgmentally adjust the outcomes.

An experiment by Carbonne et al., (1983) tried to answer four key questions regarding forecasting:

- What sort of expertise is required to use a forecasting method? Note this is not expertise in the field of interest but of the technique involved

- Which time series give more accurate results for forecaster with no training?
- Can judgmental abilities of typical users improve forecasts made by time series models?
- Can individuals with limited training provide a better forecast with a time series model than those obtained using the same model in an automatic model? So, does the quality of the forecast depend upon the time taken by the forecaster?

In their conclusion to the questions they found there was very little difference when a forecaster with little expertise ran a complex model such as Box-Jenkins. The results were like those of an expert.

When comparing the accuracy for time series models when it was run by a novice it was reported that the simpler the model was the most accurate the forecast became.

When people with limited experience applied judgment to the forecasts only the most sophisticated model (Box-Jenkins) was still good with no significant distribution found. However, for the simpler time series models, the judgmental forecasts were less accurate.

When students were given longer to apply their knowledge the results were less accurate than the black box approach (no change to the statistically derived forecast).

It would seem common sense that at forecaster’s technical knowledge would improve the performance of judgmental forecasts but the results from Carbone et al., (1993) did not support this. There was no difference between a forecaster with limited technical knowledge and a limited skilled forecaster.

This research was extended by Sanders and Ritzman (1995) who investigated the benefits in forecast accuracy by combining judgmental forecasts with those generated by statistical models. Their study differed from earlier research in this area in two ways. Two different types of judgmental forecasts are evaluated for combination with statistical forecasts - one based on contextual knowledge and one based on technical knowledge. Contextual knowledge is information obtained via job experience using time series for products being forecasted. Technical knowledge is information obtained via using formal models and data analysis and by education.

They tested two hypotheses:

- Combination forecasts which combine statistical forecasts with judgmental forecasts based on contextual knowledge are significantly more accurate than when combined with judgmental forecasts based on technical knowledge
- Forecasts generated as a combination of statistical and judgmental forecasts with contextual knowledge are significantly more accurate than forecasts generated by statistical methods alone

The results showed that combining statistically derived forecasts with those of experienced practitioners improved accuracy more than other combinations. Judgmental forecasts of students, where knowledge was purely technical in contrast, did not provide the same level of contribution to the forecast combination.

This experiment supported Sanders and Ritzman's earlier report from 1992 that technical know how does not improve accuracy amongst others, Edmundson (1990) concluded similarly. Lawrence et al., (1986) did find that technical knowledge had a positive effect when it was presented in a tabular format (rather than a graphical one).

2.10. Judgment and added opinion

In their paper "Advice taking and decision making" Bonaccio and Dalal (2006) discuss the current literature under the headings: Advice utilization; Judge and Advisor confidence; Accuracy of final decisions; Differences between advisors and "personal decision makers". The judge - advisor system (JAS) was used as a framework to explain the literature. The judge or decision maker is the person who receives the advice and must decide what to with it whilst also making the final decision. The advisor is the source of advice, usually preferring an option.

The judge would solicit advice for numerous reasons. It could be to share responsibility for the outcome of a forecast and to improve the chance that the final decision will be optimal (Yaniv, 2004a, 2004b). The advice could come from one or more sources with Yaniv (2004a) reporting that three to six would be optimal.

It has been shown that often judges do not follow advice as much as they should do (Yaniv and Kleinberger, 2000) to produce the optimal results. This behavior was called egocentric advice discounting. There have been numerous papers that advice does improve judge's accuracy but that they tend to weight their opinion much more than that of the advisor (Harvey and Fisher, 1997). Yaniv (2000a) argued that this was the case because the judge did not know the advisor's reasons for the advice it was better to keep closer to their own to justify the forecast if required. Harvey and Fisher (1997) also commented that the initial forecast from the judge is often used as an anchor from which any adjustment is made. The reason for discounting any advice could also be down to the fact that the judge prefers their own opinion and believes it to be optimal. Krueger (2003) notes that even when the judge is forecasting something that is new, and they have no experience of the scenario this can still be the case.

If the advice is given by people who the judge perceives as experts or more knowledgeable than themselves then the advice is more useful than advice from a novice (Jungermann and Fischer, 2005). This is also the case where the judge is a relative novice in comparison to the advisor (Harvey and Fischer, 1997). The quality of the advice has also seen to affect the level of discounting the judge applies. If poor advice is given then this is discounted more than good advice (Lim and O'Connor, 1995). The fact that the judge has requested advice can also be an indicator as to whether they will discount advice less. Gibbons, Sniezek and Dalal (2003) reported that is advice was given when it was not requested then is more likely to be discounted.

Confidence, either the judge's or the advisors, can be important when advice is given. The more confident an advisor is, appears to result in less discounting (Lawrence and Warren, 2003). Judges would allocate more expertise to an overconfident advisor compared to more appropriately confident one, leading to the question: could advisors act overconfidently to strategically effect and influence judges? (Yates et al., 1996). If there is more than one advisor and the advice is not the same the judge's post advice confidence is low (Budescu et al., 2003). This is more pronounced when the judge believes the advisors had access to the same information (Budescu and Rantilla, 2000).

The product of the JAS system or the output has been shown to in general improve the accuracy of the final decision (Yaniv, 2004a). Yaniv argued that the combination of uncorrelated advisors increased accuracy due to the reduction in random errors tied to each individual recommendation. This aggregating means that the variability is reduced along with the random error improving the overall forecast. Similarly, by using different sources of advice that are clearly separated from each other the helped to improve the accuracy of the forecast. An example of this was reported by Harvey, Harries and Fischer (2000) who showed that this could reveal forecasting trends where the task was unusual thus showing a bias that would stand out as opposite in the forecasts for the forecaster to recognize. It seems obvious to highlight but it is nonetheless noteworthy that where advisors have more relevant information, they tend to offer better advice, and the resulting forecast is then improved (Humphrey et al., 2002).

There has been much written about the judge who is taking the advice rather than the advice giver. Research has shown that advisors will tend to offer advice that the judge will prefer which is the opposite of what judges do who follow their own preferences (Kray, 2000). It was also found that the advisors will focus on the best alternative for market value whereas the judge will tend to equally weight all the differing attributes (Kray and Gonzalez, 1999).

Kray (2000) reported that advisors are motivated by accuracy of their own advice relative to the decision makers and show higher task related effort.

2.11. Forecasting and Decision support systems (FSS, DSS)

Forecasts that are produced with a combination of judgment and quantitative forecasting can be produced within a forecasting support system (FSS). There is evidence that this combination is often not optimal and has negative effects on forecast accuracy (Fildes et al., 2006).

A FSS should include two key components. To support a forecasters ability to know when to include judgment and to enable an accurate intervention when this is appropriate (Fildes et al., 2006).

To help the forecaster in the above tasks Silver (1991) discusses the terms ‘restrictiveness’ and ‘decisional guidance’. The restrictiveness of a system would limit the possible processes available. For example, in the cases where a statistical forecast was very good it may prohibit the use of judgmental adjustment. Guidance is explained by helping users put together and implement decision making processes by helping them choose the components.

We can describe a typical time series as having the following components: regular patterns (trends, seasonality or other stable relationships), irregular components (product promotions for example) and noise which is totally unpredictable. If the FSS was working, where there was data where statistically a regular pattern could be identified then this should be shown in the FSS model. Any irregular parts should be forecasted judgmentally. However, the evidence shows that FSS use is not ideal and the two components are usually confused. Statistical forecasts being adjusted when they did not need to be (Goodwin and Fildes, 2007, Goodwin and Wright, 1993) and when non-optimal statistical methods are chosen by the forecaster (Lawrence et al., 2006). The diagram below illustrates how a typical FFS would work.

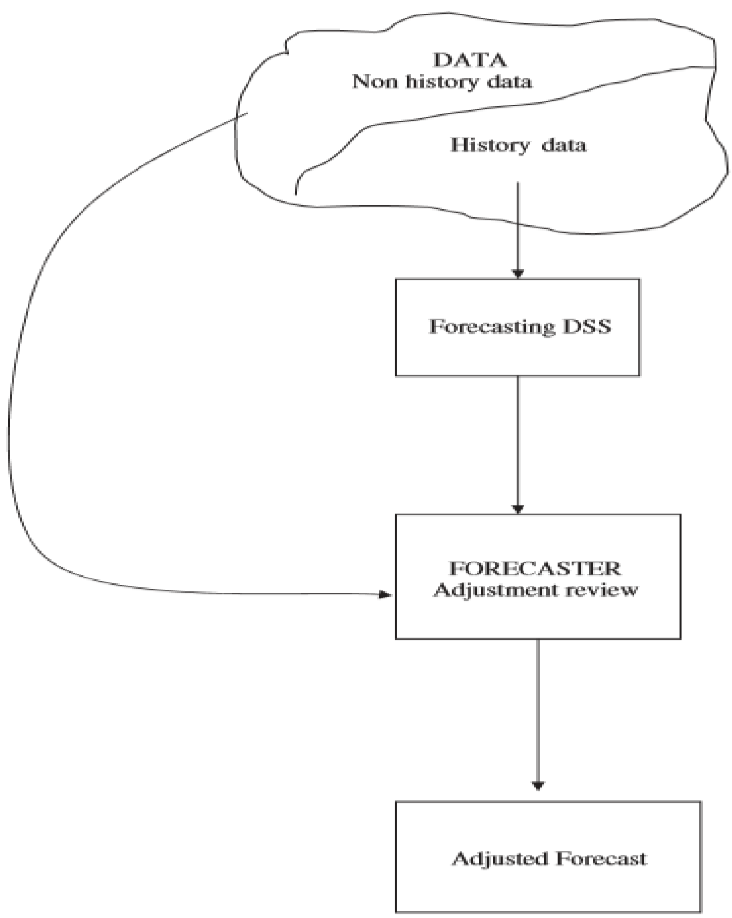


Figure 2.7. Forecasting steps

If the perfect balance could be struck between the FSS's statistical model and the application of human input, then the forecast would be more accurate. Firstly, noise would be filtered out by the statistical model (O'Connor et al, 1993) and secondly the forecaster's attention will be wholly focused on the irregular component of the data (Jones and Brown, 2002).

Fildes et al., (2006) argued that the ideal FSS system will have the following attributes:

- It will be acceptable to users
- It will offer a range of methods
- It would be commercially viable
- It would support the appropriate mix of judgment and statistical methods.

They also thought that restrictiveness would probably not allow the above elements to provide a FSS. It would lead to ease of use, but it would reduce the flexibility component of the requirements. For instance, the display of the time series could be restricted to tabular only and not allow for graphical display. This could halt any heuristic that would cause inaccuracies (Harvey and Bolger, 1996). Similarly, if the forecaster had limited understanding of statistical methods then the system could choose a specific technique (Fildes and Hastings, 1994). If the forecaster was required to make more effort to use specific methods then this also can be used as a restriction as often forecasters use the path of least effort when choosing the method, they will use (Payne et al., 1993). This could also reduce the level of judgmental adjustment as Goodwin (2000) showed when highlighting that extra effort correlated with reducing the level of harmful adjustments in this case by requesting for the adjustment. Alternatively, by easing the level of restrictiveness the forecaster can be led to the more desirable strategy. Simply by making the quantitative processes easier to use than the judgmental adjustment ones will reduce the level of adjustments. Fildes and Beard (1992) wrote that by designing easy to use facilities to choose the optimal quantitative forecast the following facilities should be used:

- Exceptional and missing observations are easily adjusted in data series
- Users could identify series by type such as trend or new product
- Enable the techniques to be compared in a test area
- Allow forecasting at SKU level and at an aggregated level
- Create menu structures that lead users to more appropriate techniques
- Highlight times series where it is appropriate to judgmentally adjust by reporting extreme errors

In contrast to restrictiveness, they thought that guidance would have the possibility to support all the elements successfully. It would be able to overcome problems users may find when confronted with statistical options and enhance user acceptability, whilst also being highly marketable. It would also be able to direct users where to use judgment in the appropriate areas. They noted that there were at the time no commercial packages that offered the correct components on the market.

The literature describes two types of FSS guidance. Decisional guidance can be defined as where guidance is intentionally built into a system. This can be in the form of informative guidance where a user is given information without any suggestion as to what is better or worse, or it can be suggestive guidance where a course of action would be proposed.

Informative guidance can be in the form of providing a record of previous strategies that users had already used as a form for memory support (Singh, 1998) which was shown to improve decision making. Guidance of this sort directed at the forecasting task was also shown to be of use. Parikh et al., (2001) reported that forecasting method selection was improved when information was provided regarding an explanation of when each technique was appropriate.

There are numerous reports of suggestive guidance having positive effects (Parikh et al., 2001) but there are also reports that point to the miscalibration that can be found if no informative advice is also given. Calibration means the level of confidence that follows the objective accuracy of the choice. By choosing a technique user's will be showing their confidence in the accuracy of alternatives. Miscalibration would be where because of overconfidence, a non-optimal choice is made, and other more accurate methods are forgone. This can be seen in the many reports of users making judgmental adjustments to quantitative forecast which result in less accuracy (Goodwin and Fildes, 1998, Lim and O'Connor 1995).

2.12. Feedback

To mitigate bias in forecasting feedback can be used to help forecasters (Lawrence et al., 2006). We can see the use of feedback in other techniques such as the Delphi method as reported by Rowe and Wright (1999) where feedback relates to the judgment of other experts. When forecasting with time series it has also been shown that feedback can improve accuracy of point forecasts (Goodwin and Fildes, 1999).

The simplest form of feedback is outcome feedback. Here the forecaster is informed of the outcome of the variable forecasted. This allows comparison of each forecast over time showing the forecast accuracy which could help the forecast improve over time. Klayman (1988) showed that this learning is slow. One reason that this could be attributed to is that each outcome would contain an amount of noise. If the forecaster looked only at a single outcome (rather than several outcomes over a set of periods) then this could lead to an overreaction in the next forecast. Outcome feedback is however easy to provide and understand and not contaminated by earlier observations which could also contain irrelevant observations (Goodwin et al., 2004). Experienced forecasters would expect to see some noise and would bear this in mind when comparing the outcome to their forecast.

Another type of feedback is cognitive feedback. This aims to provide forecasters with insights into their methods enabling them to reflect and improve (Remus and Lim, 2005). It has been reported that cognitive feedback is more effective than outcome feedback where the situation is complex with many nonlinear relations between criteria (Balzer et al., 1992). Sanders (1997) studied the impact task properties feedback on time series judgmental forecasting tasks by testing the impact of providing information on time series data patterns the degree of noise to experienced forecasters. The accuracy of both individual and group judgmental forecasting was improved when information was provided. The improvement was higher where there was particularly high noise in the series. It was thought that this could be important for statistical software packages which summarize information as it may be a good input to judgmental forecasting.

Sanders found the best forecasting method to be a combination of judgmental and statistical forecasts. Aggregation of judgmental forecasts was shown to be better than individual judgmental forecasts. Where a quantitative forecast was not available this would be the better alternative.

By using a rolling training approach to improve judgmental forecasting, Petropolous et al., (2017) found that participants in their experiment could improve accuracy. The experiment asked forecasters to make multiple judgmental forecasts for a set of time series from different time origins. The forecasters were either provided with feedback or unaided. When the outcome plus performance feedback was given the forecasters were enabled to develop and better understanding of the patterns in the data by learning continually. It was found that the longer the horizon and the higher the noise of the series produced the best improvements. It was also stated that the participants in the experiment were less confident of their own forecasts.

Certain types of feedback may be more appropriate for different elements of the forecasting tasks. Stone and Opal (2000) found that the calibration of probability forecasts was only improved when performance feedback was given. Similarly, task properties feedback was good at helping forecaster’s discrimination (to know when an event will occur) and made calibration worse. The effectiveness of a type of feedback may depend on the forecasting task itself (Fischer and Harvey 1999).

The feedback for any type must be in a format that is understandable for the receiver and in a format, that is optimal. Lim et al., (2005) found that presenting feedback in a multimedia display was less effective than sending a text message. This was said to be because the cognitive resources required to understand the text were like the resources required to improve the decision.

2.13. Experts and judgment

"The expert has difficulty explaining the basis for their intuition other than gut feel" Blattberg and Hoch (1990).

The paper lists the specific areas of weakness and strength of both experts and models.

Models are stronger than experts:

- Experts have decision bias both in perception and evaluation
- Experts suffer from over-confidence and can be influenced by politics
- Experts can tire or get bored
- Experts are not consistent

Experts are stronger than models:

- Models are reliant at what the expert has told it. Experts can identify new variables. Models only predict where experts can predict and diagnose
- Experts can subjectively evaluate difficult variables whereas models are only objective
- Models are rigid and not flexible

- Models do not have domain knowledge

The results from the experiment using a 50/50 split of database models plus managers was trying isolate the managerial intuition. The analysis concluded that the model plus manager forecasts are complimentary. Bringing together information that increase accuracy when used together. The ability of managers to adapt to new variables helped stabilize the models whilst the models provided a consistency to compensate for the more inconsistency in human judgment.

In the experiment 25% of the variance that was unexplained from the models was picked up by the managers. Where the intuition comes from is unknown but given the significant improvement it would be unwise to ignore it and rely purely on the statistical model on its own.

It was thought that due to the capacity for human processing and the number of new data sources managers should move away from using intuition as the only basis for decision making. The 50/50 model and managers decision heuristic is certainly non-optimal but it is a pragmatic solution. Blattberg and Hoch (1990) saw three advantages: as managers do not need to understand and develop models this separation should continue; managers should continue to have control over the decision-making responsibility; the combination of the model plus the manager will be more accurate than the individual inputs.

If managers could work with modelers to identify the basis of information for exceptions, then any refinement to encompass this would improve the model.

Further to this report there have been a lot of literature investigating adjustment of model-based forecasts by experts. There have been papers showing that experts do adjust model-based forecast (Sanders and Ritzman, 2001; McCarthy et al., 2006). Klassen and Flores (2001) reported that the amount of influence on the final forecast was significant. The impact to the accuracy of the forecast from expert adjustment was seen to be positive (Syntetos et al., 2009: Matthews and Diamantopolous, 1986).

The fact that statistical models cannot capture all the relevant variables and be aware of any future changes means that domain knowledge will always have some insight that the models do not. Franses (2008) agrees that incorporating expert knowledge into model-based forecasts would be beneficial. Despite the weight of evidence that supports expert knowledge being included in a forecast the report by Armstrong and Collopy (1998) shows that this is often not the case is pertinent alongside their comment that experts rarely say what they do.

The issue of forecasting SKU level sales data where a statistical model was adjusted by experts was studied by Franses and Legestree (2013). Here an analyst (manager of the experts) who has access to statistical model, the expert forecasts and the outcomes was tasked with suggesting ways to improve them. This could be in the form of suggesting changes to the model design or to the experts and their judgmental adjustments.

They examined whether expert knowledge could be included in the statistical model. If it could be incorporated as an explanatory value, then this could help in increasing acceptance of a forecast support system (Rangaswamy and van Bruggen

2009). In their conclusion, they stated that including the past expert forecasts and the difference between the expert and model forecasts, which could be the variables which were missing in the models, made the model forecasts worse. This suggested that the FSS in the data set used (which was a large pharmaceutical data set of over 1000 SKU's) did not lack a set of variables which were suggested by the experts. Indeed, the results which were worse could indicate that the variables themselves were redundant.

However, when the model and expert forecasts were not performing well a systematic improvement could be found when past judgment was included. The results indicated that the SKU level forecasts can be biased systematically due to omitted variables and the experts due to domain knowledge are aware of this. The experts place too much weight on their own expertise which harms the quality of the forecast. It would, as a form of guidance (Fildes et al., 2006) in the FSS be useful to include the indication of the performance of the past FSS model and the part expert forecasts. In this case the company could have a rule which was if the past forecasts are not accurate (both model and expert) then change model. The experts would feel that their expertise was being considered if it is adding to the forecast accuracy whilst the weight of the added expertise could be reduced for the next forecast. The ability to bring past performance into the FSS and switching the input variables if necessary is that is bring some balance into the interaction between experts and models. Given that the literature shows experts put too much weight on their own input (Sanders and Manrodt, 1994) then this new strategy could improve forecast accuracy.

2.14. Character of Judgment adjustment

It was found that forecasters were more effective when making larger adjustments than smaller ones (Diamantopoulos and Mathews, 1989). This could be related to the fact that large adjustments are often related to specific events outside the data available. If information that was reliable (company pricing policy for example) was available, then this sort of event that will have a significant effect would not be included in any quantitative method of forecasting. Because of this it would be expected that the accuracy for the forecast would be improved by such intervention. This is also shown by Sanders and Ritzman (1992) when examining the effect of time series on forecast adjustments. Here the series with the most volatility were seen to much better judgmentally adjusted than the series with a more constant character. Again, this could be because of significant information available outside the time series meaning larger adjustments were more accurate.

Using the same rationale, smaller adjustments will be made if the information was less reliable. If this was the case, then forecasters would probably like to be more cautious and adjust less. Smaller adjustments have also been shown to be the result of human's desire to adjust even when given excellent statistical forecasts (Lim and O'Connor, 1995). This tendency to adjust statistical forecasts was also reported by Yaniv (2004) where users again discounted good advice. The literature show that forecasters make small adjustments to statistical forecasts even if there is no logic and the subsequent result is a reduction in accuracy. Fildes et al., (2009) commented that from observations and discussions with forecasters that this tendency to adjust (tinker) is down to the employees wanting to show they are reviewing forecasts and doing their job.

The question of whether the direction of adjustments has any bearing on the accuracy of the forecast was tested by Syntetos et al., (2009). The research focusing on intermittent demand at a pharmaceutical company showed conclusively that negative adjustments performed better that positive ones. Positive adjustments were found to be less accurate and were explained by optimism bias or pressure from senior management who wanted to see increased sales. In the company where the test was taken there was also evidence that the forecasters were trying to ensure priority from suppliers within the supply chain.

In summary, the literature finds that large adjustments generally have a positive effect on accuracy whether upwards or downwards. Whilst small adjustments not as effective negative ones add to accuracy more than positive ones.

2.15. Judgmental bias

In 1981 Hogarth and Makridakis stated that a forecaster can see a pattern in data even when one does not exist. Lawrence and Makridakis (1989) also reported that a forecaster can see random variability where the data is very stable.

When discussing bias, the literature points to a human's personality, work locus of control and motivation as affecting individuals work performance and cognitive processing (Eroglu and Iles, 2010). Evans (1992) defined bias as a systematic deviation as opposed to a random one.

In a study into how individual differences effect biases in judgmental adjustments Eroglu and Croxton (2010) focused on three kinds of bias:

- Optimism bias which describes a tendency for forecasters to adjust forecasts upwards
- Anchoring bias which refers to how forecasters do not deviate from an anchor value despite forecasting the direction of the necessary adjustment
- Overreaction bias which happens when the size of an adjustment although correct, creates a larger error.

Personality was described as a set of traits that drives an individual's behavior consistently over time (Levy et al., 2004). Judge and Iles (2002) described personality with respect to five traits: conscientiousness, neuroticism, extraversion, agreeableness and openness to experience. It is suggested in the literature that personality can influence judgment and decision making via information processing and affect (mood-states). Epstein (1994) suggested that individuals reach decisions via the interaction of two information systems. One rational, controlled and conscious and the other emotional, intuitive and experiential the former being more cognitive driven than the latter.

A person's work performance and cognitive performance can be affected by motivation (Amabile and Kramer, 2007). Motivation can come from different areas for different people. Extrinsic motivation refers to people who are motivated by social standing or by rewards for performing better when working towards something that is external to their work. Others are motivated by the work itself which they find challenging or enjoyable which is intrinsic motivation (the two types not being mutually exclusive). Studies find that extrinsic motivation has a detrimental effect on critical thinking if the task is complex and engaging (Cheung et al., 2001) and further studies have shown that problem solving, and creativity are also negatively affected (Amabile, 1995). However, for intrinsic motivation the results are different. A positive relationship was found between intrinsic motivation and cognitive performance by Walker et al., (2006). Similarly, challenge seeking was seen to improve behavioral performance and thus outcome (Maio et al., 2009).

If a person believes that rewards and outcomes are governed by their own actions, then this is termed internal work locus of control. If on the other hand a person believes that external forces are in control and outside of their influence, then this is termed external locus of control. This idea was first discussed by Rotter (1954, 1965). It is suggested that locus of control is important in relation to motivation and work performance (Erez and Judge, 2001). Hough (1992) argued that forecasters with an internal locus of control would improve faster.

In their findings Eroglu and Croxton (2010) found that from an optimism perspective, experience and a challenge seeking mentality increased the level of judgmental adjustment whilst external locus of control decreased it. When analyzing anchoring bias, they found the experience within the current position was a factor in increasing judgmental adjustment. If a forecaster is college educated or has experience in the current position and was challenge seeking, then the amount of judgmental adjustments was increased with an overreaction bias. If an individual had external work locus of control, then the level of adjustment decreased.

The implications for management suggested were that there was a relationship between a person's characteristics and their susceptibility to forecasting bias. This could be of use when assigning task to individuals within an organization. The assessment of an employee's characteristics could be included in the variables when deciding who should be given a task involving judgmental adjustment. If a person was identified to have certain bias traits, then it could be useful to make them aware of the implications. Feedback on a routine basis may also be worthwhile in order that forecasters can check to see if they are indeed able to keep certain biases in check (Lawrence et al., 2006).

The main driver that indicated whether a forecaster made judgmental adjustments was experience. Once a decision to make an adjustment was made then the personality and motivational orientation were the best indicators of the level of bias.

2.16. Judgment and spare parts inventory

In their investigation into spare parts demand forecasting Bacchetti and Saccani (2012) described the difficulties that are inherent in their demand and thus inventory management. The high number of parts, the occurrence of intermittent and lumpy demand, the importance of having availability due to possible downtime and the risk of obsolescence were highlighted as specific issues.

Although fast moving parts can be suitably forecasted using time series methods the slow-moving SKU's will be characterized by intermittent and lumpy demand which requires special attention (Boylan and Syntetos, 2010). When a moving average is used for intermittent demand has been shown not to be optimal. Syntetos and Boylan (2005a) when updating a forecasting technique created by Croston (1972) by reducing bias, showed that using their technique forecast errors were reduced in comparison to an exponentially weighted smoothed forecast (as well as Croston's technique). The added complexity of intermittent demand and the statistical models available will support Lawrence et al., (2006) in that forecasters will most probably be unaware or unwilling to try the method in the real world resulting in less accuracy and thus inventory problems in the form of unavailability or over stocking.

We know that many quantitative forecasts are judgmentally adjusted (Sanders and Manrodt, 1994) but there have not been many studies in a practical environment and particularly focusing on fast and slow-moving items. Syntetos et al., (2009) argued that research did not consider integrating judgment with statistical forecasts for slow and intermittent items. Furthermore, no study has looked at the implications of these judgments on inventory control. The investigation into the intermittent demand forecasts of a pharmaceutical company considered whether the judgmentally adjusted statistical forecasts using input from intelligence from forecasters had benefits to the forecast accuracy and the inventory control.

The results found that judgmental adjustments can have a positive effect on forecast accuracy when applied to intermittent demand series. In concurrence with previous findings the character and size of the adjustments did effect whether the adjustment was improving accuracy. The findings supported Fildes et al., (2009) in that the positive adjustments were less accurate than the negative ones perhaps displaying optimism bias.

It was also seen that the results did not improve over time showing a lack of learning which was consistent with items that were not intermittent in their demand type (Nikolopoulos et al., 2005). It was thought that given the intermittent nature of demand it would be unlikely that a feedback system would improve the accuracy of the forecasts.

The results also showed that the improvement in forecast accuracy did influence the stock control performance. Items that were adjusted were reported to be closer to the service levels and targets required than the system forecasts.

When summarizing the current practitioner orientated method of forecasting practice when dealing with spare parts Bacchetti and Saccani (2012) note there is no conclusive evidence of best practice. There are some studies that do not support Croston or its variants as the best method for forecasting intermittent demand (Willemain et al,. 2004) for example, advocates bootstrapping. The studies that are available do not have a relevant amount of data to give a conclusive answer. The results of many studies are based on the performance accuracy per period and this may not be the best method to test series with many zero demand periods. Teunter and Sani (2009) suggest inventory and service levels as better accuracy measures. The differentiation of techniques is an issue also raised. Whilst many studies focus on lumpy demand Syntetos et al., (2005) propose using different methods per demand-based classification: simple exponential smoothing for smoother items and Croston or a variant to the more intermittent ones. Baccheti and Saccani (2012) observe that there have been very few studied which dealt with the practical applicability for spare parts management. User skills and support systems, availability of data and implementation have been overlooked in the academic research.

There is some research that suggests that after sales services do not receive the amount attention that it should from companies. The level of services in many companies was described as poor despite half a century of research by Wagner (2006). In a further study by Wagner and Liedermann (2008) it was commented that there was a lack of awareness invested by senior management despite the significant contribution to profits in the spare parts area of business.

2.17. Conclusion

The literature review highlights a gap in the literature regarding the experiments that have underpinned many of the findings regarding judgmental forecasting. The papers that have been reviewed have used data from sources such as: surveys, retrospective studies in the field and experiments usually with generated data. The participants have been from different groups: undergraduates, experts, senior managers and forecasting managers. None of the experiments have been real time and the participants have usually not been the real-world managers of the data which itself is usually generated or at best months old when analysed.

That this is the case does cause issues regarding the validity of the results. The fact that this PHD uses forecasters who are working with the data that is their real-world task in real time means the results are more valid than using a group of undergraduates looking at generated data for example. This is underlined by the fact that all the SKU's which are in the experiment are "live" SKU's and 1 of the company-based participants will use the forecast to produce and Order Up To level (OUT). This depending on the stock level and purchase orders already placed may generate a purchase order.

The company participants will also better reflect a real-world environment outcome as this is a task that all of them have undertaken monthly for between 2 and 20 years. For example, an undergraduate forecasting with generated numbers is aware that any error will not result in real word consequences such as back orders and overstocking. They are also not drawing from the experience gained from forecasting this way over years of employment with the additional insight that this adds.

In their research Alvarado-Valencia and Barrero (2014) summarised the research into judgmental forecasting showing an exhaustive list covering the years 1994 - 2014 in their table below. It shows clearly what the different data collection / analysis methods and the population / forecasting information sources were for the papers focused on judgmental forecasting during this period.

Methodological characteristics in reviewed papers.

Manuscript	Data collection and analysis methods	Population/forecasting information source	Sample	Factors evaluated on response variables	Response variables
Sanders and Manrodt (1994)	Survey; Descriptive analysis	Sales forecasting responsible on US companies	96		Methods familiarity, usage, and satisfaction
Sanders (1997)	Survey; Descriptive analysis and statistical tests	Forecasting responsible on US Manufacturers	86		Methods & software usage, problematic areas in forecasting
Mentzer and Kahn (1995)	Survey; Descriptive analysis and statistical tests	Forecasting executives on US companies	207	Methods familiarity, usage and satisfaction	Accuracy and forecasting performance
Sanders and Manrodt (2003)	Survey; Descriptive analysis and statistical tests	Marketing heads on US companies	240		Accuracy, methods usage, information/software access & use
Smith and Mentzer (2010)	Survey; Structural equations	Senior responsible of forecasting on companies	216	Forecasting procedures, support systems quality and access	Performance (accuracy)
(Winklhofer & Diamantopoulos (2002)	Survey; Structural equations	Sales forecasting responsible of UK exporters	180	Accuracy, bias, timeliness, cost	Effectiveness
Gonul et al. (2009)	Survey; Descriptive analysis and statistical tests	Top- management executives in Turkish companies	124	Accuracy, expertise, explanations, source quality	Perceptions, expectations and adjustment behavior of provided forecasts
Lawrence, O'Connor, and Edmundson (2000)	Field retrospective study	Australian companies/ products	13/304	Company, not intermittent demand	Accuracy, bias and inefficiency
Fildes et al. (2009)	Field retrospective study	UK companies/SKU	4/1536	Company, not intermittent demand	Accuracy, judgmental adjustments, bias and inefficiency
Syntetos et al. (2009)	Field retrospective study	UK companies/SKU	4/138	Intermittent demand	Accuracy, judgmental adjustments, learning and stock control
Trapero et al. (2013)	Field retrospective study	Manufacturing company/SKU	1/169	Promotions, forecast type (judgmental vs. statistic)	Accuracy, bias
Franses and Legerstee (2011)	Field study	Experts by country/SKU	37	Age, gender, position experience, number of products.	Optimal judgmental and statistical combination for accuracy
Franses and Legerstee (2013)	Field study	Product by country	1078	Statistical and judgmental forecasts	Accuracy
Lawrence and O'Connor (2005)	Experiment. Reward: Individual, by performance	Graduate students/M3 competition	100/10	Loss function symmetry, loss function kindness, and time	Learning asymmetric losses
Lim and O'Connor (1995)	Experiment. Reward: Best takes all by treatment	Graduate students/M competition series	64/2	Support type, seasonality, feedback and time	Accuracy improvement over initial judgment
Lim and O'Connor (1996)	Experiment. Reward: Individual, by performance	Graduate students/Generated time series	40/1	Support (statistical and causal)	Accuracy improvement over initial judgment
Lawrence and O'Connor (1995)	Experiment Reward: None	Undergraduate students/M competition	Not reported/	Forecasting horizon	Presence of anchor and adjustment
Goodwin et al. (2007)	Experiment. Reward: Equal for the 50% superior	Undergraduate students/ Generated time series	32/20		Forecasting process
Arnott and O'Donnell (2008)	Experiment Reward: Best takes all	Graduate students/Product	178/1	Support	Accuracy
Lee et al. (2007)	Experiment. Reward: All, 50% superior even more.	Graduate & undergraduate students/Generated problem	54/1	Support, noise, promotion level.	Accuracy
Goodwin et al. (2011)	Experiment. Reward: None	Graduate & undergraduate students/Generated time series	130/1	Support, series nature, noise & type of period	Accuracy, learning, time and adjustments
Webby et al. (2005)	Experiment. Reward: Best take all on each treatment	Graduate students/Generated time series	64/16	Support, trend and information load	Accuracy
Lawrence et al. (2002)	Experiment. Reward: None	Undergraduate students/ Generated time series	180/20	Participation	Confidence, satisfaction and accuracy
Gupta (1994)	Experiment. Reward: Individual, by performance	Graduate students/Generated problem	80/1	Support, task and situation	Accuracy
Goodwin (2005)	Experiment. Reward: Equal for the 50% superior	Students/Generated time series	60/10	Support, asymmetry	Time, cost
Edmundson (1990)	Experiment. Reward: None	Graduate students/M competition	38	Support, experience, forecasting method	Accuracy and time
Onkal et al. (2012)	Experiment	Graduate students groups/ Generated time series	13/20	Role playing	Accuracy and adjustment behavior

Goodwin et al. (2013)	Experiment			Noise, Interval format, trend, Provision of point forecast	Trust and adjustment behavior
Jiang and Muhanna (1996)	Experiment	Students- managers/ Generated forecasting problems	75–48/ 12	Weight and strength of past performance; type of subject	Deviations from normative behavior
Onkal et al. (2009)	Experiment	Undergraduate students/real stock time series	76/30	Source of advice	Adjustment behavior
Önkal et al. (2008)	Experiment	Undergraduate students/ generated time series	214/18	Previous adjustments information; explanations	Adjustment behavior
Gonul et al. (2006)	Experiment	Undergraduate students/ generated time series	116/30	Explanation length; explanation confidence	Adjustment behavior
Goodwin et al. (2010)	Experiment	Undergraduate students/ generated time series	56/10	Type of statistical advice; asymmetric loss type	Deviations from normative behavior
Harvey and Harries (2004)	Experiment	Undergraduate students/ generated time series	144/144	Task order; previous forecast label	Accuracy and bias
Goodwin and Fildes (1999)	Experiment	Undergraduate students/ generated time series	48/4	Series complexity; noise level; cue reliability; type of advice/feedback	Accuracy and deviation from normative behavior
Jones et al. (2006)	Experiment	Undergraduate and graduate students/generated forecasting problem	270/20	Information provided; cue reliability; time spent in non-modeled information, support type	Accuracy and adjustment behavior
Reimers and Harvey (2011)	Experiment	General population/ Generated time series	2982/12	Noise; autocorrelation; carry-over effects; series label	Sensitivity to autocorrelation
Harvey (1996)	Experiment	Undergraduate students/ generated time series	104/42	Information format; trend type; noise	Accuracy and bias
Thomson et al. (2013)	Experiment	Undergraduate students/real exchange rates	86/28	Trend strength/trend direction/ manipulations salience	Accuracy and bias of direction and forecast; consistency
Harvey (1995)	Experiment	Undergraduate students/ generated time series	40+/4+	Noise/frequency of time series/task type	Accuracy; consistency

Table 2.8. Methodological characteristics in reviewed papers

The fact that there are no real-time experiments reflects the difficulty in arranging such research whilst participants are full time workers. The additional factor of extending the experiment over 12 months makes this even more onerous for the participants to donate the time required to take part. This PHD utilizes the opportunity to run an experiment with actual data that the participants are forecasting within their jobs. Thus, the impact and additional workload was minimized by this fact and that the process used was identical to what they were using.

The papers that have been discussed are primarily areas that have an influence on judgmental forecasting and its results. There are many that are reasonably old which is a reflection on the recent level of research in the PHD's specific area.

3. Methodology

3.1. Introduction

The term methodology refers to the theory of how research is undertaken. This contrasts with the term "research methods" which refers to techniques which can be used to obtain and analyse data. The research method is only applicable when the methodological approach has been decided and the research methodology has come to a consensus.

The first section off the introduction will discuss the underlying philosophical considerations which underpins the research. Following on, the approach taken to the research and its context and limitations will be discussed in the second section. The third section will look at the strategy and design of the thesis and the research methods used. Data collection and analysis is discussed in the fourth section and in the final section the ethics, reliability, validity, generalizability and limitations will be examined.

3.2. Philosophy

The philosophical approach for this thesis is positivism. Described by Remenyi et al. (1998) as "*working with an observable social reality and that the product of such research can be law like generalisations like those produced by the physical and natural scientists*". Gill and Johnson (1997) state that positivism has an emphasis on a structured methodology to facilitate replication and has quantifiable observations that allow statistical analysis.

Positivism aims to mirror scientific method using deductive reasoning, empirical evidence and hypothesis testing. Bryman (2001) described positivism as an epistemological position that uses natural sciences to study social reality.

Positivistic philosophy within management research because of this scientific approach is usually associated with quantitative methods and data collection. The researcher would attempt via the data to explain relationships between variable (Saunders et al., 2003). Positivism holds a deterministic philosophy in which causes effect outcomes and thus many problems studied in a positivistic way show the need to identify and understand the causes that influence outcomes in experiments.

The key assumptions of positivism were outlined by Phillips and Burbles (2000) as: knowledge is conjectural and absolute truth can never be found thus researchers try to indicate a failure to reject a hypothesis; research is the process of making claims, data, evidence and considerations shape knowledge; research seeks to develop true statements; being objective is an essential part of competent enquiry (validity and reliability are important questions in research).

This need to explain causal relationships between variable makes data collection through samples of enough size an appealing approach when law-like conclusions are sought an appealing approach with the researcher an independent observer.

The experimental (positivist) approach has a realist ontology (we can discover reality) and an empiricist epistemology (I will collate data to find the reality).

The participants of the experiments are given a clear target. To forecast the future requirement of the SKU's whilst keeping stock to a given inventory target (based on the ABC classification of the SKU). This stock target is included in the statistical forecast when calculated. There is no need to interpret these requirements as they are not open to debate. The participants adjust the statistical forecast (which is the same for all of them) and then have the chance to make a final adjustment before the forecast is completed. The results of all the changes are logged each month and then the process is repeated for 12 months.

All participants have a spreadsheet which is updated with new demand each month. The spreadsheet creates an alternative reality for all but 1 of the participants for each SKU. This is because 1 of the company participants forecast will be the one used to generate the companies purchase requirements. The academic participants forecasts will not be used for the company forecasts. For each of the participants whose forecast is not used by the company the "alternative" forecast is input into the spreadsheet and this is then rolled forward to the next month. Unless the participants forecasts are the same then the order up to level will be different each month which could produce purchase orders if stock were not enough.

Each month the accuracy, direction and size of the forecast was measured, and the level of inventory was noted. The positivistic nature of the reporting was specific. Each participant made an adjustment to the forecast (or not) and this had an impact the given statistical forecast. Similarly, the final adjustment.

3.3. Approach

This thesis will be using the experiment research strategy. This is a classical form of research owing much to the natural sciences. Sanders et al. (1997) described this approach as typically involving: a definition of a theoretical hypothesis; samples of individuals from a known population; allocation of samples to different experimental conditions; introduction of planned change on one or more of the variables; measurement on a small number of the variables; control of the variables.

The experimental design was said to be *the* approach for obtaining information about causal relationships (Robson, 1993). It allows researchers to investigate the relationship between one variable and another. The design in principle allows one variable to be altered to see if it has any impact on others. The variable that is changed is known as the independent variable and the change that results from the that is the dependent variable.

In this thesis, the design can be described as a longitudinal time series experiment. The experiment consists of a time series for 90 SKU's over a 12-month period. The participants are not randomly chosen but are members of the company's inventory planning department responsible for spare parts procurement and inventory.

As this is a real world (demand is real for each month), real time (the participants receive the statistical forecast in real time), experiment, there is no manipulation of the data it is purely the real demand for the time buckets (4 weeks) which the participants are receiving and deciding if they want to adjust the statistical forecast which in turn will produce a purchase order (depending upon the order up to level for each SKU). Due to the design of the experiment there is no placebo as with medical trials, for example, the demand is real and there is no placebo time series to compare against.

Using the research process onion (Saunders et al., 2003) the approach could be described from outer ring to center as: Positivism, Deductive, Experiment, Longitudinal and Data collection. The thesis is positivistic (quantitative) by its approach, deductive in that we are testing a theory (do judgmental adjustments improve a statistical forecast?), an experiment in its design, longitudinal to the extent that the experiment was not cross sectional but over 12 months and data collection as the findings are based on the measurement of forecast accuracy and inventory values over the period.

There is an inductive element to the research as from the literature and the knowledge in the area which is where the PHD starts, information is collected in an inductive fashion to allow reflection on what is already known in this area. This could be termed semi-inductive as move from the theory to the data and then back to the theory to amend it.

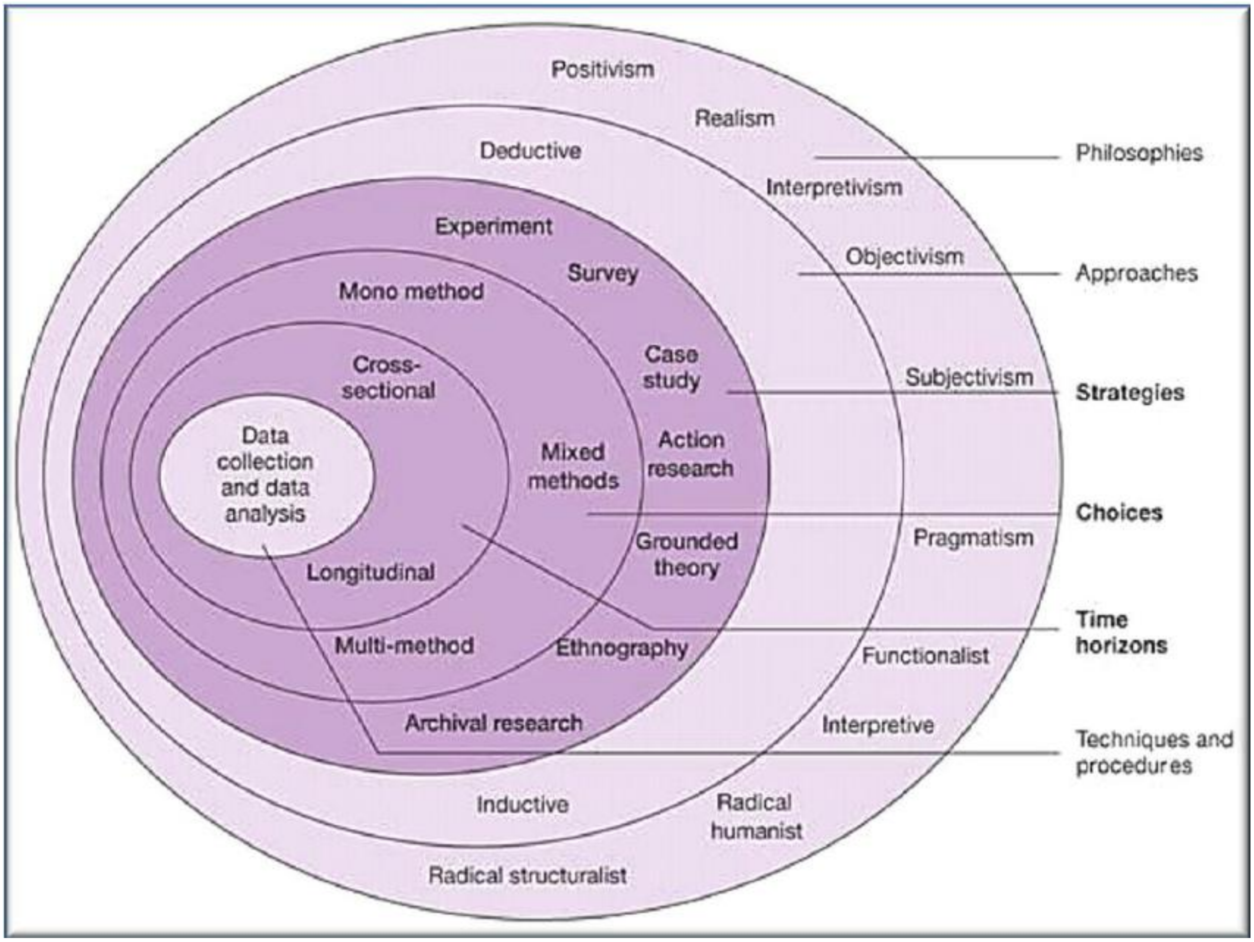


Figure 3.1. The research process onion

3.4. Research design

The thesis will empirically investigate what the implications are for forecast accuracy and inventory when statistical forecasts are judgmentally adjusted.

The experiment consists of 6 participants. 5 of the participants come from the company department which controls the SKU's included in the study. The manager is the PhD candidate. One participant is a University Senior academic (Chair in

Operational Research and Operations Management). As the participants were not randomly chosen this can be called a quasi – experiment.

- 6 participants
    - Mark - Manager (Researcher/ long experience / male)
    - Michelle - Senior inventory controller (Degree level / short experience / female)
    - Dominic - inventory controller (long service / long experience / male)
    - Dave - Inventory controller (no qualifications / medium experience / male)
    - Jane - Inventory controller (no qualifications / short experience / female)
    - Professor - University senior academic (male)

The participants are listed in order of academic qualifications firstly then by experience within the forecasting department. It is noted that Michelle who was a graduate is listed above Dominic who has greater experience. This is not a definitive indication of the probable forecasting accuracy performance (that is academic qualification may not be as productive as experience).

The total of stock keeping units in the experiment is 90. They have been chosen to reflect different time series and value characteristics. This is because they may produce different judgmental forecasts from the participants. For example, with either of the trending time series the way this is perceived and acted upon will be different between the participants. This could affect the forecast accuracy of the judgmental adjustment and affect the inventory targets for each SKU group.

9 groups (10 SKU's per group)

- High frequency / low value
- High frequency / high value
- Low frequency / low value
- Low frequency / high value
- Very low frequency
- New SKU's
- SKU's with a long lead-time
- Increasing demand
- Decreasing demand

The SKU's are forecasted and ordered monthly and require a forecast for the next three months. The period of the study is 12 months long.

The participants will start from the same month and have 24 weeks' history of demand. The company's ERP system (SAP) outputs demand buckets in weekly periods. The Managers participation was known to the group

The SKU's were spare parts of the products the company manufactured. The product range was printers (both for small / medium business and small office / home office), multi-function centres (comprising of printing, scanning and copying functions), personal sewing machines and Labelling machines. Predominantly manufactured in the Far East (China, Vietnam, Taiwan and the Philippines).

Each participant is given a statistical forecast for the next 3 months (based on the second 12 weeks' demand average plus an average of the difference of the second 12 weeks and first 12 weeks to reflect a trend). This was the company's standard forecasting procedure. The participant was then able to judgmentally adjust the statistical forecast for the any of the 3 months if desired. The Managers participation was known to the group. This did not compromise the independence of the experiment as all the participants did exactly what they were doing in the day job. The fact that the Manager also forecasted the SKU's involved was irrelevant to them and the monthly forecasts from the manager or any of the other participants were not available to view until the end of the 12-month period.

Regarding the participation of the manager, it could be construed that this put the participants under extra pressure to produce better forecasts or to possibly judgmentally adjust more than normal. The fact that all the other participants already produced the forecasts as part of their responsibilities as part of their weekly tasks added no time pressure. Also, that the results of the experiment were not going to be scrutinised by the company regarding personal performance (only learning outcomes would be delivered) meant that it was not a competitive scenario either. It was felt that there were no negative impacts from Mark's participation in fact it was felt equitable that he should do the same as all the other participants.

The inventory at the start point of the experiment was the same for all participants for the SKU's involved (reflecting actual real inventory level at the start point of the experiment). The inventory for each future month then changes to reflect purchase order arrivals (including outstanding purchase orders at the time of the start of the study) minus the actual demand for the month passed.

To reflect reality as best as possible, the purchase orders placed by each participant arrive at the same time as the real purchase order for the SKU involved (one of the company participants is placing the real purchase order for the company to negate duplication in their real-world tasks). If no real purchase order exists, then the standard lead-time (average lead-time based on history) for that SKU is used. This could be the case if participant forecasts were higher than the participant who was forecasting for real. Each participant can see the previous 24 weeks' actual demand each month.

The experiment is visually shown in the figure below:

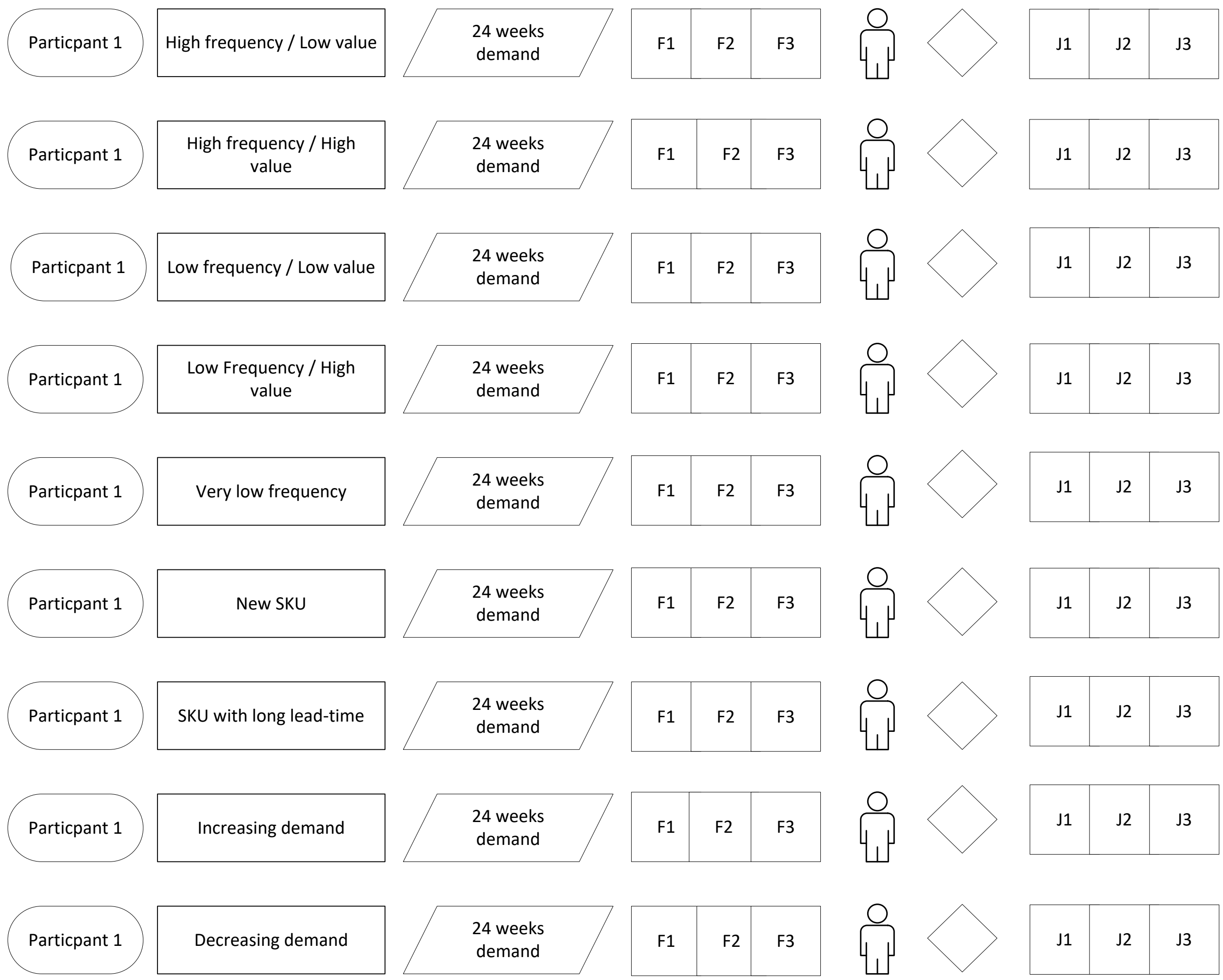


Figure 3.2. Experimental design table – Single participant single month task

As shown in the table each participant each month receives 90 SKU's details (9 sets of 10SKU's). For every SKU, they can see 24 weeks' demand and are given the next 3 months statistical forecast (F1, F2, F3). Each participant is then requested to input their 3 months final forecast for the SKU's (J1, J2, J3). This may be the same as the statistical forecast if no adjustment is made.

The participants were each given an excel spreadsheet with each month presented as in Figure 3.2. The final forecasts were added monthly at the same time for the company participants (the academic participant filled in the spreadsheets remotely).

3.5. Data collection and analysis method

The company data was held in an Enterprise Resource Planning system (ERP). SAP R/3 is a business management system which integrates different functions across a business (Shehab et al., 2004). The system enables companies to have real time visibility of their operations (Gargeya and Brady, 2005). Although the SAP system did have some forecasting functionality the company did not use it due to it being poor when dealing with intermittent demand. Typically, the system was seen to over forecast when demand was preceded by zero demand periods. The decision was taken to forecast all spare parts demand outside the SAP system.

The experiment data was created using extracted files from the company's ERP system to Microsoft excel files. This was updated monthly with the new demand information of 4 weeks' demand and removing the oldest 4 weeks. The demand quantity of the SKU's and the statistical forecasts were presented as quantifiable data. As oppose to categorical data whose values cannot be measured numerically spare parts demand is quantifiable in whole numbers (it is impossible to order half a spare part) and in order quantities of single units (there were no master cartons where a single order unit equates to more than 1pc of the spare part). Data which can be measured precisely is termed discrete data.

The spreadsheet contained 90 SKU part numbers. These were 9-digit codes that were discrete for each SKU. The data which filled the Microsoft Excel spreadsheet was extracted from SAP R3 from different areas (material master, warehouse, sales and inventory). To give the participants a statistical forecast, calculations were added into the Excel spreadsheet which will be explained later.

Breaking the spreadsheet down into sections: SKU information; demand information; forecast information the sheet was set out as follows:

SKU information:

Division	NEW ABC	Material	Description	PRICE (EUR)	CREATION DATE	1 MTH STK	RSL	OUT	Rounding Value

Figure 3.3. Experiment spreadsheet headings

For each SKU, the participants can see the division (e.g. 30) which shows the product division each SKU belongs to.

The next column NEW ABC indicates the ABC categorisation of the SKU. The company used order frequency to rank the SKU's with A items being the highest frequency and C items being the slowest. There were also two other categories. E items indicated that the SKU was a new part number and that the respective product had not been in the market for more than 6 months. Z items were SKU's where there was a close correlation between the demand of the part and the number of machines in the field (MIF) from which they were a constituent, with a live warranty (the company's standard warranty across Europe ranged from 1 to 3 years). For these items, the company used the percentage of MIF to forecast the SKU demand (at the time in a trial period of testing). This technique was available to only one of the participants who had been involved in the development of the trial.

The material column showed the 9-digit code unique to each SKU, followed by the description of the part (maximum of 30 alphanumeric). The price of the SKU was showed in Euro which was the reporting currency. The date the SKU was entered onto the SAP system was shown as the creation date.

Each SKU had a 1-month stock figure. This was based on the sum of the last 12 weeks' demand divided by 12 and multiplied by 4 to indicate 1 month's stock (1 MTH STK). The next figure was the required stock level (RSL). This was calculated by multiplying the single month's stock figure by the stock target for the SKU based on ABC categorization. The company planned to stock 2 months of A, E and Z items (reflecting stable demand for A and Z category and a tentative approach for E items), 3 months of B items and 4 months of C items. So, for example an A item would be the single month's stock figure multiplied by 2.

The order up to (OUT) level in the last column shown was calculated by multiplying the last 12 month's weekly average (used in calculating the 1-month stock figure) by the total OUT level in weeks. The OUT was calculated for each SKU the company ordered (due to the variance in lead-time this was considered necessary). The OUT was determined by summing the targeted stock index, the shipping time (using SF), the supplier's lead-time to pick and pack and an order adjustment factor to reflect that the SKU's were ordered monthly of 2 weeks. The range of OUT's in weeks was between 15 and 35 reflecting the varying stock index targets and the supplier's historical ability to supply quickly or not.

The company when ordering spare parts liaised with suppliers to facilitate easier picking and packing and subsequently could indicate rounding values which participants could use. This was also for ease of goods receipt at the European warehouse.

Demand information:

Legacy PO Arrivals	PO Arrivals	Stock-Sales Orders	Purchase Orders	Port of Loading	2013.WEEK 30	2013.WEEK 52	2014.WEEK 1

Figure 3.4. Experiment spreadsheet headings (continued)

For the experiment, it was necessary to calculate the quantities for each SKU that were already placed on purchase orders and had arrived that month. The column 'Legacy PO Arrivals' indicates the quantity that arrived in the month from these orders. The 'PO Arrival' column shows the quantity arriving that the participant had ordered in the experiment period. Thus, the 'Legacy PO Arrivals' reduce in frequency as the 'PO Arrival' quantities increase as the experiment moves through the 12-month period.

The 'Stock minus Sales' column showed the current stock quantity this was calculated by taking away 4 weeks' demand from the previous monthly total and adding any purchase order arrivals (either legacy or not).

Column 'Purchase Orders' shows the total number of purchase orders outstanding for the SKU (this total will include legacy purchase orders at the start of the experiment).

‘Port of Loading’ indicates where the supplier is Japan (Nagoya). This was used to highlight a possible longer lead-time due to the shipment volume being smaller resulting in longer time to fill 20ft or 40ft containers.

The next three columns showing weekly demand for SKU’s in weekly buckets. The spreadsheet had 24 weekly buckets so from the table 3.4 the column titled 2013. Week 30 is the oldest weekly demand bucket and the next two columns are the most recent demand buckets showing week 52 of 2013 and week 1 of 2014. The participants could see all 24 weeks in the spreadsheet each month.

Forecast information:

Frequency	SI	Value	1st half	2nd half	1st half average	2nd half average	Diff %	Ratio	Difference of averages

Figure 3.5. Experiment spreadsheet headings (continued)

‘Frequency’ was indicated for each SKU. This was the sum of the number of weeks in the 24-week period in which the SKU had a demand. The ‘SI’ (stock index) was calculated by dividing the ‘Stock - Sales Orders’ figure by the ‘1 MTH STK’ figure. The ‘Value’ column showed the 1pc cost for each SKU. The ‘1st Half’ and ‘2nd Half’ columns were the sum of the first and second 12-week periods. The first and second half averages were calculated dividing the sums by 12. The next three columns compare the two 12-week periods. Firstly, by a % (‘Diff %’), then by a ratio (‘Ratio’) and finally by a number difference (‘Difference of Averages’). The comparison was always the 2nd 12-week average to the first 12 week.

Forecast information:

Month 1	Month 2	Month 3	Month 1 Forecast	Month 2 Forecast	Month 3 forecast	OUT (Wks)	**Calculated Order Qty**	**Final Order Qty**

Figure 3.6. Experiment spreadsheet headings (continued)

The 'Month 1' column showed the next month's statistical forecast. This was calculated by adding the 'Difference of the Averages' divided by 12 to the 'Second half average' of the 24 weeks' demand. The company were including this to try and reflect any trend in the demand over 24 weeks. The 'Month 2' column then showed this figure minus the 'Difference of the averages' and the 'Month 3 'column showed the 'Month 2 figure' again minus the 'Difference of the Averages'. This was the 3-month statistical forecast given to the participants for all 90 SKU's. The Month 1,2,3 forecast columns were the forecast that the participants inputted for the 3 monthly periods.

The total 'OUT' in weeks was then shown. The company placed this information close to the forecast columns too highlight where SKU's had a long lead-time.

The 'Calculated order quantity' was automatically populated when the Month 1, 2 and 3 forecasts were entered. The cell calculates the figure by summing the three forecasted months and dividing by 12, this is then multiplied by the OUT and finally the Stock – Sales and the Purchase order quantities are subtracted.

The participants then had an opportunity to adjust the calculated order quantity before saving the forecast. After this adjustment, there was one more opportunity to adjust the final forecast before placing the purchase order. This, in practice is a common occurrence and reflects the desire to round the order quantity to a figure divisible by 10 or 100 for example. The nature of both these adjustments will be reflected upon.

Accuracy measurement and analysis of forecasts needs to reflect the intermittent nature of many of the time series and provide a scale independent result as the time series data demand quantities vary from one group to another. A common method used to measure accuracy for non-intermittent demand is Mean Absolute Percentage Error (MAPE). Wherever there are zero observations any attempt to divide the absolute error (the difference between the forecast and the actual demand notated as only a positive figure) by the actual demand is not useful (any error figure is not divisible by zero).

By using the Mean Absolute Error (MAE) the issue of zero observations is negated. As long at the series has at least 1 observation of more than zero then the mean will always be useable.

When looking at numerous series (this experiment has 9) it is important the accuracy measure is scale dependent, that is, its own scale is not dependent on the scale of the data series. Hoover (2006), recommended using the MAE as a ratio of the mean demand as a solution to the scale dependency problem.

The formula for the MAE: Mean ratio for an individual series is as below:

$$MAE : Mean = \frac{\dfrac{\sum_{i=1}^{N} \left| \hat{d}_{i,i-1} - d_i \right|}{N}}{\dfrac{\sum_{i=1}^{N} d_i}{N}} = \frac{\sum_{i=1}^{N} \left| \hat{d}_{i,i-1} - d_i \right|}{\sum_{i=1}^{N} d_i}$$

Figure 3.7. Mean Absolute Demand: Mean equation

There was a cautionary note that care should be taken with seasonal intermittent data which could make the data unstable. The time series in the experiment are not affected by seasonality and are for 12 months only.

The Senior academic who was remotely forecasting used a statistical method different to the company.

The equation was as below (SBA):

$$y"_t = \left(1 - \frac{x}{2}\right)\frac{z"_t}{p"_t}$$

This equation was an adaption of the method first devised by Croston (1972) as explained in section 2.16 earlier.

3.6. Ethics, Reliability, Validity, Generalizability and Limitations

3.6.1. Ethics

In the context of research ethics refers to the appropriateness of behavior in relation to the rights of those who are affected by the work. Wells (1994) defined ethics as “*a code of behavior appropriate to academics and the conduct of research*”.

Saunders (2003) highlights privacy as the cornerstone of the issues for the researcher. Some of the implications of respecting privacy in business and management research are listed as rights: not to participate; not to be harassed; of participants to determine when they will participate; not to be subject to issues that create stress or discomfort; to expect anonymity and confidentiality.

The relative importance of this issue was stressed by Marshall and Rossman (1999), who argued that evidence that ethical issues should be considered and evaluated and that they should be a criterion against which research proposals are judged.

For this thesis, the participants provided informed consent given freely and based on full information about participation rights and use of data. The company members were part of the team managed by the researcher and were fully aware of the requirements and effort required to participate. Each were responsible for 100's of SKU's as part of the daily departmental operation and the process used for the experiment was completely mirroring the day to day task so each member was already knowledgeable about the work involved.

3.6.2. Reliability

The reliability of a measure indicates whether it is without bias or error and allows a consistency of measurement across items and time. The "goodness" of a measure reflects the consistency with which the instrument measures the concept (Sekaran and Bougie, 2009).

The instrument used to collect the data was the current spreadsheet that was in use at the company for the forecasting procedure. The reliability of the instrument is demonstrated by the fact that this is how the company deals with the day to day procedure of forecasting the spare parts demand. Foddy (1994) discusses the importance of questions and answers making sense emphasizing that "*the question must be understood by the respondent in the way the intended by the researcher and the answer given by the respondent must be understood by* *the researcher in the way intended by the respondent".* As the tasks required in the experiment were daily requirements for the company participants in their job roles this was the case.

There was no need to discuss and explain the spreadsheets and the respective calculations and columns as this was the spreadsheet that the company participants used monthly. This was not the case with the academic outsider who needed the process to be explained.

The process that produces the statistical forecast needs to be understood by the participants otherwise the black box issue arises. If this is the case, then any calculation from other companies would be comparable before a participant were to make judgmentally adjustments. There is a question regarding the workings of the calculation (what is entailed) and whether the participant accepts the validity of it but generally a relatively simple average would be understood and accepted as valid.

3.6.3. Validity and Generalizability

The procedure and experience of the participants was the same as their day to day roles thus the conclusions drawn about causes and effect are valid internally.

Humbley and Zumbo (1996), when discussing construct validity focused on whether the data serves a useful purpose and whether it is used in practice. This is the case as the experiment did not change the procedure that existed already for forecasting spare parts.

The constraints regarding population size (there are 6 participants) were the number of people in the company who were responsible for spare parts forecasting. External validity is therefore comparable to other companies who have a similar sized forecasting department.

The number of participants is comparable to many other companies both within the sector of this company but also across other business sectors.

3.6.4. Limitations

As the experiment was ran in a real-world environment by participants who were employed by company the participant size was naturally constrained. However, the fact that they were practitioners, who were in the real-world environment, using a method that they were familiar with enhances the generalizability of the experiment when compared to other laboratory tests. This pool of forecasting resource is typical of similar sized companies.

The number of SKU's and the length of the experiment were governed somewhat by the practicalities of participants doing their employee duties and finding time to complete the monthly task required. By selecting SKU's by their characteristics, the number of time series types was maximized (9 groups of 10 SKU's). Which would give a maximum of 3,240 forecasts per participant (9 x 10 x 3 x 12) and 19,440 forecasts in total (3,240 x 6). The total could be lower than this figure if any SKU becomes withdrawn (no longer available to purchase). The maximum number of possible final adjustments (adjustments made to the final forecast) per participant is 1,080 (90 x 12) and 6,480 in total (1,080 x 6).

The number of forecasts is large considering that the participants can apply judgment to all the forecasts and the time that this decision-making process takes. The participants reflected the total actual number of spare parts forecasters within the business.

The SKU numbers are reflective of a substantial portfolio of products and the subsequent spare parts therein. This level of forecasts per month is comparable to a medium sized business machine manufacturer selling into the European market.

The opportunity to have participants that were willing and able to participate for a full year is unusual (I am not aware of any other research that had this length of involvement). That the demand and SKU's were 'live' in a real-time environment also makes the experiment unique.

4. Data findings and analysis

4.1. Introduction

The aim of the experiment was to assess the impact of judgement on statistical forecasts and to evaluate the impact on inventory management.

The experiment purpose was to analyse six participant's judgmental adjustments to a statistical forecast over a 12-month period. The participants also had the opportunity to make a final judgmental adjustment to the forecast also. The impact these judgments have on inventory control was also investigated. At the start point of the experiment the participants have a given stock level, 12-month demand series and outstanding purchase order total, which was identical for all. For the next 12 months, the implications of their forecasts and subsequent purchase orders are simulated. Unless participants did not adjust the statistical forecasts at any point in the 12-month period for a SKU (in which case the SKU data would be identical for the participants) any judgmental adjustment to the statistical forecast or the final order quantity would result in different purchase orders in the future and different inventory levels.

The spreadsheet for each participant was collected at the end of the 12-month period. Each month the new 4 weeks of demand were added to the spreadsheet and the old 4 weeks of demand were deleted off. Each month the spreadsheet showed the latest 24 weeks' demand.

Five of the participants worked as inventory analysists in the company, responsible for the forecasting and order placement of the spare parts involved in the experiment and one participant was external. The external participant did not have context of the spare parts involved and therefore was not able to judgmentally adjust from a context perspective. Instead the participant could use a statistical forecast technique that they felt was appropriate to the time series. This external statistical forecast was used to reflect on the company's own method by analyzing the results regarding forecast accuracy and inventory.

In this chapter the results will be shown as follows:

Forecast judgmental adjustment accuracy

- By expertise
- By SKU group

Final forecast adjustment accuracy

- By expertise
- By SKU group

Effect of horizon on forecast judgmental adjustment

Direction of adjustment

- Judgmental adjustment
- Final forecast adjustment

Direction / size of adjustment

- Judgmental forecast
- Final forecast

Impact on inventory

- Judgmental forecast

- Final forecast

4.2. What effect did expertise have on the accuracy of forecasts?

The participants were as follows:

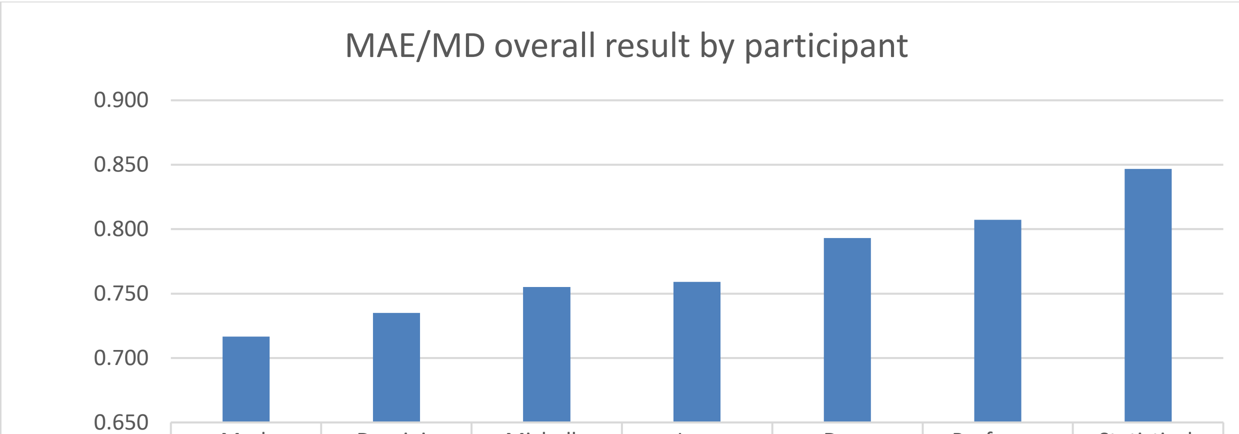
- Mark Departmental Manager (20 years' experience)
- Dominic Senior Inventory Analyst (10 years' experience)
- Michelle Senior Inventory Analyst (5 years' experience)
- Dave Inventory Analyst (5 years' experience)
- Jane Inventory Analyst (1 years' experience)
- Academic Senior academic (not in the company)

The experiment contained 90 SKU's for which there were 3 forecasts (Month 1, Month 2 and Month 3) each month for 12 months. A total of 3240 forecasts were entered from each participant.

The range of expertise in the participant was large. Ranging from 20 years of experience to just 1 year. There was also the Senior academic who was working purely with the data and had no other information regarding the SKU's involved. One of the aims of the experiment was to investigate if this would affect the accuracy of the judgmental adjustment and the final adjustment (for the company participants only).

The accuracy measure that was applied was the ratio Mean Average Error / Mean Demand. The ratio was calculated for each forecast made and then each forecast was compared to the actual demand. The statistical forecast which included no adjustments was also included to provide insight into whether the adjustments participants had made were improving the companies own statistical forecast.

To compare forecast accuracy across all the forecasts the average of the MAE:MD was taken to report the month 1,2 and 3 figures. This is possible due to the scale independent nature of the error measure.



	Mark	Dominic	Michelle	Jane	Dave	Professor	Statistical
MAE/MD	0.717	0.735	0.755	0.759	0.793	0.807	0.847

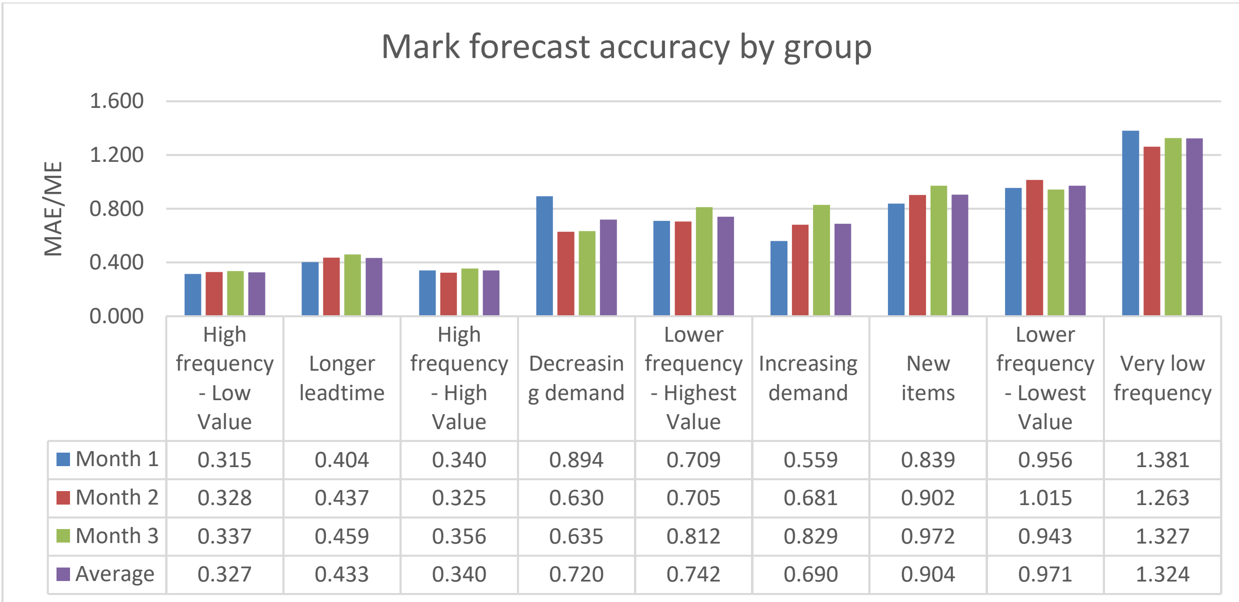
Table 4.1. Forecast accuracy by participant across all groups.

In table 4.1. the lowest ratio relates to the least error. The participant to the left of the table created the most accurate forecasts on average across the whole experiment (the lowest MAE:MD).

All the participants performed better than the statistical forecast. All the company participants performed better than the academic who had no causal information but applied a different forecasting method to the companies.

The ranking show that the company participants were ranked according to experience (the lowest forecast error was the most experienced participant) apart from the 4th and 5th placed participants who were reversely placed.

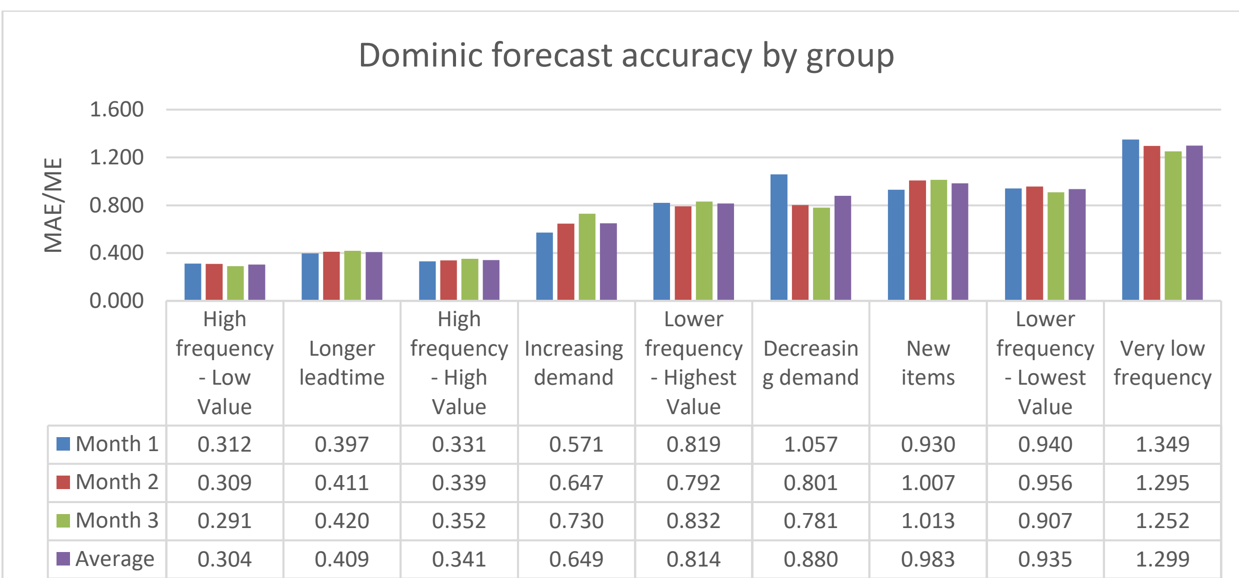
The forecast accuracy for each participant by SKU group was:



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Decreasing demand	Lower frequency - Highest Value	Increasing demand	New items	Lower frequency - Lowest Value	Very low frequency
Month 1	0.315	0.404	0.340	0.894	0.709	0.559	0.839	0.956	1.381
Month 2	0.328	0.437	0.325	0.630	0.705	0.681	0.902	1.015	1.263
Month 3	0.337	0.459	0.356	0.635	0.812	0.829	0.972	0.943	1.327
Average	0.327	0.433	0.340	0.720	0.742	0.690	0.904	0.971	1.324

Table 4.2. Mark forecast accuracy by SKU group.

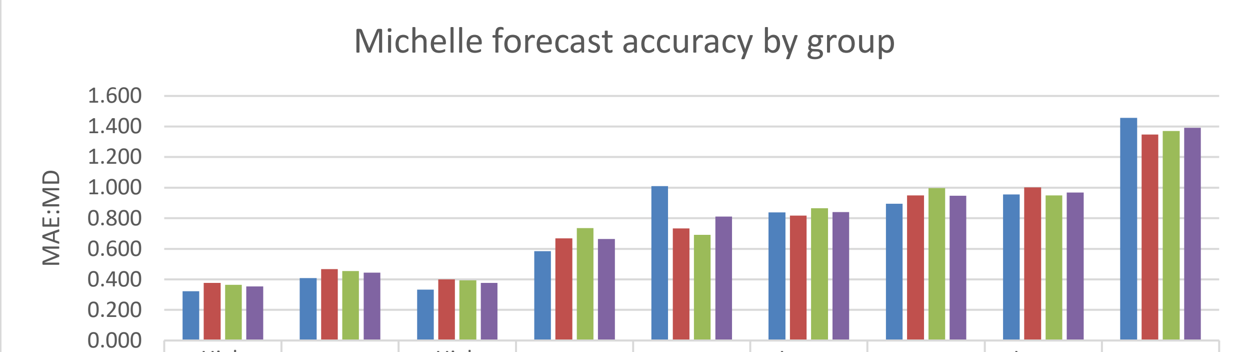
For Mark, the most senior expert the higher frequency and longer lead-time groups were the best forecasts and the lower frequency groups the worse. The overall average was the best of the participants.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Increasing demand	Lower frequency - Highest Value	Decreasing demand	New items	Lower frequency - Lowest Value	Very low frequency
Month 1	0.312	0.397	0.331	0.571	0.819	1.057	0.930	0.940	1.349
Month 2	0.309	0.411	0.339	0.647	0.792	0.801	1.007	0.956	1.295
Month 3	0.291	0.420	0.352	0.730	0.832	0.781	1.013	0.907	1.252
Average	0.304	0.409	0.341	0.649	0.814	0.880	0.983	0.935	1.299

Table 4.3. Dominic forecast accuracy by SKU group.

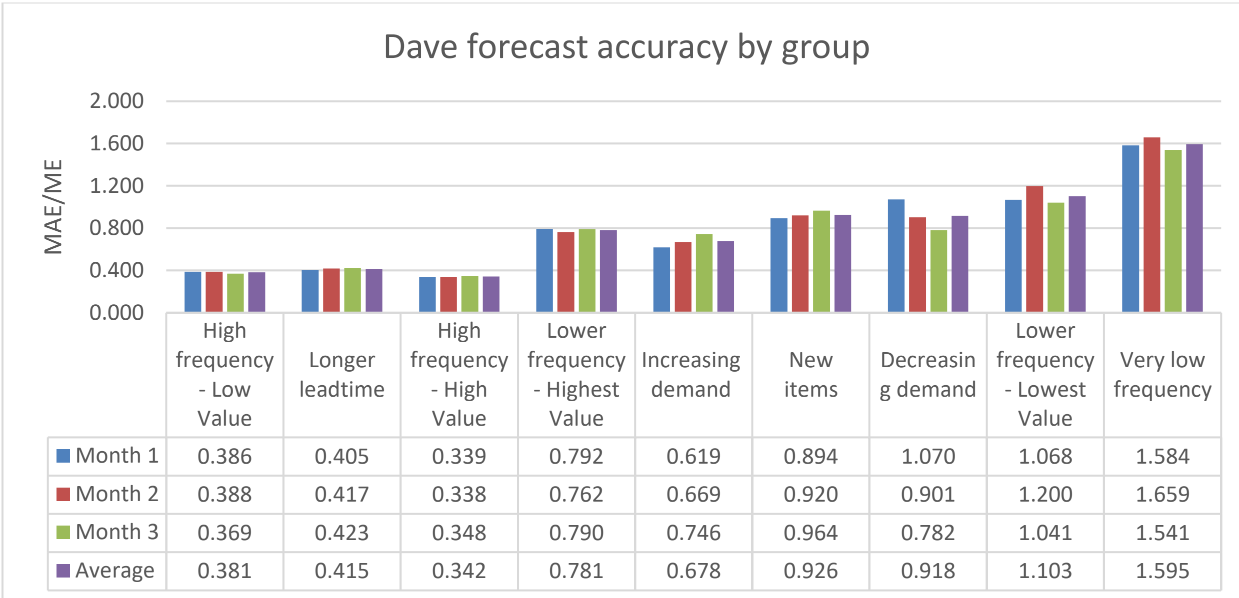
Dominic the second most expert of the participants the result was similar. High frequency best forecast and low frequency worse forecast. Dominic's overall average was the second best of the participants.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Increasing demand	Decreasin g demand	Lower frequency - Highest Value	New items	Lower frequency - Lowest Value	Very low frequency
Month 1	0.323	0.408	0.334	0.585	1.009	0.839	0.894	0.956	1.456
Month 2	0.377	0.468	0.400	0.668	0.734	0.816	0.950	1.002	1.349
Month 3	0.364	0.455	0.395	0.736	0.692	0.865	0.997	0.949	1.370
Average	0.355	0.444	0.376	0.663	0.812	0.840	0.947	0.969	1.391

Table 4.4. Michelle forecast accuracy by SKU group.

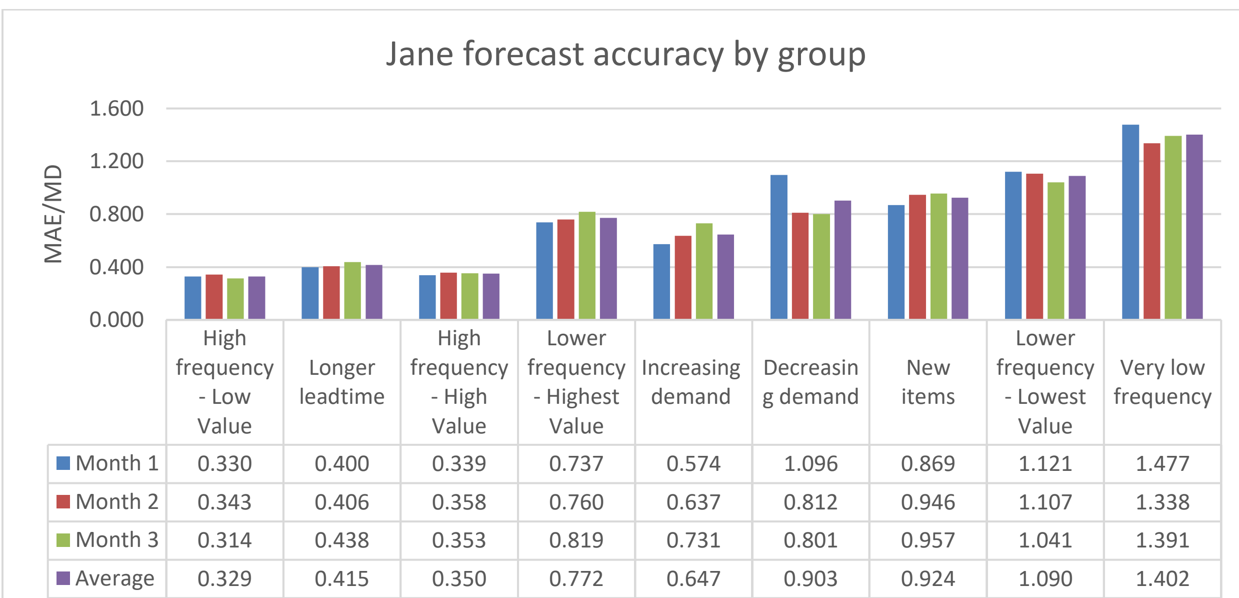
Michelle the third most expert of the participants the result was similar. High frequency best forecast and low frequency worse forecast. Michelle's overall average was the third best of the participants.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Lower frequency - Highest Value	Increasing demand	New items	Decreasing demand	Lower frequency - Lowest Value	Very low frequency
Month 1	0.386	0.405	0.339	0.792	0.619	0.894	1.070	1.068	1.584
Month 2	0.388	0.417	0.338	0.762	0.669	0.920	0.901	1.200	1.659
Month 3	0.369	0.423	0.348	0.790	0.746	0.964	0.782	1.041	1.541
Average	0.381	0.415	0.342	0.781	0.678	0.926	0.918	1.103	1.595

Table 4.5. Dave forecast accuracy by SKU group.

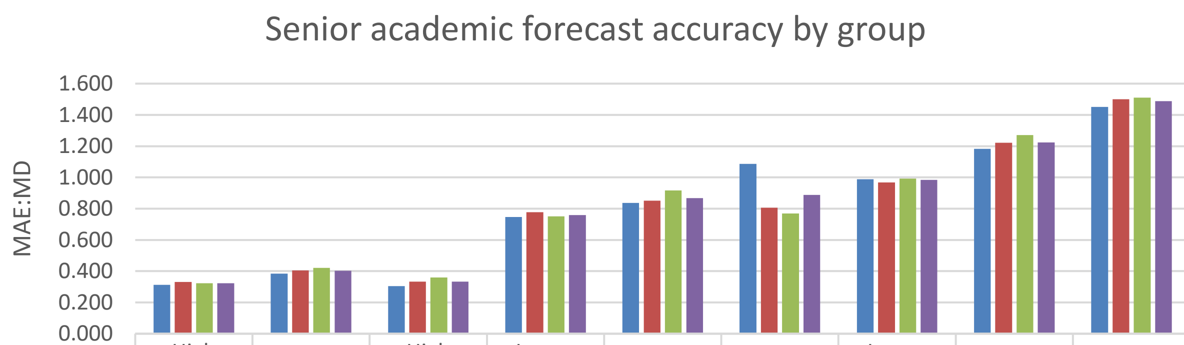
Dave the fourth most expert of the participants also had a similar ranking of errors. His average made him the fouth best of the participants.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Lower frequency - Highest Value	Increasing demand	Decreasing demand	New items	Lower frequency - Lowest Value	Very low frequency
Month 1	0.330	0.400	0.339	0.737	0.574	1.096	0.869	1.121	1.477
Month 2	0.343	0.406	0.358	0.760	0.637	0.812	0.946	1.107	1.338
Month 3	0.314	0.438	0.353	0.819	0.731	0.801	0.957	1.041	1.391
Average	0.329	0.415	0.350	0.772	0.647	0.903	0.924	1.090	1.402

Table 4.6. Jane forecast accuracy by SKU group.

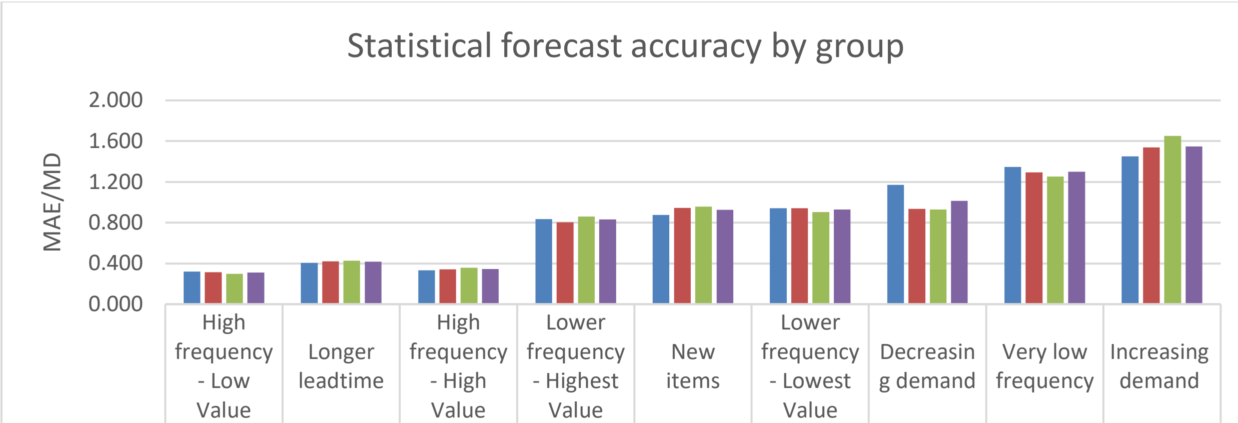
Jane the fifth most expert of the participants also had a similar ranking of errors. The overall average for Jane made her the fourth best participant.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Lower frequency - Highest Value	New items	Decreasing demand	Lower frequency - Lowest Value	Increasing demand	Very low frequency
Month 1	0.314	0.383	0.304	0.746	0.836	1.087	0.989	1.183	1.451
Month 2	0.331	0.406	0.334	0.777	0.852	0.805	0.969	1.221	1.499
Month 3	0.322	0.421	0.360	0.751	0.917	0.769	0.993	1.270	1.511
Average	0.322	0.403	0.333	0.758	0.868	0.887	0.984	1.225	1.487

Table 4.7. Senior academic forecast accuracy by SKU group.

The academic participant varied from the rest of the participants in that the increasing demand group was the second worse error forecast. The overall average for the Senior academic meant his was the least accurate of the participants.



	High frequency - Low Value	Longer leadtime	High frequency - High Value	Lower frequency - Highest Value	New items	Lower frequency - Lowest Value	Decreasing demand	Very low frequency	Increasing demand
Month 1	0.322	0.405	0.334	0.834	0.876	0.943	1.172	1.349	1.450
Month 2	0.315	0.421	0.342	0.805	0.946	0.943	0.937	1.295	1.537
Month 3	0.297	0.427	0.357	0.861	0.957	0.903	0.931	1.252	1.653
Average	0.311	0.418	0.344	0.833	0.926	0.929	1.013	1.299	1.547

Table 4.8. Statistic forecast accuracy by SKU group.

The statistical forecast provided by the company had a similar grouping in the best forecast groups but had the increasing and decreasing groups as the worse forecast of the groups and not the lower frequency groups. The statistical forecast average result was the worse than all the participants.

By ranking each participant forecast error results in the tables by SKU groups the order for each is similar (tables 4.1. – 4.8.)

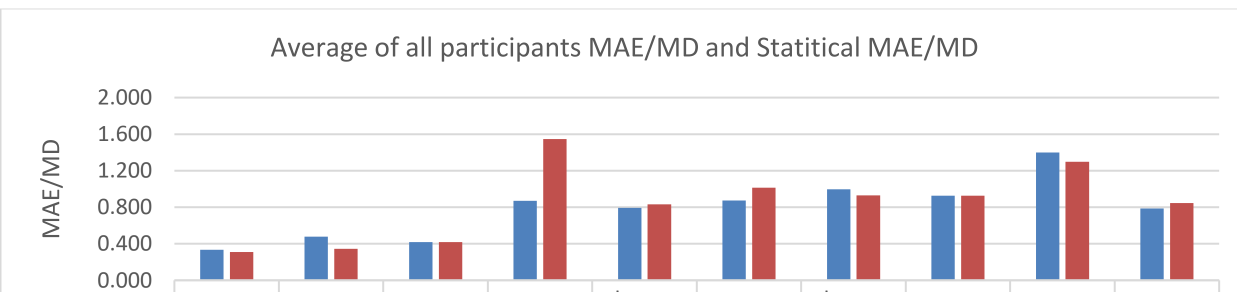
By ranking in order of smallest error using 1-9 for each participant the table 4.9. shows there was only a small variance in the ranking of SKU group by error size.

SKU Group	Mark	Dom	Dave	Jane	Michelle	Professor	Statistic	Sum
High frequency - Low Value	1	1	1	1	1	1	1	7
Longer leadtime	2	2	2	2	2	2	2	14
High frequency - High Value	3	3	3	3	3	3	3	21
Lower frequency - Highest Value	5	5	4	4	6	4	4	32
Increasing demand	6	4	5	5	4	8	9	41
Decreasing demand	4	6	7	6	5	6	7	41
New items	7	7	6	7	7	5	5	44
Lower frequency - Lowest Value	8	8	8	8	8	7	6	53
Very low frequency	9	9	9	9	9	9	8	62

Table 4.9. SKU group error ranking.

The first three SKU groups in table 4.9. are the same for all participants. The remaining six groups are not the same however the final two groups are the same for five out of the six human participants.

The result reflects the difference in the time series. Lower frequency is often hard to forecast due to the intermittency of the demand quantities and these groups were at the bottom of the table. Higher frequency can be easier to forecast as there is a demand each month which often means averages can be useful for forecasting and these groups are at the top of the table. The middle SKU groups were groups with trends or unusual traits (such as new items) which average based calculations are not so good a tool to forecast with.



	High frequency - Low Value	High frequency - High Value	Longer leadtime	Increasing demand	Lower frequency - Highest Value	Decreasing demand	Lower frequency - Lowest Value	New items	Very low frequency	Average
Average	0.333	0.478	0.420	0.870	0.793	0.875	0.996	0.928	1.398	0.788
Statistical	0.311	0.344	0.418	1.547	0.833	1.013	0.929	0.926	1.299	0.847

Table 4.10. Comparison of average MAE/MD for participant and statistic.

Whilst from table 4.1. the statistical forecast performs the worst over the whole series as an average, taken individually table 4.10. shows that on average per group it performed better than the participants on six occasions. The statistical forecast used by the company performed better for the high frequency groups and worse for the very low and lower frequency ones which was expected as described earlier.

4.3. How did accuracy vary between SKU groups?

The SKU groups were chosen to highlight different characteristics both in the time series an also the inventory importance from a value perspective. The aim was to see if the participants approached the groups differently due to either off the variables.

Data tables were created showing the ratio for each of the SKU groups as shown in Table 4.11.

The table reports the participants MEA/MD across each group by forecast horizon. An average was then taken to show an aggregated accuracy over the three months.

High frequency - Low Value	**Month 1**	**Month 2**	**Month 3**	**Average**
Dominic	0.312	0.309	0.291	**0.304**
Statistical	0.322	0.315	0.297	**0.311**
Professor	0.314	0.331	0.322	**0.322**
Mark	0.315	0.328	0.337	**0.327**
Jane	0.330	0.343	0.314	**0.329**
Michelle	0.325	0.377	0.361	**0.354**
Dave	0.386	0.388	0.369	**0.381**
				0.333

Lower frequency - Highest Value	**Month 1**	**Month 2**	**Month 3**	**Average**
Mark	0.709	0.705	0.829	**0.748**
Professor	0.746	0.777	0.751	**0.758**
Dave	0.792	0.762	0.790	**0.781**
Jane	0.737	0.760	0.819	**0.772**
Dominic	0.819	0.792	0.832	**0.814**
Statistical	0.834	0.805	0.861	**0.833**
Michelle	0.846	0.824	0.872	**0.847**
				0.793

Very low frequency	**Month 1**	**Month 2**	**Month 3**	**Average**
Mark	1.381	1.263	1.327	**1.324**
Statistical	1.349	1.295	1.252	**1.299**
Dominic	1.349	1.295	1.252	**1.299**
Michelle	1.456	1.349	1.349	**1.384**
Jane	1.477	1.338	1.391	**1.402**
Professor	1.451	1.499	1.511	**1.487**
Dave	1.584	1.659	1.541	**1.595**
				1.398

Longer leadtime	**Month 1**	**Month 2**	**Month 3**	**Average**
Professor	0.383	0.406	0.421	**0.403**
Dominic	0.397	0.411	0.420	**0.409**
Jane	0.400	0.406	0.438	**0.415**
Dave	0.405	0.417	0.423	**0.415**
Statistical	0.405	0.421	0.427	**0.418**
Mark	0.404	0.437	0.459	**0.433**
Michelle	0.412	0.464	0.453	**0.443**
				0.420

Decreasing demand	**Month 1**	**Month 2**	**Month 3**	**Average**
Mark	0.894	0.630	0.635	**0.720**
Michelle	1.004	0.723	0.681	**0.803**
Dominic	1.057	0.801	0.781	**0.880**
Jane	1.096	0.812	0.801	**0.903**
Dave	1.070	0.901	0.782	**0.918**
Professor	1.087	0.805	0.769	**0.887**
Statistical	1.172	0.937	0.931	**1.013**
				0.875

High frequency - High Value	**Month 1**	**Month 2**	**Month 3**	**Average**
Professor	0.304	0.334	0.360	**0.333**
Mark	0.340	0.325	0.356	**0.340**
Dominic	0.331	0.339	0.352	**0.341**
Dave	0.339	0.338	0.348	**0.342**
Statistical	0.334	0.342	0.357	**0.344**
Michelle	0.344	0.409	0.404	**0.386**
Jane	0.737	0.760	0.819	**0.772**
				0.408

Lower frequency - Lowest Value	**Month 1**	**Month 2**	**Month 3**	**Average**
Statistical	0.943	0.943	0.903	**0.929**
Dominic	0.940	0.956	0.907	**0.935**
Michelle	0.948	0.993	0.941	**0.960**
Mark	0.956	1.015	0.943	**0.971**
Professor	0.989	0.969	0.993	**0.984**
Jane	1.121	1.107	1.041	**1.090**
Dave	1.068	1.200	1.041	**1.103**
				0.996

New items	**Month 1**	**Month 2**	**Month 3**	**Average**
Professor	0.836	0.852	0.917	**0.868**
Mark	0.839	0.902	0.972	**0.904**
Dave	0.894	0.920	0.964	**0.926**
Statistical	0.876	0.946	0.957	**0.926**
Jane	0.869	0.946	0.957	**0.924**
Michelle	0.897	0.961	1.025	**0.961**
Dominic	0.930	1.007	1.013	**0.983**
				0.928

Increasing demand	**Month 1**	**Month 2**	**Month 3**	**Average**
Jane	0.574	0.637	0.731	**0.647**
Dominic	0.571	0.647	0.730	**0.649**
Michelle	0.589	0.654	0.736	**0.660**
Dave	0.619	0.669	0.746	**0.678**
Mark	0.559	0.681	0.812	**0.684**
Professor	1.183	1.221	1.270	**1.225**
Statistical	1.450	1.537	1.653	**1.547**
				0.870

Table 4.11. Participants accuracy by SKU group (ranked by smallest MAE/MD).

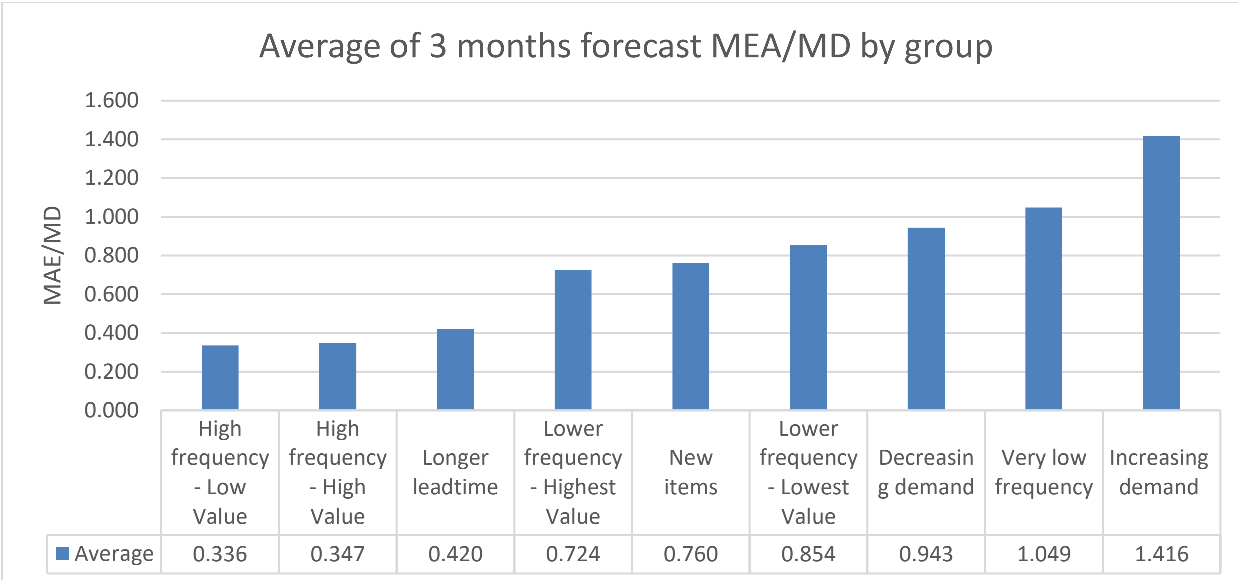


Table 4.12. Accuracy by SKU group ranked.

The results show that when the SKU group characteristic contained high frequency or high value then the errors were smaller. For the groups with trends, lower frequency and no history (new items) the errors were higher.

The coefficient of variation (SD/Mean) for each SKU group shows the relative magnitude of the standard deviation. This gives some insight into the difficulty of forecasting for each SKU group.

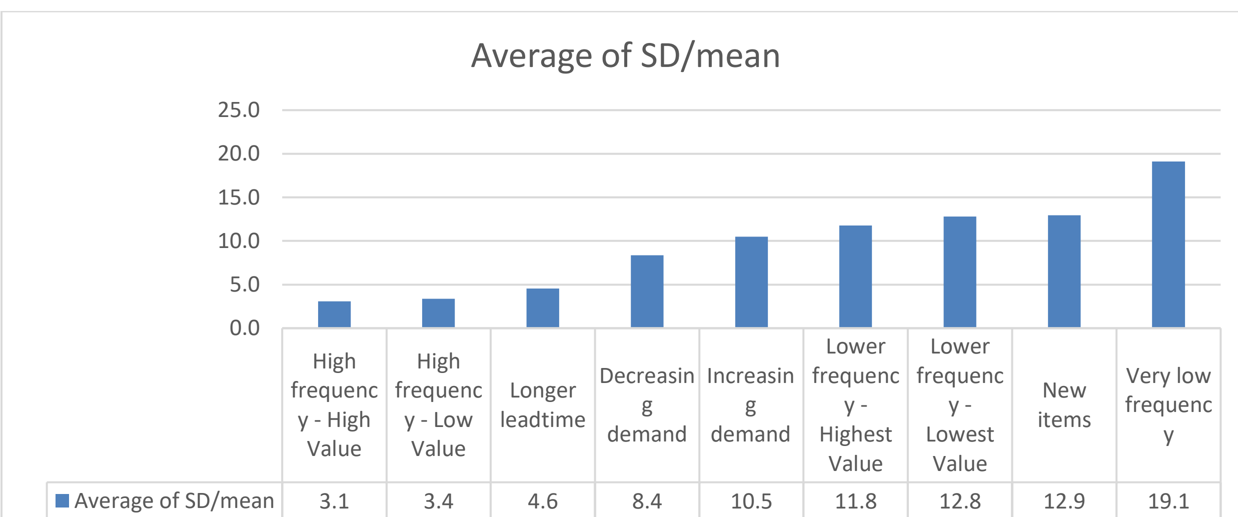
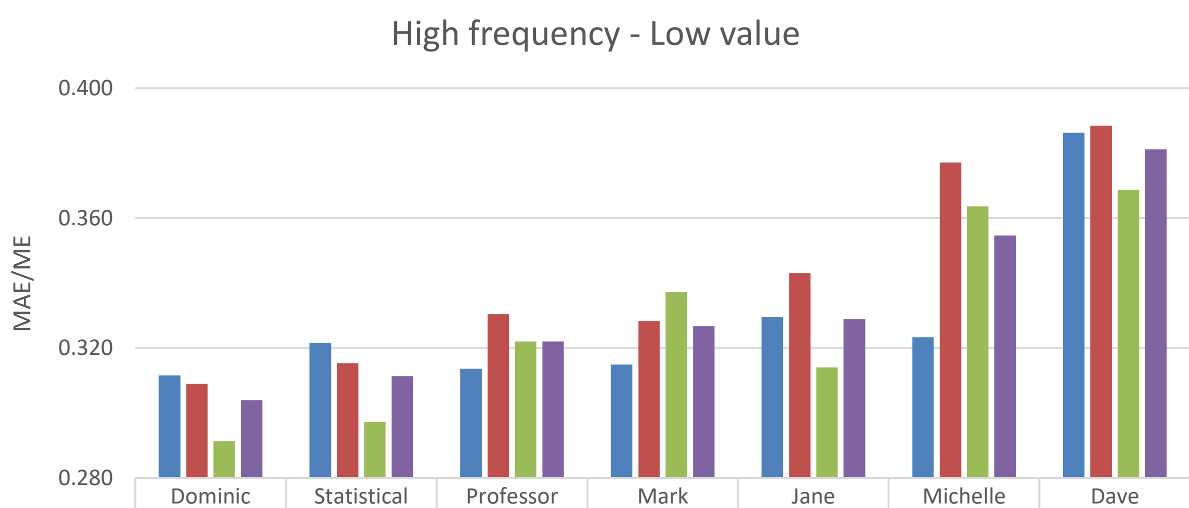


Table 4.13. Coefficient of variation per SKU group.

The forecast accuracy for each SKU group by participant including the statistical forecast was as follows.

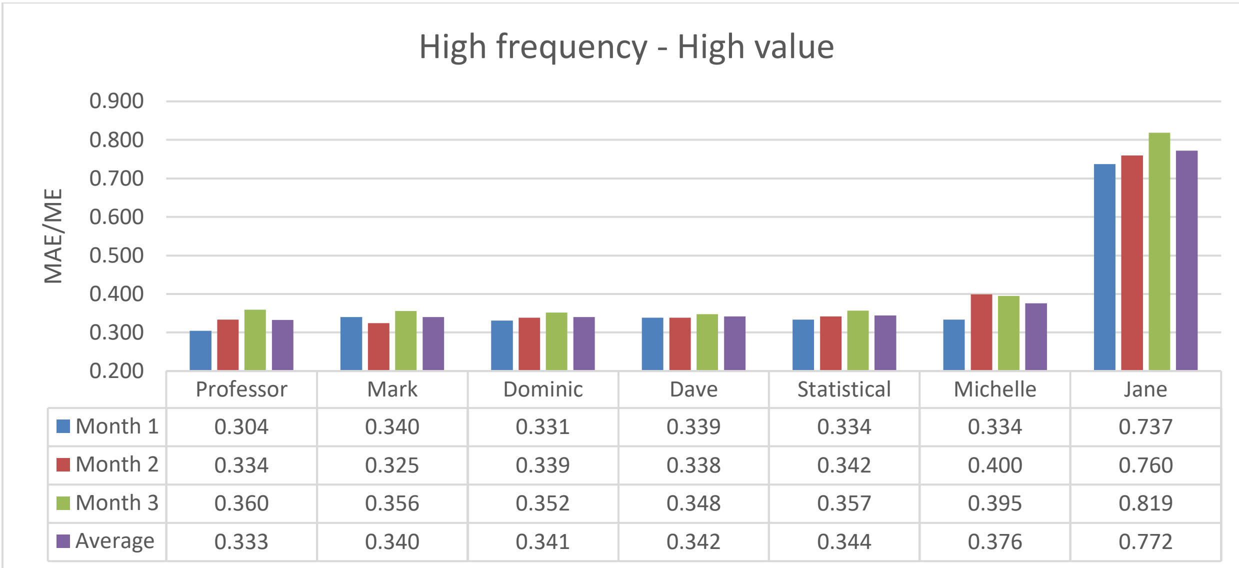


	Dominic	Statistical	Professor	Mark	Jane	Michelle	Dave
Month 1	0.312	0.322	0.314	0.315	0.330	0.323	0.386
Month 2	0.309	0.315	0.331	0.328	0.343	0.377	0.388
Month 3	0.291	0.297	0.322	0.337	0.314	0.364	0.369
Average	0.304	0.311	0.322	0.327	0.329	0.355	0.381

Table 4.14. Participant performance for High Frequency – Low Value.

The two most experienced participants along with the Senior academic and the statistical forecast performed the best. This category was important as it was a high frequency category. High frequency means that from an operational perspective there would be more orders and any stock outs (orders which could not be fulfilled) would have a large impact on back orders.

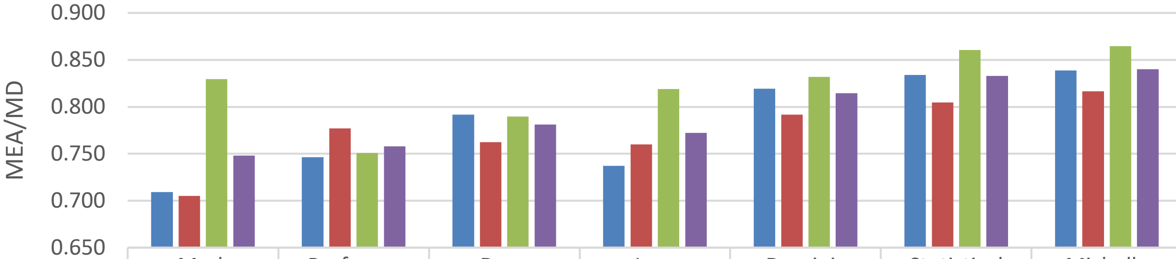
This means that these SKU's would make up a higher percentage of daily requirements than other slow-moving categories. Using a simple ABC Pareto classification, they would be 'A' items.



	Professor	Mark	Dominic	Dave	Statistical	Michelle	Jane
Month 1	0.304	0.340	0.331	0.339	0.334	0.334	0.737
Month 2	0.334	0.325	0.339	0.338	0.342	0.400	0.760
Month 3	0.360	0.356	0.352	0.348	0.357	0.395	0.819
Average	0.333	0.340	0.341	0.342	0.344	0.376	0.772

Table 4.15. Participant performance for High Frequency – High Value.

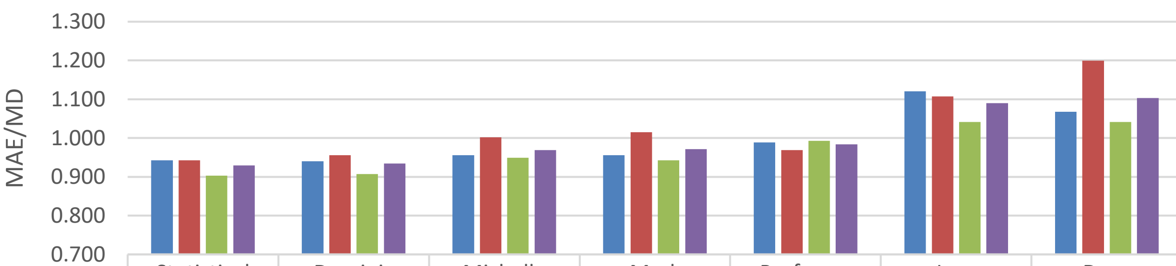
The senior academic and the two most experienced company participants performed the best along with Dave who was the 2nd best in this category. The statistical forecast is relatively close to all the participants forecasts apart from Jane (the least experienced of the company participants). This category was important as it was a high frequency category. These items would also be ‘A’ items as they are fast moving. However, they are also high value, so they are important from an inventory perspective. From an operational perspective this group was the most important as it was both high order volume and high cost. This could have been a factor when Jane was adjusting (as the least experienced of the company participants) as she may have hedged upwards in the knowledge that this group of SKUs’ was the most important from an availability and sales perspective (although high stock could adversely affect the inventory target).



	Mark	Professor	Dave	Jane	Dominic	Statistical	Michelle
Month 1	0.709	0.746	0.792	0.737	0.819	0.834	0.839
Month 2	0.705	0.777	0.762	0.760	0.792	0.805	0.816
Month 3	0.829	0.751	0.790	0.819	0.832	0.861	0.865
Average	0.748	0.758	0.781	0.772	0.814	0.833	0.840

Table 4.16. Participant performance for Low Frequency – High Value.

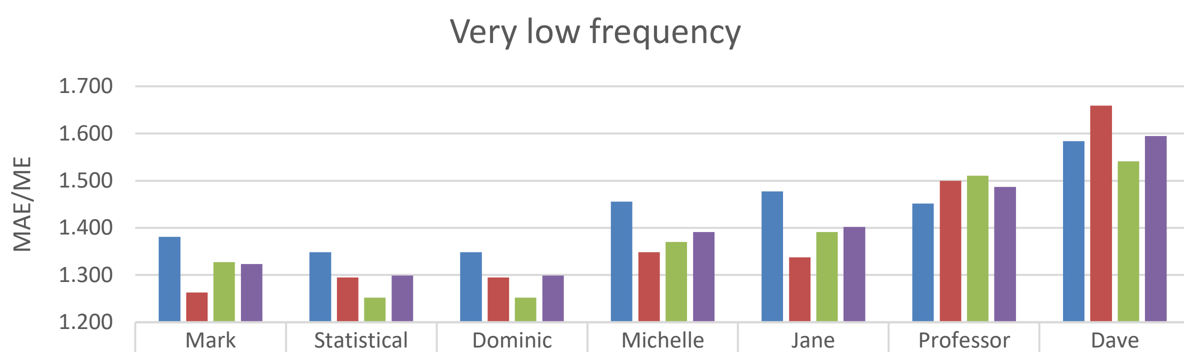
The most experienced company participant and the senior academic performed the best in this category. Although these items were low frequency, they were high value, so the inventory target is important.



	Statistical	Dominic	Michelle	Mark	Professor	Jane	Dave
Month 1	0.943	0.940	0.956	0.956	0.989	1.121	1.068
Month 2	0.943	0.956	1.002	1.015	0.969	1.107	1.200
Month 3	0.903	0.907	0.949	0.943	0.993	1.041	1.041
Average	0.929	0.935	0.969	0.971	0.984	1.090	1.103

Table 4.17. Participant performance for Lower Frequency – Lowest Value.

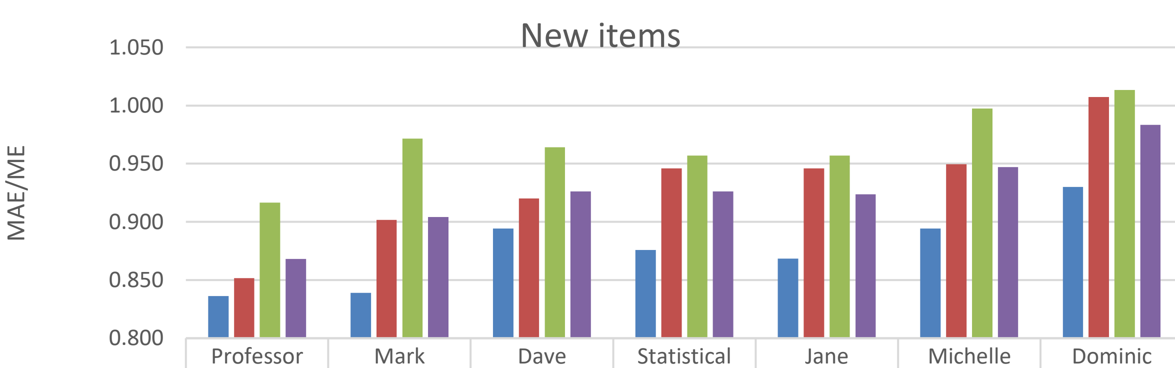
The statistical forecast performed best in this SKU group. This category by being both low in frequency and value was the least most important from an operation perspective. Neither having a large impact on daily orders or inventory value.



	Mark	Statistical	Dominic	Michelle	Jane	Professor	Dave
Month 1	1.381	1.349	1.349	1.456	1.477	1.451	1.584
Month 2	1.263	1.295	1.295	1.349	1.338	1.499	1.659
Month 3	1.327	1.252	1.252	1.370	1.391	1.511	1.541
Average	1.324	1.299	1.299	1.391	1.402	1.487	1.595

Table 4.18. Participant performance for Very Low Frequency.

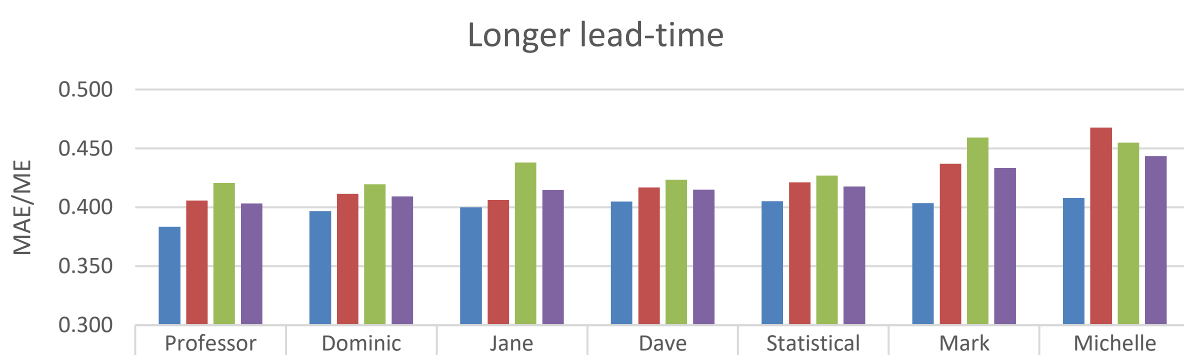
The most experienced company participant and the statistical forecast performed best in this SKU group. The average error lever was the highest in this group due to the difficultly in forecasting very low frequency items.



	Professor	Mark	Dave	Statistical	Jane	Michelle	Dominic
Month 1	0.836	0.839	0.894	0.876	0.869	0.894	0.930
Month 2	0.852	0.902	0.920	0.946	0.946	0.950	1.007
Month 3	0.917	0.972	0.964	0.957	0.957	0.997	1.013
Average	0.868	0.904	0.926	0.926	0.924	0.947	0.983

Table 4.19. Participant performance for New Items.

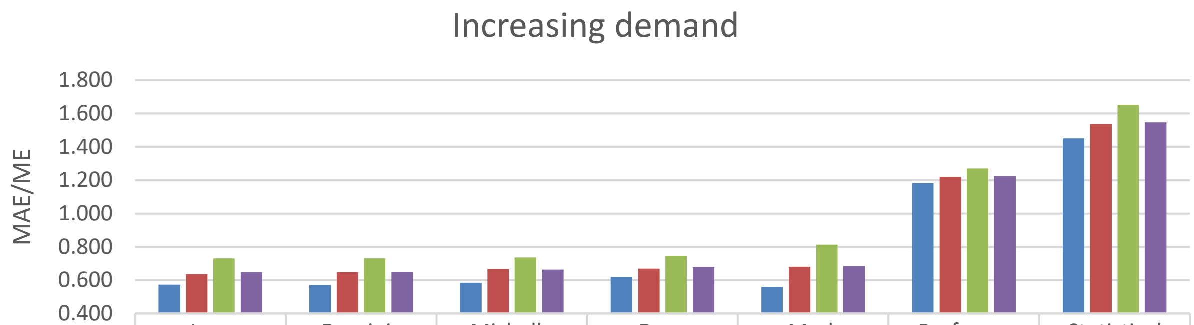
The senior academic and the most experienced company participant performed best in this group. The SKU's were new so there was very little contextual information available.



	Professor	Dominic	Jane	Dave	Statistical	Mark	Michelle
Month 1	0.383	0.397	0.400	0.405	0.405	0.404	0.408
Month 2	0.406	0.411	0.406	0.417	0.421	0.437	0.468
Month 3	0.421	0.420	0.438	0.423	0.427	0.459	0.455
Average	0.403	0.409	0.415	0.415	0.418	0.433	0.444

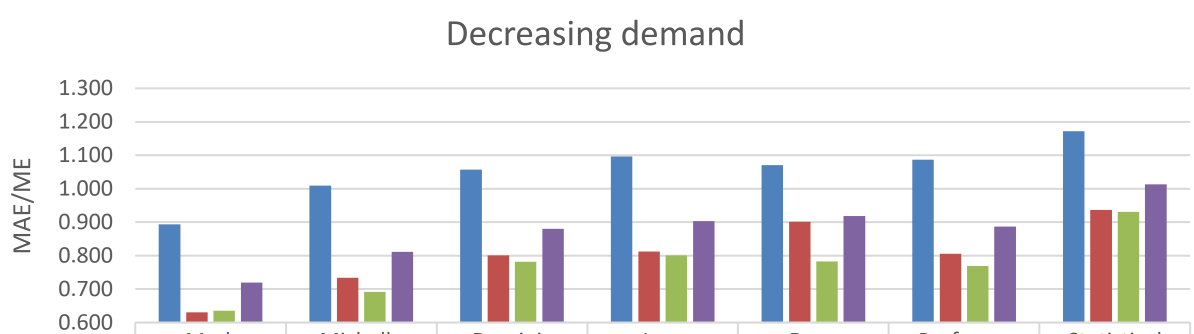
Table 4.20. Participant performance for Longer Lead-time.

The senior academic and the 2nd most experienced company participant performed best. There was little variance across all the participants for this group. This was surprising as the longer lead-time group should have meant that the contextual information was important. The fact that the Senior academic performed best was not predicted.



	Jane	Dominic	Michelle	Dave	Mark	Professor	Statistical
Month 1	0.574	0.571	0.585	0.619	0.559	1.183	1.450
Month 2	0.637	0.647	0.668	0.669	0.681	1.221	1.537
Month 3	0.731	0.730	0.736	0.746	0.812	1.270	1.653
Average	0.647	0.649	0.663	0.678	0.684	1.225	1.547

Table 4.21. Participant performance for Increasing Demand.



	Mark	Michelle	Dominic	Jane	Dave	Professor	Statistical
Month 1	0.894	1.009	1.057	1.096	1.070	1.087	1.172
Month 2	0.630	0.734	0.801	0.812	0.901	0.805	0.937
Month 3	0.635	0.692	0.781	0.801	0.782	0.769	0.931
Average	0.720	0.812	0.880	0.903	0.918	0.887	1.013

Table 4.22. Participant performance for Decreasing Demand.

The company participants performed better than the senior academic and the statistical forecasts for the increasing and decreasing SKU groups. The statistical forecast was the least accurate for both the trending groups underlining the issue of averages being slow to pick up increases and decreases in time series.

In contrast to the table showing participants performance by SKU groups when the participant performance is reported within each group there is more variance.

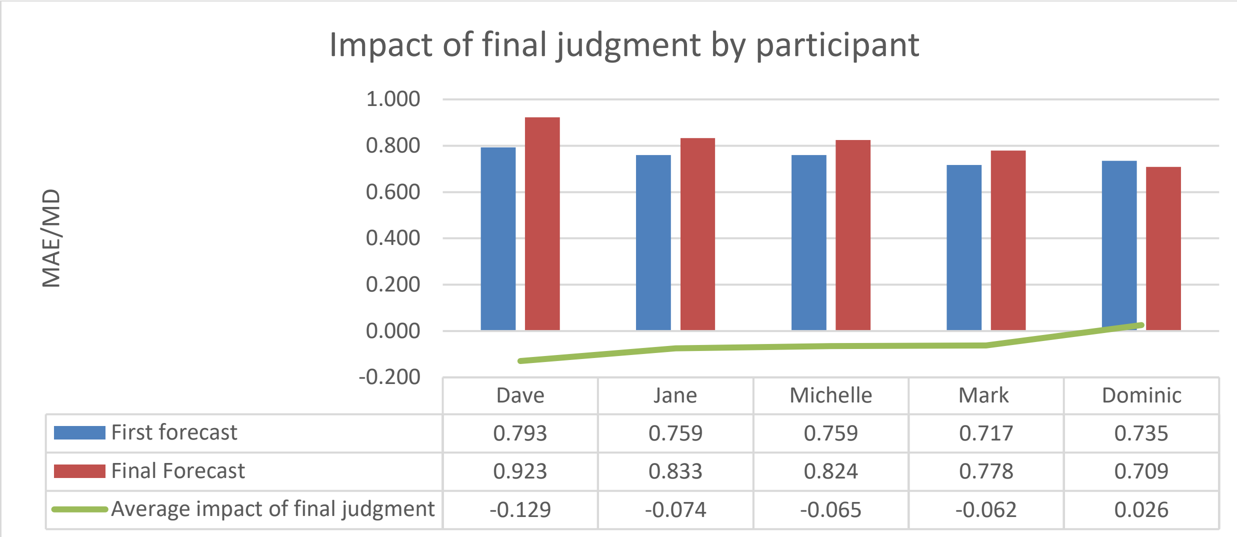
Using the same method as previously but scoring 1-6 for participant by error rank the result is more variable across the participants.

Particpant	High frequency - Low Value	Longer leadtime	High frequency - High Value	Lower frequency - Highest Value	Increasing demand	Decreasing demand	New items	Lower frequency - Lowest Value	Very low frequency	Sum
Mark	4	6	2	1	5	1	2	4	1	26
Dom	1	2	3	5	2	3	7	2	3	28
Professor	3	1	1	2	6	6	1	5	6	31
Statistic	2	5	5	6	7	7	4	1	2	39
Jane	5	3	7	4	1	4	5	6	5	40
Dave	7	4	4	3	4	5	3	7	7	44
Michelle	6	7	6	7	3	2	6	3	4	44

Table 4.23. Participants error ranking.

The table also shows the statistical error ratio producing a result of fourth if scored in this manner. This contrasts with the overall average ratio where the statistical forecast is the largest of the seven.

The company participants also had the opportunity to adjust the adjusted forecast before inputting the final forecast. The impact of these final adjustments was as in table 4.24. Note the senior academic nor the statistical forecast did not make any final adjustments.

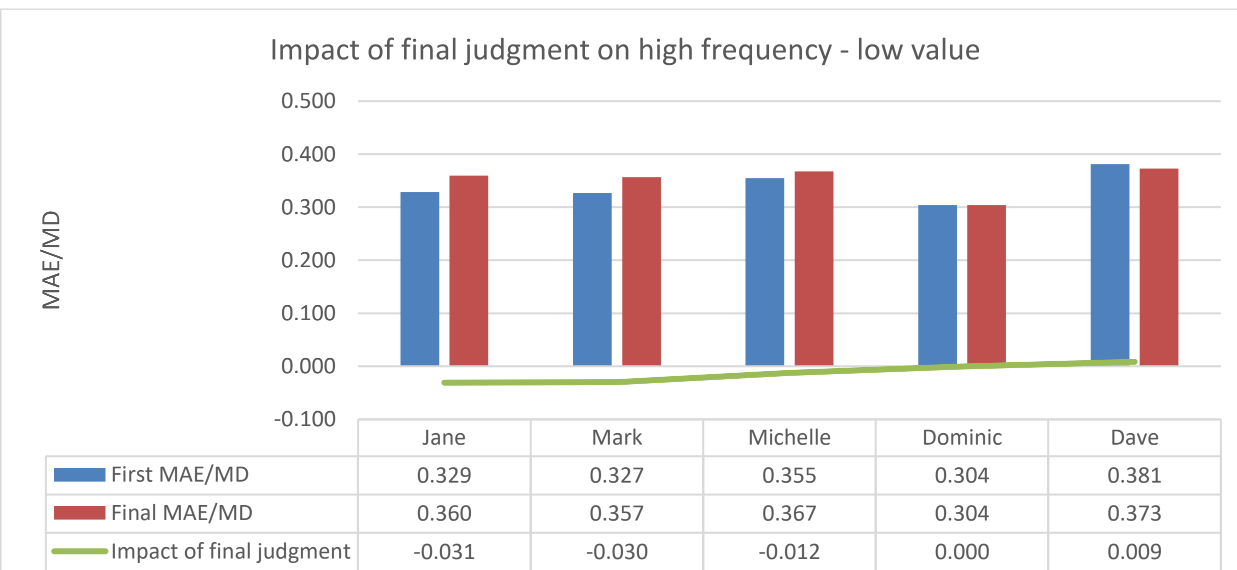


	Dave	Jane	Michelle	Mark	Dominic
First forecast	0.793	0.759	0.759	0.717	0.735
Final Forecast	0.923	0.833	0.824	0.778	0.709
Average impact of final judgment	-0.129	-0.074	-0.065	-0.062	0.026

Table 4.24. The impact of the final adjustment.

The table shows that one of the five company participant made final adjustments that were positive (Dominic) and for the remaining four the impact was negative.

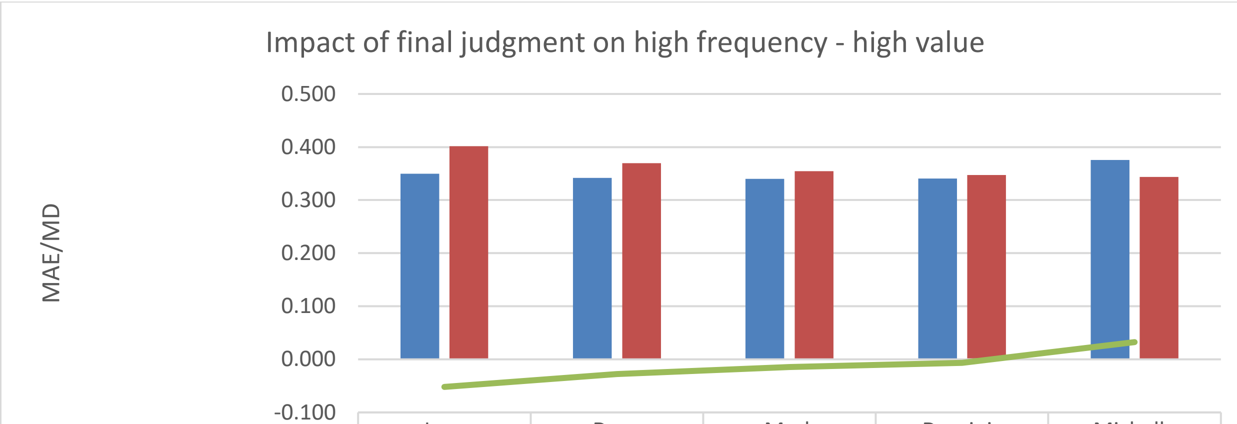
By individual SKU groups as below.



	Jane	Mark	Michelle	Dominic	Dave
First MAE/MD	0.329	0.327	0.355	0.304	0.381
Final MAE/MD	0.360	0.357	0.367	0.304	0.373
Impact of final judgment	-0.031	-0.030	-0.012	0.000	0.009

Table 4.25. Impact of final judgment on high frequency – low value.

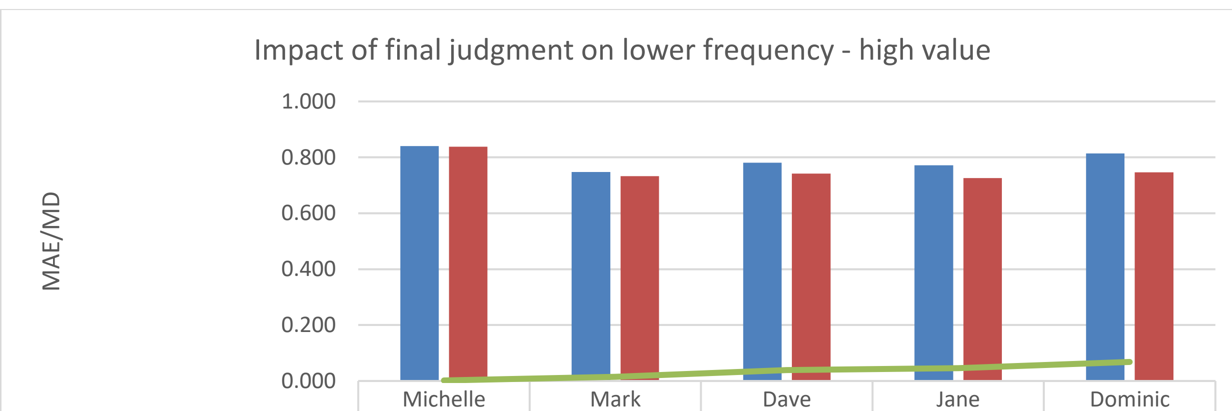
Two of the participants made positive final adjustments. This was an important group as previously explained.



	Jane	Dave	Mark	Dominic	Michelle
First MAE/MD	0.350	0.342	0.340	0.341	0.376
Final MAE/MD	0.402	0.370	0.355	0.347	0.343
Impact of final judgment	-0.052	-0.028	-0.014	-0.006	0.032

Table 4.26. Impact of final judgment on high frequency – high value.

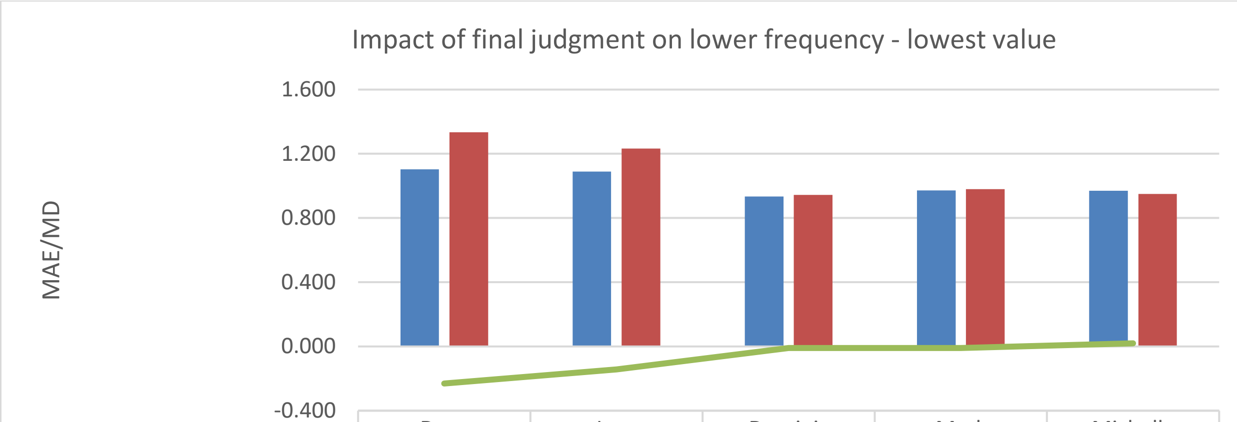
One of the participants made a positive final adjustment. Again, this was an important group.



	Michelle	Mark	Dave	Jane	Dominic
First MAE/MD	0.840	0.748	0.781	0.772	0.814
Final MAE/MD	0.838	0.733	0.742	0.726	0.746
Impact of final judgment	0.002	0.015	0.039	0.046	0.068

Table 4.27. Impact of final judgment on lower frequency – high value.

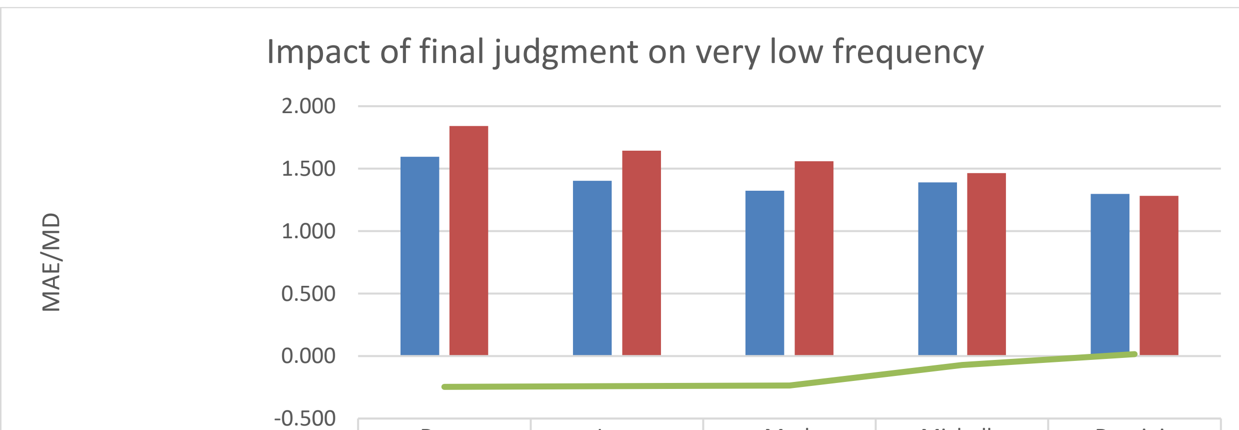
All the final adjustments were positive for this SKU group. This groups stands out as one of only two groups where all the final adjustments were positive.



	Dave	Jane	Dominic	Mark	Michelle
First MAE/MD	1.103	1.090	0.935	0.971	0.969
Final MAE/MD	1.334	1.234	0.945	0.980	0.949
Impact of final judgment	-0.231	-0.144	-0.010	-0.009	0.020

Table 4.28. Impact of final judgment on lower frequency – lowest value.

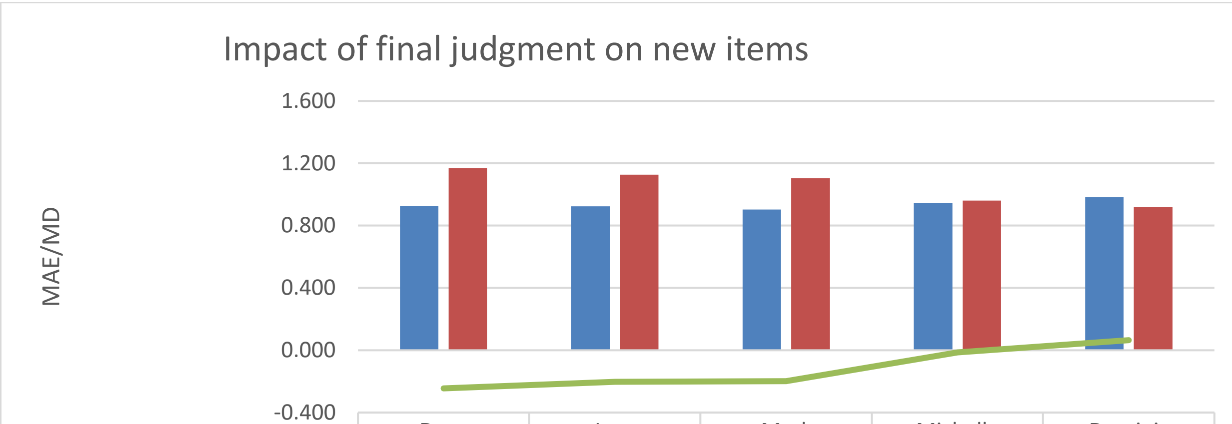
One of the participants made a positive final adjustment.



	Dave	Jane	Mark	Michelle	Dominic
First MAE/MD	1.595	1.402	1.324	1.391	1.299
Final MAE/MD	1.841	1.644	1.558	1.463	1.283
Impact of final judgment	-0.247	-0.241	-0.235	-0.072	0.015

Table 4.29. Impact of final judgment on very low frequency.

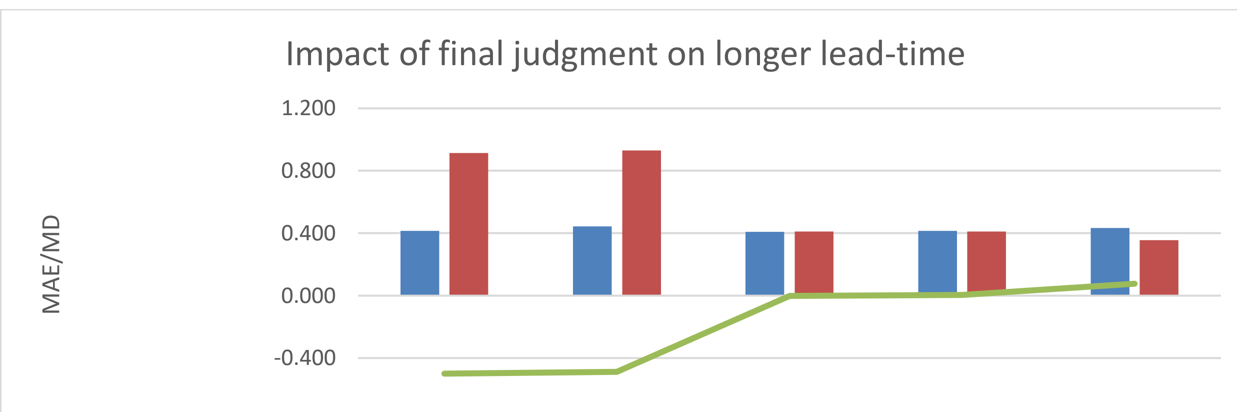
One of the participants made a positive final adjustment.



	Dave	Jane	Mark	Michelle	Dominic
First MAE/MD	0.926	0.924	0.904	0.947	0.983
Final MAE/MD	1.171	1.126	1.103	0.961	0.919
Impact of final judgment	-0.245	-0.202	-0.199	-0.014	0.065

Table 4.30. Impact of final judgment on new items.

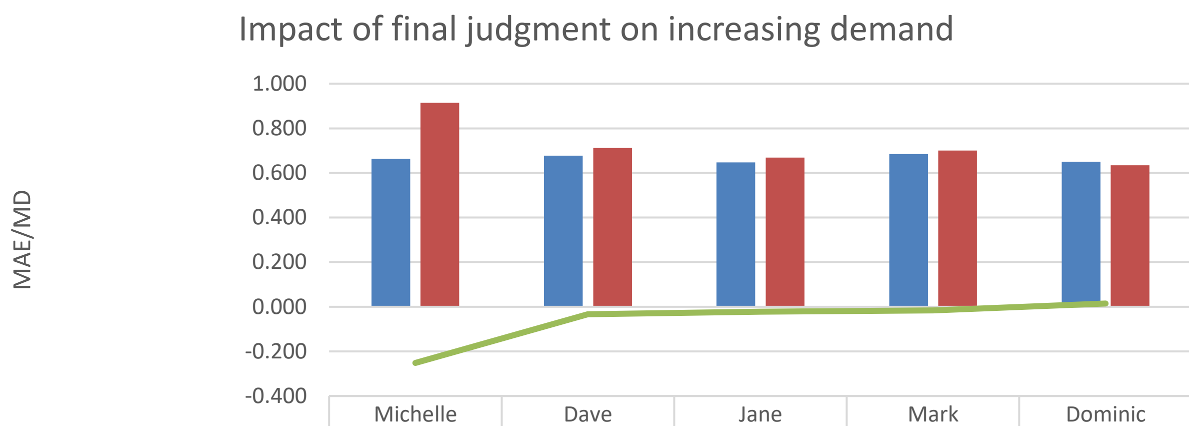
One of the participants made a positive final adjustment.



	Dave	Michelle	Dominic	Jane	Mark
First MAE/MD	0.415	0.444	0.409	0.415	0.433
Final MAE/MD	0.914	0.931	0.411	0.411	0.357
Impact of final judgment	-0.499	-0.488	-0.002	0.004	0.077

Table 4.31. Impact of final judgment on longer lead-time.

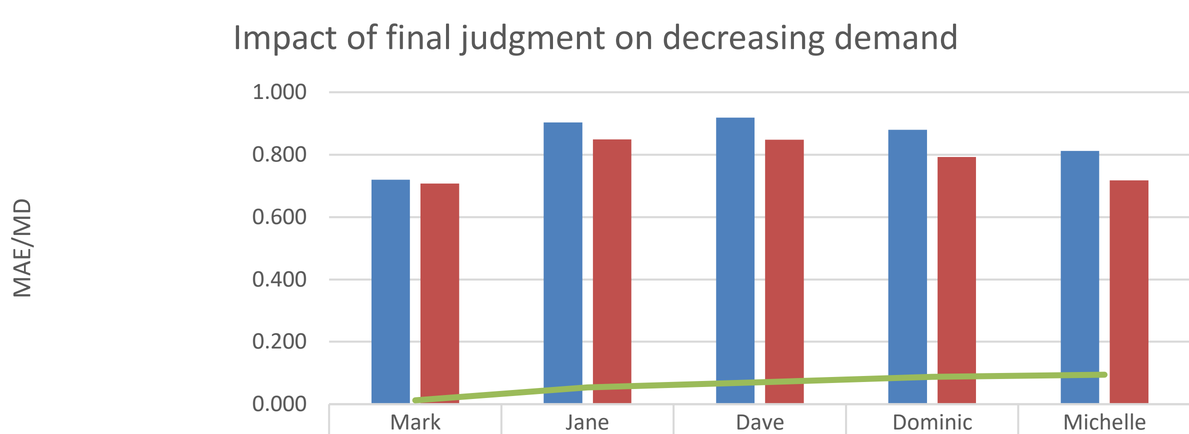
Two of the participants made positive final adjustments.



	Michelle	Dave	Jane	Mark	Dominic
First MAE/MD	0.663	0.678	0.647	0.684	0.649
Final MAE/MD	0.915	0.711	0.669	0.700	0.635
Impact of final judgment	-0.252	-0.034	-0.022	-0.016	0.015

Table 4.32. Impact of final judgment on longer lead-time.

Once of the participants made a positive final adjustment.



	Mark	Jane	Dave	Dominic	Michelle
First MAE/MD	0.720	0.903	0.918	0.880	0.812
Final MAE/MD	0.707	0.849	0.848	0.793	0.717
Impact of final judgment	0.012	0.054	0.070	0.087	0.095

Table 4.33. Impact of final judgment on decreasing demand.

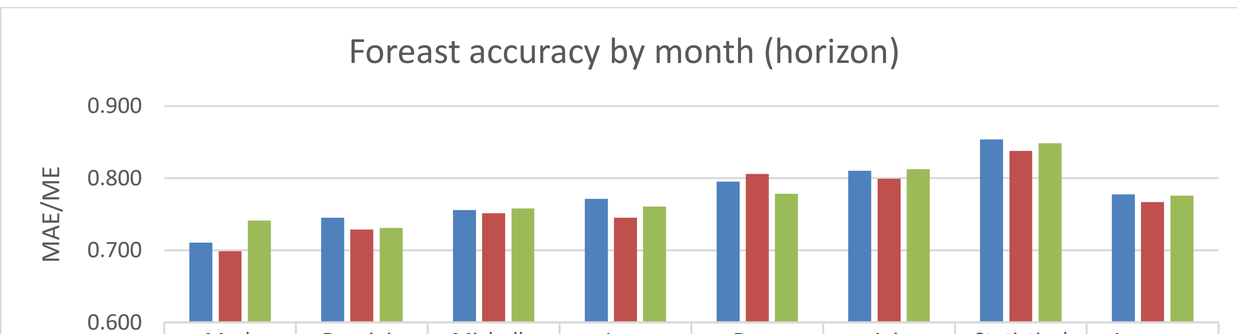
All the participants made positive final adjustments. This was the second group where all the final adjustments were positive.

The fact that the final adjustments were overwhelmingly negative when compared to the judgmentally adjusted forecasts suggests that there was a different reason behind the adjustment than accuracy.

4.4. Does forecast accuracy change with horizon?

By looking at each month's forecast for the next three months at each month the aim was to see if the forecasts were different in accuracy over the 3 horizons. Would participants accuracy improve over the horizon or would it get worse?

The results comparing accuracy change of horizon by participant are shown in Table 4.34.



	Mark	Dominic	Michelle	Jane	Dave	Aris	Statistical	Average
Month 1	0.711	0.745	0.756	0.771	0.795	0.810	0.854	0.777
Month 2	0.698	0.729	0.751	0.745	0.806	0.799	0.838	0.767
Month 3	0.741	0.731	0.758	0.761	0.778	0.813	0.849	0.776

Table 4.34. Forecast accuracy by horizon.

The table shows the average MEA/MD for each of the participants for months 1,2 and 3. There are only small variations at the aggregate level with no clear indication of a meaningful difference.

From the result per SKU groups horizon is a factor relating to forecast accuracy for new items (table 4.19.), longer lead-time (table 4.20.), Increasing demand (table 4.21) and decreasing demand (4.22.). All these SKU groups show the forecast getting progressively worse the further the horizon. No group showed results where the forecast accuracy improved over the 3 months progressively.

4.5. What was the direction of the forecast adjustment?

By recording the direction of the forecast, it is possible to see whether each participant was judgmentally adjusting in one direction more than the other and whether this was linked to expertise or SKU group characteristic.

Table 4.35. shows that the overall number of adjustments and what direction they were in.

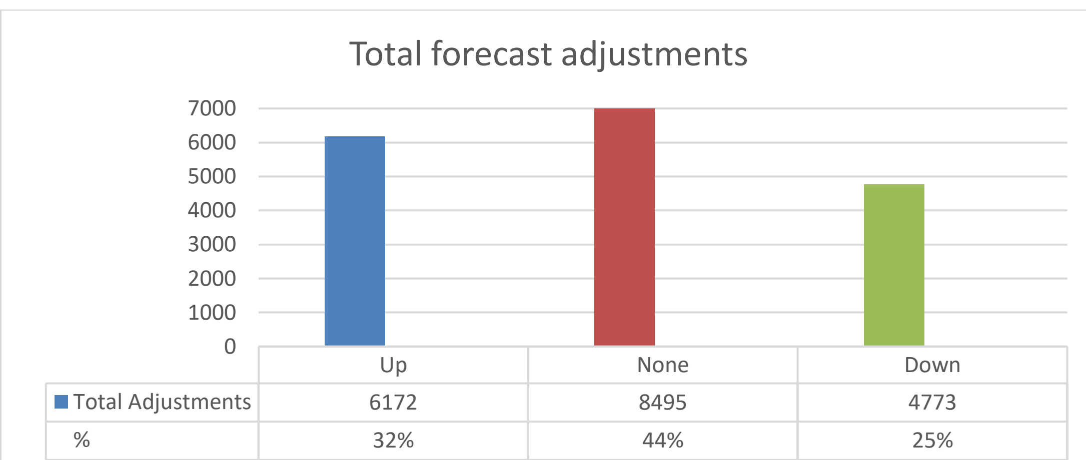
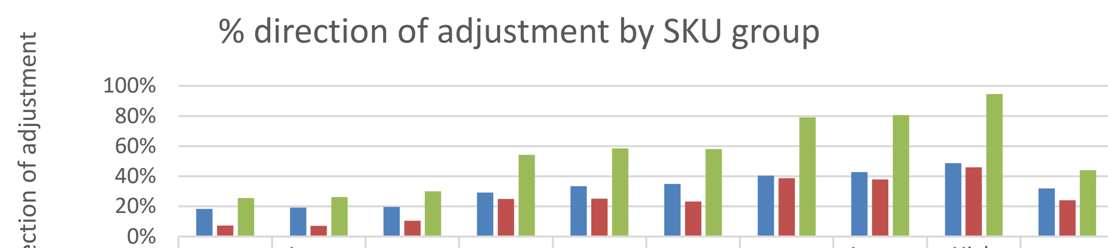


Table 4.35. Total adjustment direction.

The table shows that there were more positive adjustments than negative adjustments. 60% of all statistical adjustments were adjusted.

When this data is reported by the SKU groups (Table 4.36.) it shows that there was variance in the forecasts adjusted per group. The table is expressed a % of the total forecasts available to adjust per group (480 forecasts).

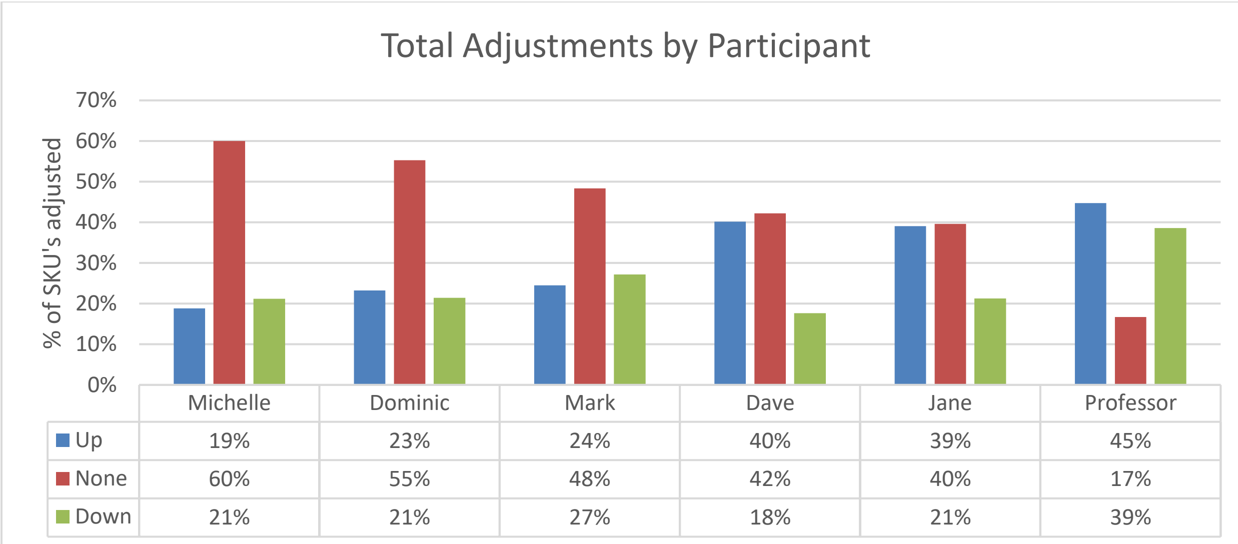


	Very low frequency	Lower frequency - Lowest Value	New items	Decreasing demand	High frequency - Low Value	Increasing demand	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment	18%	19%	20%	29%	33%	35%	40%	43%	49%	32%
Negative adjustment	7%	7%	10%	25%	25%	23%	39%	38%	46%	24%
Total adjustment	26%	26%	30%	54%	58%	58%	79%	81%	95%	44%

Table 4.36. Direction of adjustments by SKU group.

All groups show more positive adjustments than negative ones. Forecasts of high value SKU groups were highly adjusted (levels of 95% and 81%) in comparison to the lowest value SKU group (level of 26%). The range of variance in adjusted levels may reflect the perceived importance of each group to the participant.

There was also a variance in the adjusted forecasts among the participants.

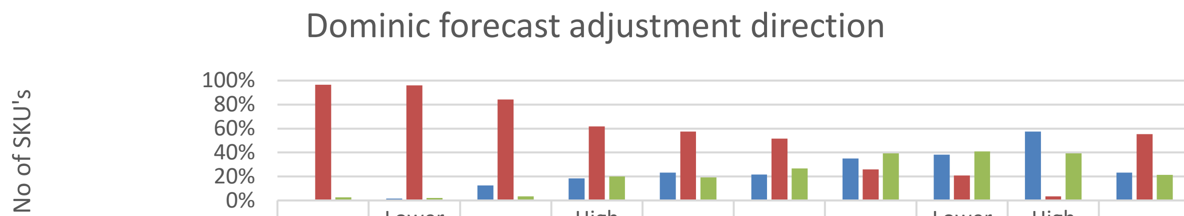


	Michelle	Dominic	Mark	Dave	Jane	Professor
Up	19%	23%	24%	40%	39%	45%
None	60%	55%	48%	42%	40%	17%
Down	21%	21%	27%	18%	21%	39%

Table 4.37. Adjustment totals by participant.

The level of adjustment ranged from 83% to 40%. The most adjustments were made by the offsite academic who applied a different statistical forecast to the data and therefore only where the two statistical forecasts were the same was there no adjustment. Only Michelle and Mark made more negative adjustments than positive ones. The three most experienced company participants made the least amount of adjustments.

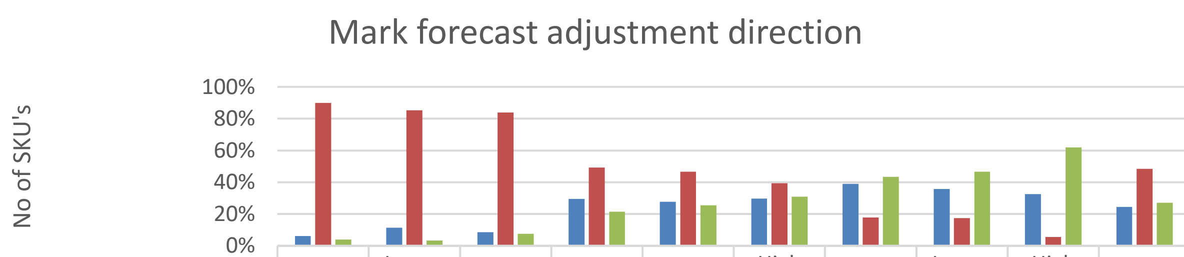
Adjustment by each participant by SKU group was as follows.



	Very low frequency	Lower frequency - Lowest Value	New items	High frequency - Low Value	Increasing demand	Decreasing demand	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment %	1%	2%	13%	18%	23%	22%	35%	38%	58%	23%
No Adjustment %	97%	96%	84%	62%	58%	52%	26%	21%	3%	55%
Negative adjustment %	3%	2%	3%	20%	19%	27%	39%	41%	39%	21%

Table 4.38. Dominic forecast adjustment direction.

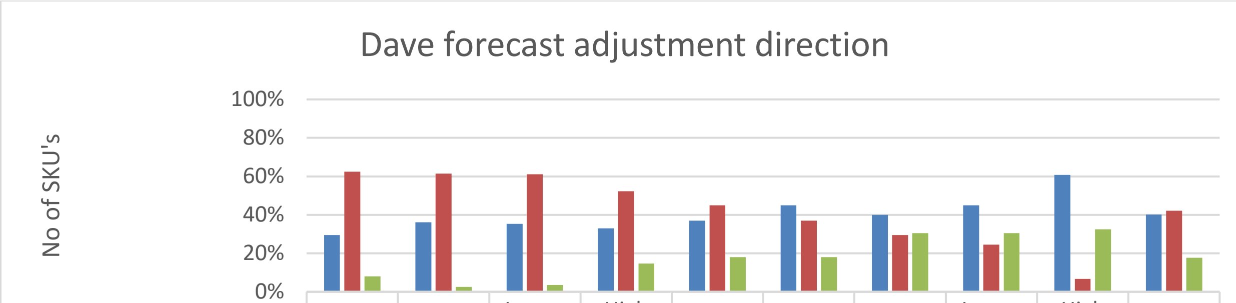
The high frequency – high value SKU group shows a 97% adjustment in contrast to very low frequency SKU group which shows 4%.



	Very low frequency	Lower frequency - Lowest Value	New items	Decreasing demand	Increasing demand	High frequency - Low Value	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment %	6%	11%	9%	29%	28%	30%	39%	36%	33%	24%
No Adjustment %	90%	85%	84%	49%	47%	39%	18%	18%	6%	48%
Negative adjustment %	4%	3%	8%	21%	26%	31%	43%	47%	62%	27%

Table 4.39. Mark forecast adjustment direction.

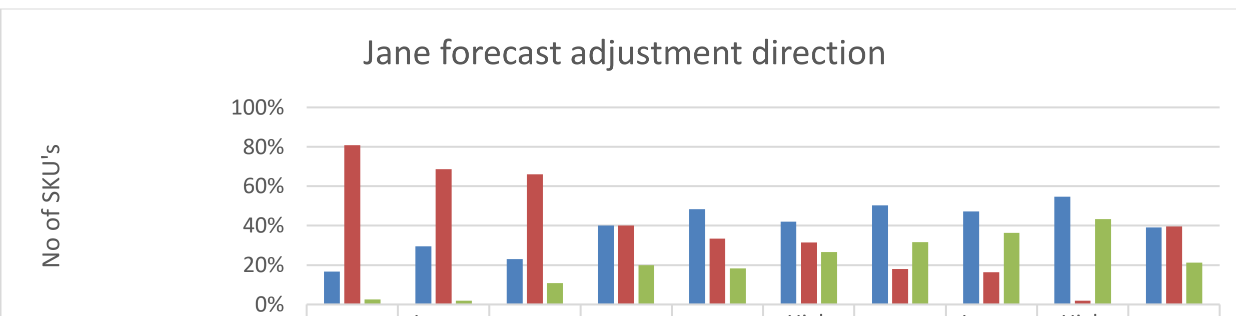
The high frequency – high value SKU group shows a 94% adjustment in contrast to the very low frequency SKU group which showed a 10% adjustment.



	New items	Very low frequency	Lower frequency - Lowest Value	High frequency - Low Value	Decreasing demand	Increasing demand	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment %	29%	36%	35%	33%	37%	45%	40%	45%	61%	40%
No Adjustment %	63%	61%	61%	52%	45%	37%	29%	24%	7%	42%
Negative adjustment %	8%	3%	4%	15%	18%	18%	31%	31%	33%	18%

Table 4.40. Dave forecast adjustment direction.

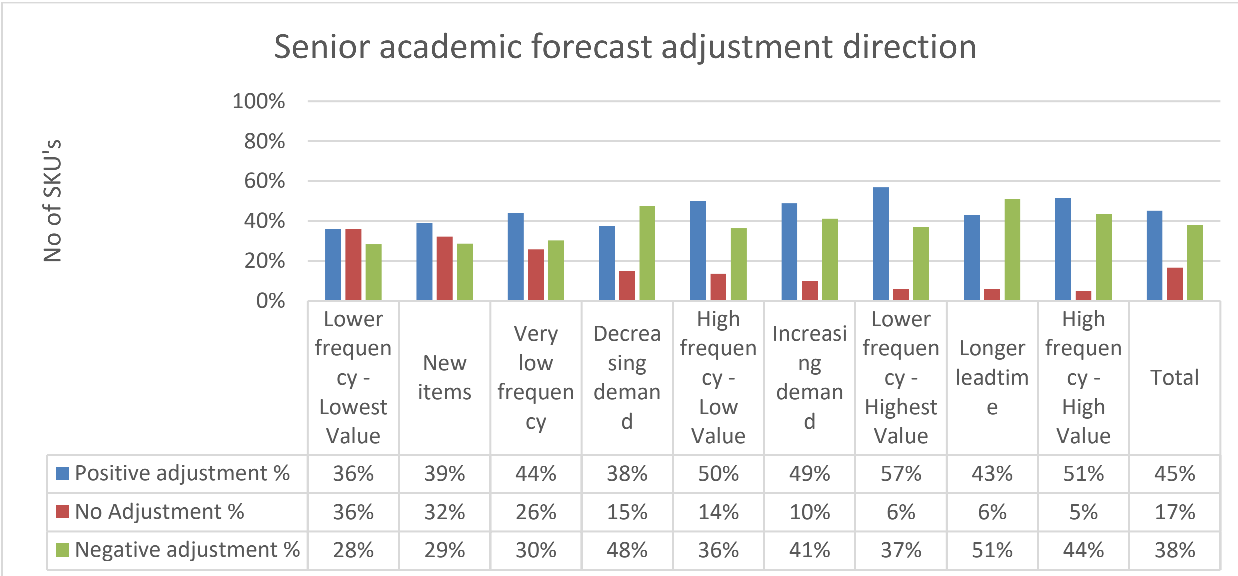
The high value – high frequency SKU group shows a 93% adjustment. The lowest adjusted SKU group for Dave was new items at 37%.



	Very low frequency	Lower frequency - Lowest Value	New items	Decreasing demand	Increasing demand	High frequency - Low Value	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment %	17%	29%	23%	40%	48%	42%	50%	47%	55%	39%
No Adjustment %	81%	69%	66%	40%	33%	31%	18%	16%	2%	40%
Negative adjustment %	3%	2%	11%	20%	18%	27%	32%	36%	43%	21%

Table 4.41. Jane forecast adjustment direction.

The high frequency – high value SKU group shows a 98% adjustment in contrast to the very low frequency SKU group which showed a 19% adjustment.

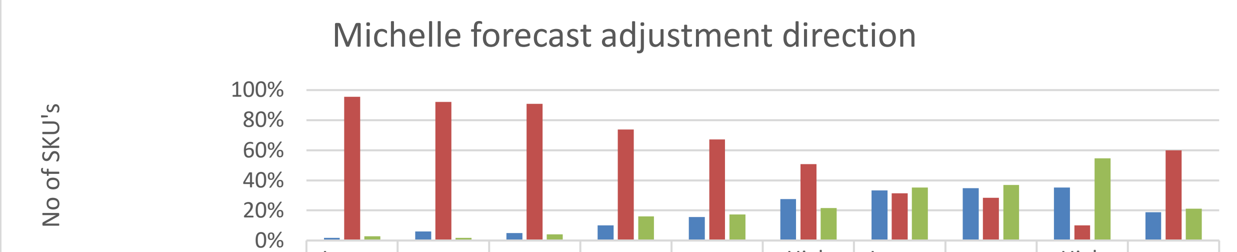


	Lower frequency - Lowest Value	New items	Very low frequency	Decreasing demand	High frequency - Low Value	Increasing demand	Lower frequency - Highest Value	Longer leadtime	High frequency - High Value	Total
Positive adjustment %	36%	39%	44%	38%	50%	49%	57%	43%	51%	45%
No Adjustment %	36%	32%	26%	15%	14%	10%	6%	6%	5%	17%
Negative adjustment %	28%	29%	30%	48%	36%	41%	37%	51%	44%	38%

Table 4.42. Senior academic forecast adjustment direction.

The high frequency – high value SKU group shows a 95% adjustment in contrast to the lower frequency – lowest value SKU group which showed a 64% adjustment.

The total of 17% of no adjustment is much lower than all the company participants.

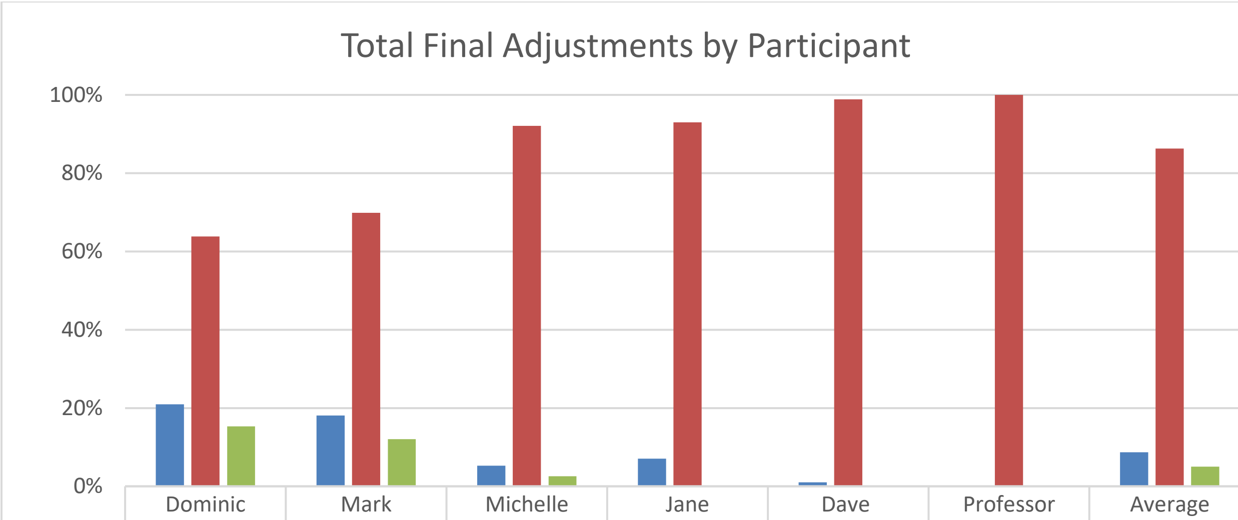


	Lower frequency - Lowest Value	Very low frequency	New items	Decreasing demand	Increasing demand	High frequency - Low Value	Lower frequency - Highest Value	Longer leadtime	High frequency - High Value	Total
Positive adjustment %	2%	6%	5%	10%	16%	28%	33%	35%	35%	19%
No Adjustment %	96%	92%	91%	74%	67%	51%	31%	28%	10%	60%
Negative adjustment %	3%	2%	4%	16%	17%	22%	35%	37%	55%	21%

Table 4.43. Michelle forecast adjustment direction.

The High frequency – High value SKU group shows a 90% adjustment in contrast to the Lowest frequency – Lowest value SKU group which showed a 4% adjustment. Michelle made less adjustments than all the participants at 40% adjusted.

The final forecast made by the company participants showed adjustment direction as below.

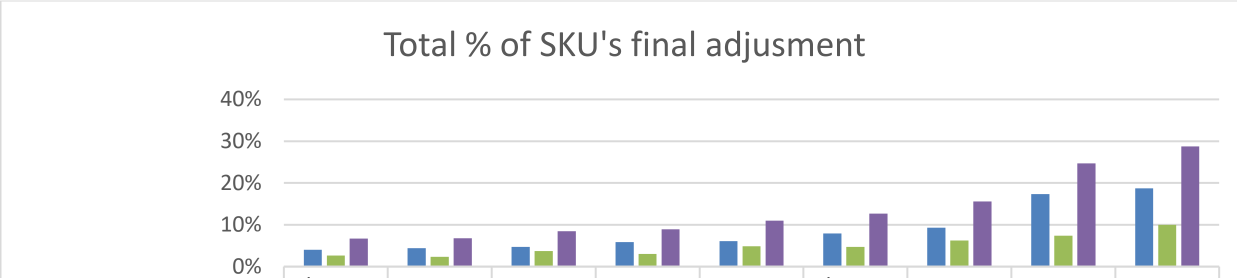


	Dominic	Mark	Michelle	Jane	Dave	Professor	Average
Up	21%	18%	5%	7%	1%	0%	9%
None	64%	70%	92%	93%	99%	100%	86%
Down	15%	12%	3%	0%	0%	0%	5%

Table 4.44. Total final adjustment direction.

The senior academic made no final adjustments. The range between the company participants was high. Dave made 12 adjustments and Dominic made 391 which were 1% and 36% respectively of all forecasts (1080). The graph shows that only Dominic and Mark made significant levels of final adjustments with 36% and 30% respectively the rest of the company participants made less than 8%.

The final adjustment by group was also captured. Note the graphs show the totals of the SKU's that were adjusted per group splitting into positive, negative and total adjustments.



	Lower frequency - Highest Value	Decreasing demand	Very low frequency	New items	Increasing demand	Lower frequency - Lowest Value	Longer leadtime	High frequency - Low Value	High frequency - High Value
Positive adjustment	4%	4%	5%	6%	6%	8%	9%	17%	19%
Negative adjustment	3%	2%	4%	3%	5%	5%	6%	7%	10%
Total adjustment	7%	7%	8%	9%	11%	13%	16%	25%	29%

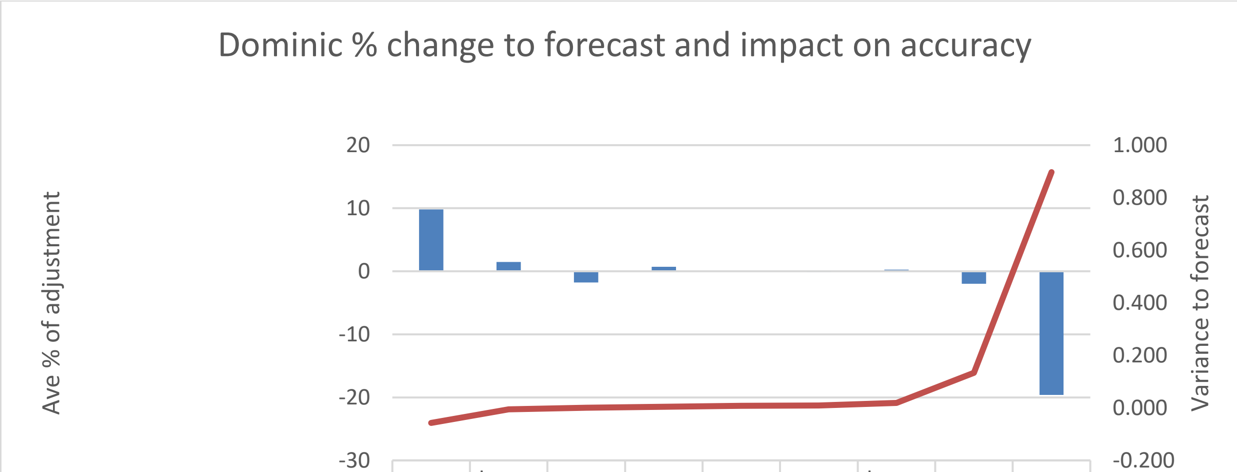
Table 4.45. Total % of SKU final adjustments by group.

The high frequency – high value SKU group was the highest final adjusted group at 29% and the lowest final adjusted SKU group was low frequency – highest value at 7%. The level of final adjustment as an aggregate was much lower than the forecast adjustment with all but two SKU groups adjusted less than a quarter of the SKU's.

4.6. What was the impact on accuracy from the size and direction on the judgmental adjustments?

The question of whether larger adjustments were more successful than smaller adjustments when improving the forecast accuracy was reported. Also, whether the direction of the reduction also have a bearing on the accuracy of the adjustment.

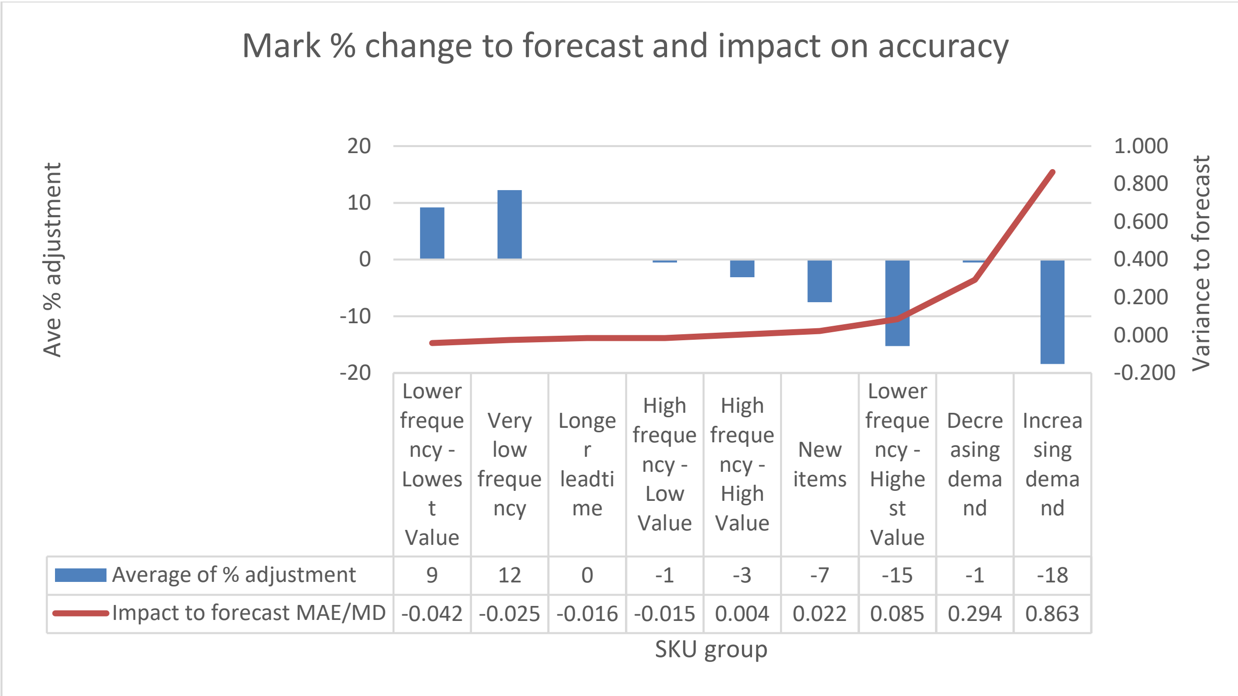
For each participant the average % change was calculated across the SKU groups and then compared to the impact to forecast accuracy (MAE/MD). Where a negative impact is shown the judgmental adjustment made the forecast accuracy worse.



	New items	Lower frequency - Lowest Value	Very low frequency	High frequency - High Value	High frequency - Low Value	Longer leadtime	Lower frequency - Highest Value	Decreasing demand	Increasing demand
Average of % adjustment	10	1	-2	1	0	0	0	-2	-20
Impact to forecast MAE/MD	-0.057	-0.005	0.000	0.004	0.007	0.009	0.019	0.133	0.897

Table 4.46. Dominic % change to forecast and impact on accuracy.

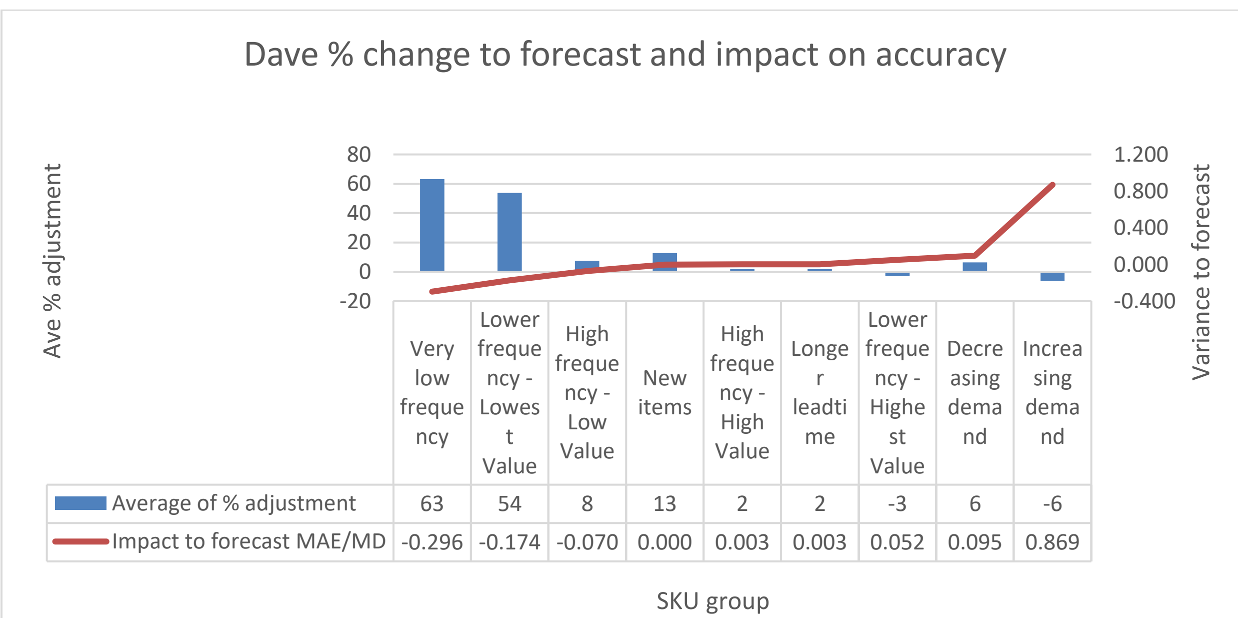
For 6 of the SKU groups the forecast adjustment had a positive effect on forecast accuracy with one group not affected. From the % average adjustments we can see that for Dominic larger negative changes had the most positive impact to forecast accuracy with the reverse for positive ones.



SKU group	Lower frequency - Lowest Value	Very low frequency	Longer leadtime	High frequency - Low Value	High frequency - High Value	New items	Lower frequency - Highest Value	Decreasing demand	Increasing demand
Average of % adjustment	9	12	0	-1	-3	-7	-15	-1	-18
Impact to forecast MAE/MD	-0.042	-0.025	-0.016	-0.015	0.004	0.022	0.085	0.294	0.863

Table 4.47. Mark % change to forecast and impact on accuracy.

For 5 of the groups the forecast adjustment had a positive effect on forecast accuracy. For Mark the most positive adjustments came from the largest negative % change to forecasts.



SKU group	Very low frequency	Lower frequency - Lowest Value	High frequency - Low Value	New items	High frequency - High Value	Longer leadtime	Lower frequency - Highest Value	Decreasing demand	Increasing demand
Average of % adjustment	63	54	8	13	2	2	-3	6	-6
Impact to forecast MAE/MD	-0.296	-0.174	-0.070	0.000	0.003	0.003	0.052	0.095	0.869

Table 4.48. Dave % change to forecast and impact on accuracy.

For 5 of the groups the forecast adjustment had a positive effect on forecast accuracy. Dave's% forecast adjustments were predominantly positive with the largest positive % adjustments bringing about the worse affects to forecast accuracy.

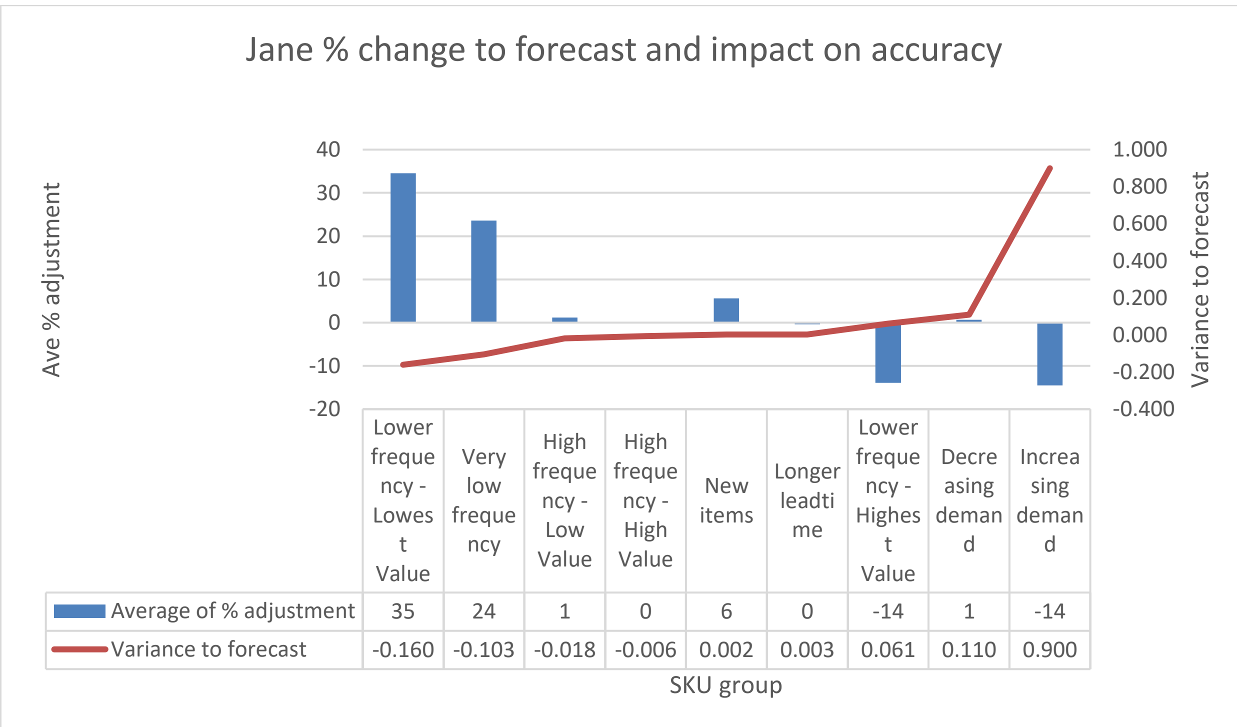


Table 4.49. Jane % change to forecast and impact on accuracy.

For 5 of the groups the forecast adjustment had a positive effect on forecast accuracy. For Jane, the highest positive adjustments provided the worse accuracy results and the largest negative adjustments the best ones.

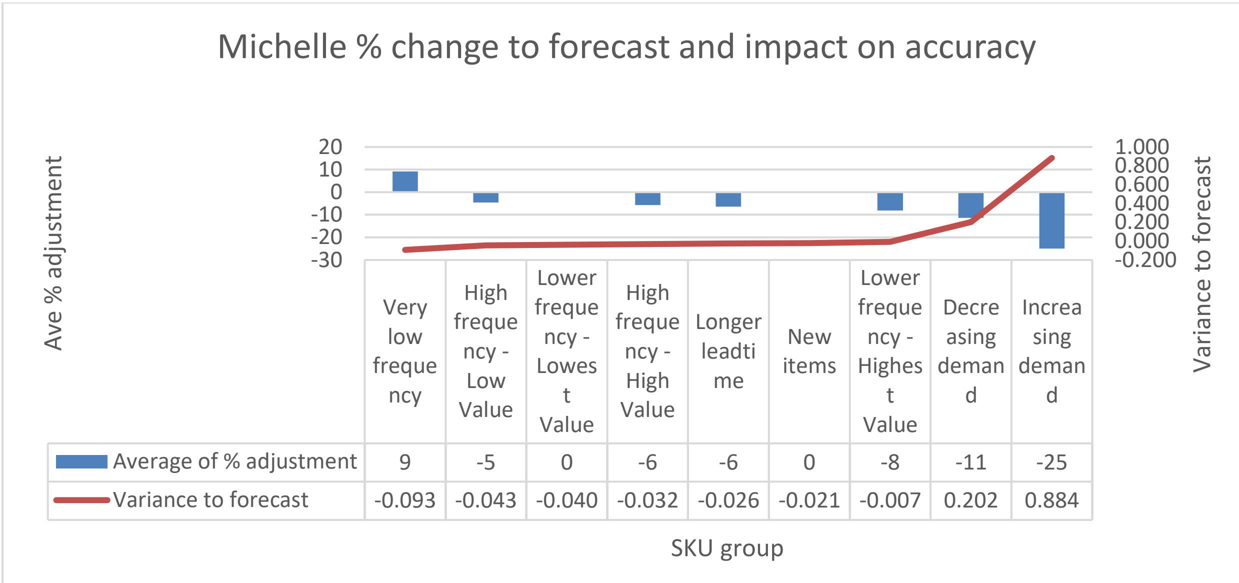


Table 4.50. Michelle % change to forecast and impact on accuracy.

For 2 of the groups the forecast adjustment had a positive effect on forecast accuracy. Michelle's results show the largest positive affect from the highest negative adjustments.

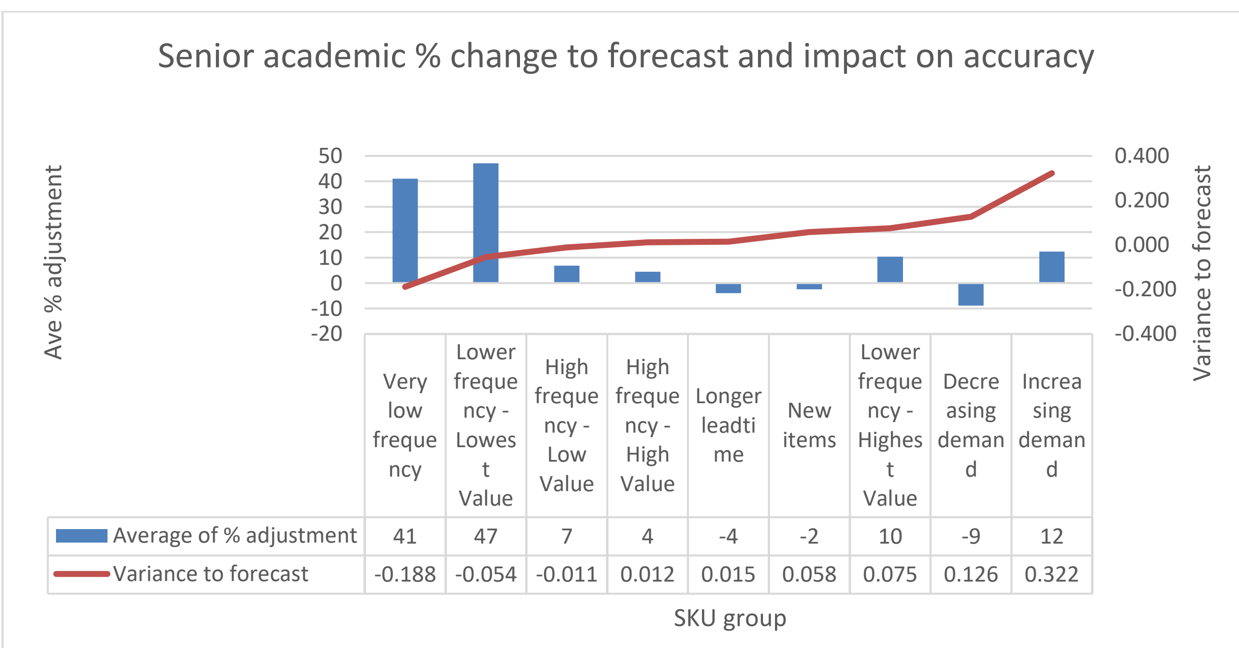


Table 4.51. Senior academic % change to forecast and impact on accuracy

For 6 of the groups the forecast adjustment had a positive effect on forecast accuracy. The Senior academic's results were not as clear as the other participants. Although the largest positive % changes provided the worse accuracy adjustments the best improvement to forecast accuracy was also provided by a positive adjustment.

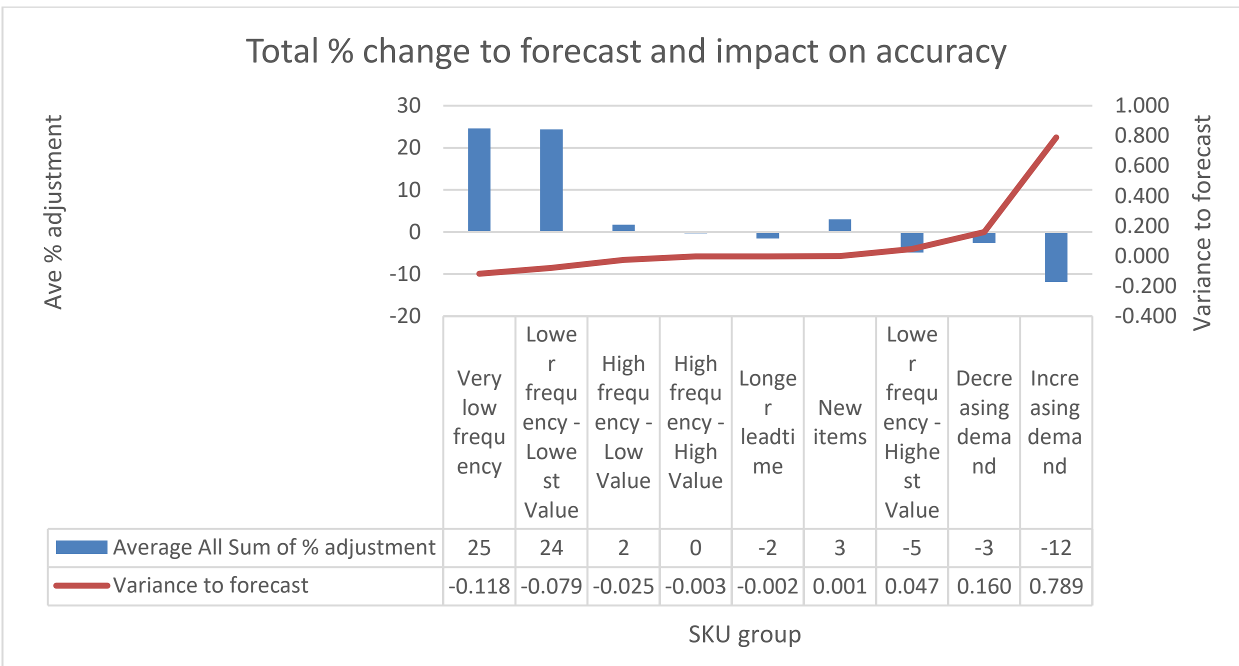
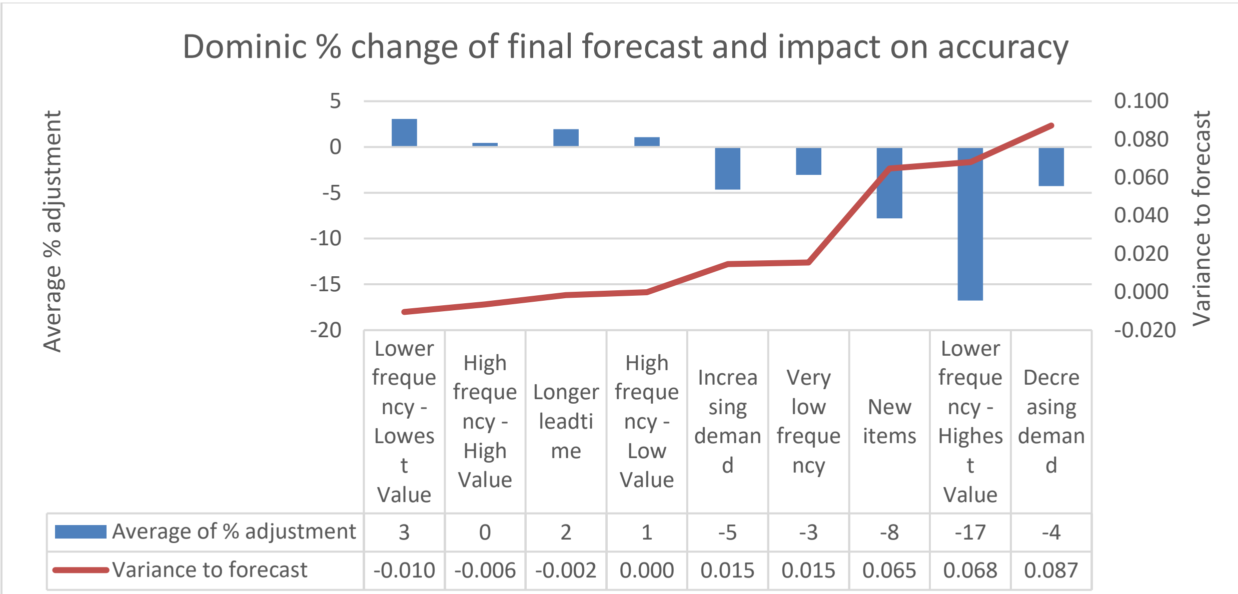


Table 4.52. Aggregate % change to forecast and impact on accuracy

When the results were aggregated by SKU group 4 of the groups showed an improvement to forecast accuracy after adjustment. The table clearly shows that the larger the positive % change the worse the effect on forecast accuracy and vice versa.

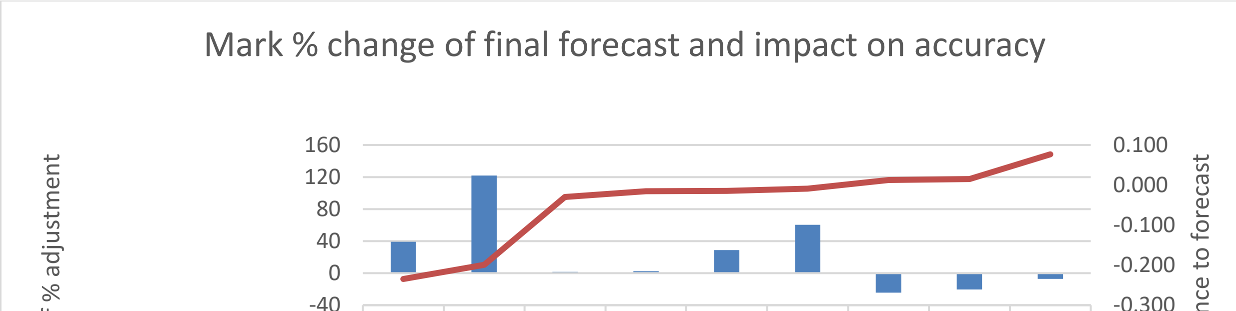
The impact from the final adjustment was shown by participant and the impact to the MAE/MD in the same format (negatives being a worsening of the statistical forecast).



	Lower frequency - Lowest Value	High frequency - High Value	Longer leadtime	High frequency - Low Value	Increasing demand	Very low frequency	New items	Lower frequency - Highest Value	Decreasing demand
Average of % adjustment	3	0	2	1	-5	-3	-8	-17	-4
Variance to forecast	-0.010	-0.006	-0.002	0.000	0.015	0.015	0.065	0.068	0.087

Table 4.53. Dominic Impact of final adjustment

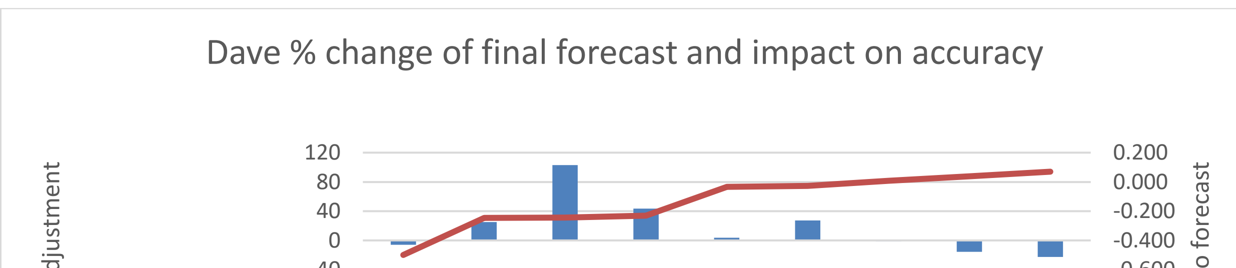
For 5 of the groups the final adjustment had a positive effect on forecast accuracy. The table conforming to the judgmental adjustment results.



	Very low frequency	New items	High frequency - Low Value	Increasing demand	High frequency - High Value	Lower frequency - Lowest Value	Decreasing demand	Lower frequency - Highest Value	Longer leadtime
Average of % adjustment	39	122	2	2	29	60	-24	-20	-7
Variance to forecast	-0.235	-0.199	-0.030	-0.016	-0.014	-0.009	0.012	0.015	0.077

Table 4.54. Mark Impact of final adjustment.

For 3 of the groups the final adjustment had a positive effect on forecast accuracy. The results conformed to the judgmental adjustments table.



	Longer leadtime	Very low frequency	New items	Lower frequency - Lowest Value	Increasing demand	High frequency - High Value	High frequency - Low Value	Lower frequency - Highest Value	Decreasing demand
Average of % adjustment	-6	25	103	44	4	27	-1	-16	-23
Variance to forecast	-0.499	-0.247	-0.245	-0.231	-0.034	-0.028	0.009	0.039	0.070

Table 4.55. Dave Impact of final adjustment.

For 3 of the groups the final adjustment had a positive effect on forecast accuracy. Dave's results in general conformed to the judgmental adjustments result table.

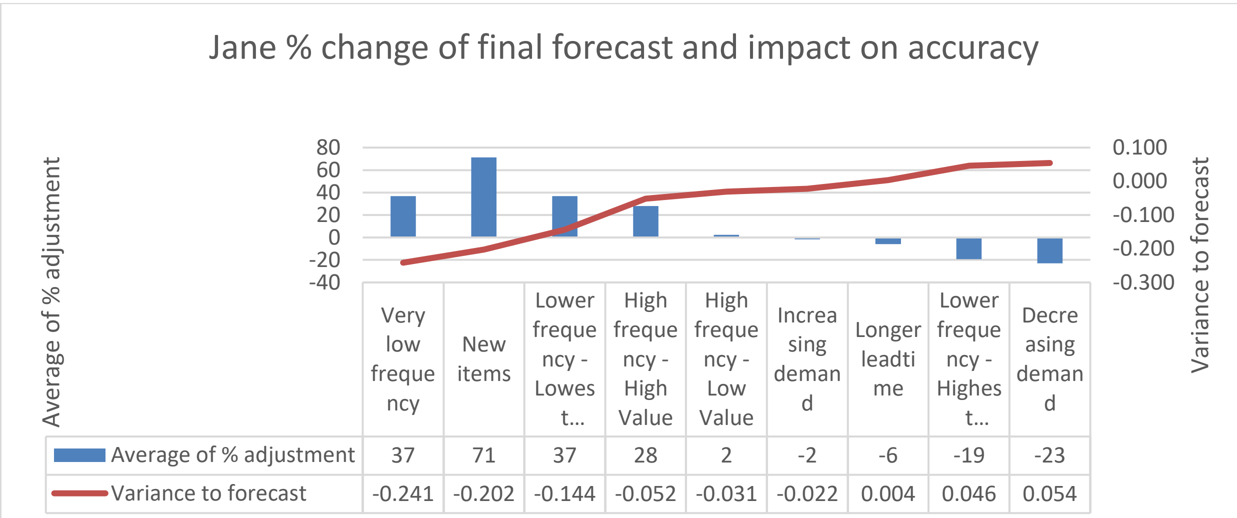


Table 4.56. Jane Impact of final adjustment.

For 3 of the groups the final adjustment had a positive effect on forecast accuracy. Jane's results conformed to the judgmental adjusted results.

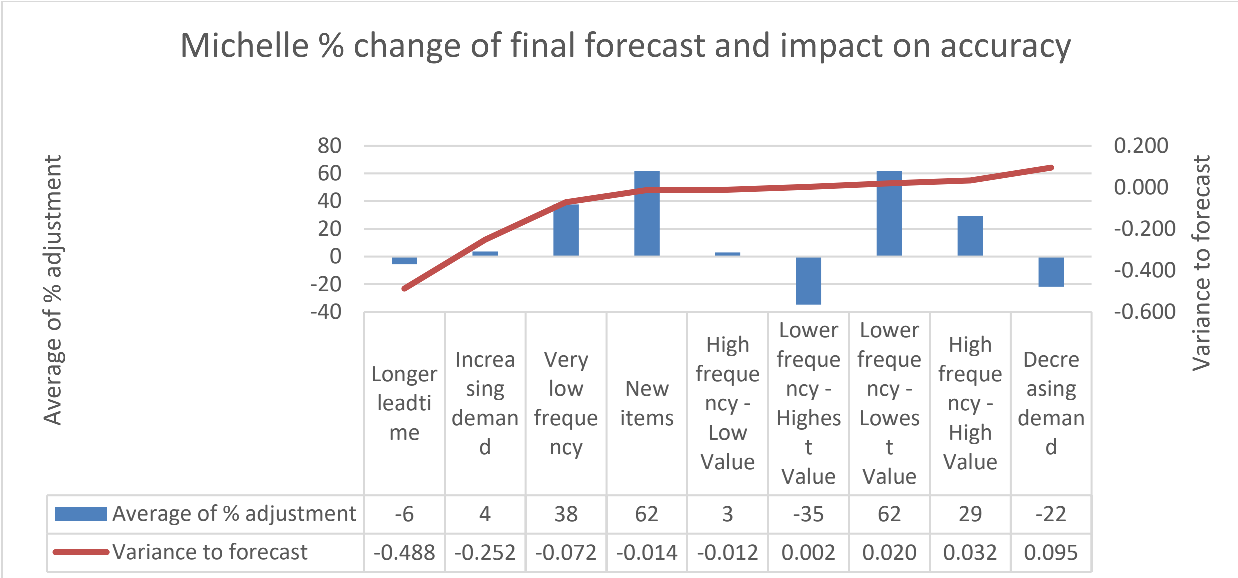


Table 4.57. Michelle impact of final adjustment.

For 4 of the groups the final adjustment had a positive effect on forecast accuracy. Michelle's results were erratic. Some large positive % changes were positive, and some were negative. The worse result was also a negative % change.

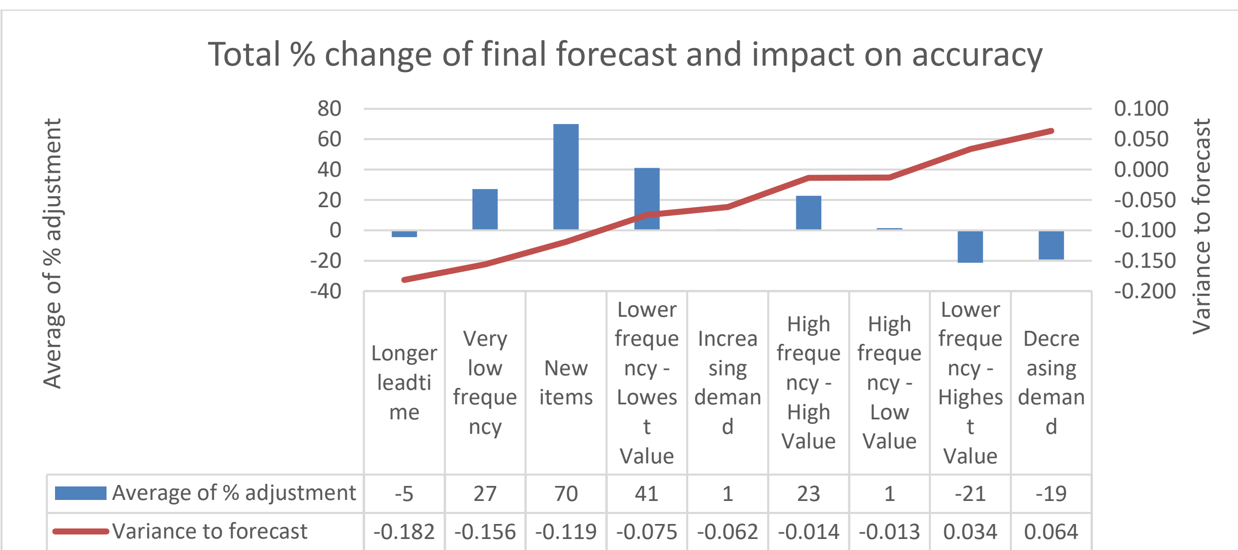
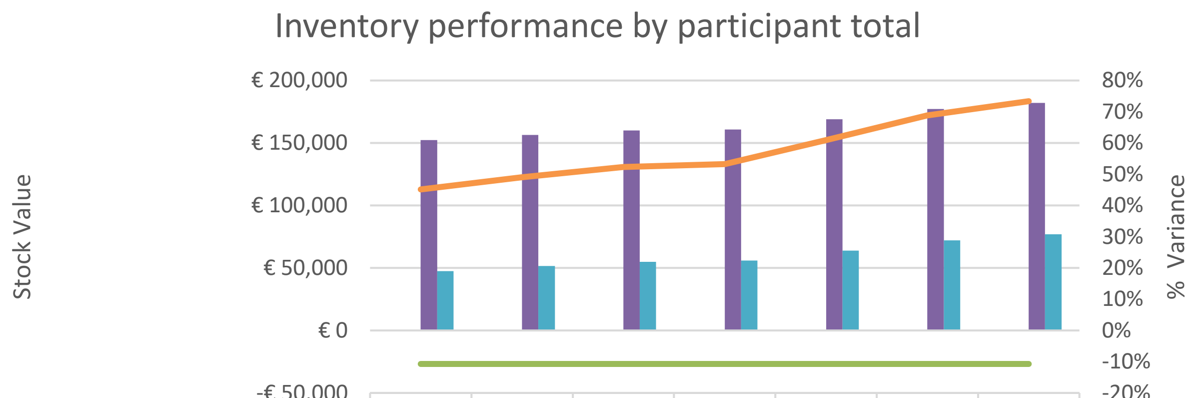


Table 4.58. Aggregated impact of final adjustment.

When the results were aggregated by SKU group 2 of the groups showed an improvement to forecast accuracy after adjustment. The aggregated results showed that the large negative % changes had a positive effect on the final forecast. The large positive % changes generally had a negative impact to forecast accuracy however three positive % adjustments did have small improvements to the accuracy. The largest impact was a negative % change with was also negative to the forecast accuracy.

4.7. Impact of adjustment on inventory

To compare the stock level of the SKU's at the beginning and the end of the experiment it is necessary to show as a ratio of the target stock. The target stock for each SKU was calculated by the company using a set number of weeks multiplied by the average of the last 12 weeks' demand. Although the company reclassified usually every 6 months the SKU's in the experiment were not changed so the target stock remained the same number of weeks.



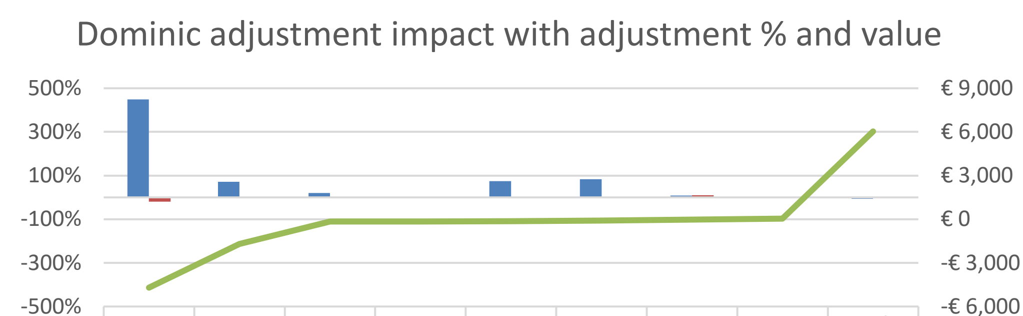
	Michelle	Mark	Dominic	Statistic	Dave	Jane	Professor
Start of Period	€ 138,891	€ 138,891	€ 138,891	€ 138,891	€ 138,891	€ 138,891	€ 138,891
Variance start	-€ 16,677	-€ 16,677	-€ 16,677	-€ 16,677	-€ 16,677	-€ 16,677	-€ 16,677
End of Period	€ 152,398	€ 156,544	€ 159,928	€ 160,855	€ 168,948	€ 177,219	€ 182,049
Variance end	€ 47,399	€ 51,544	€ 54,928	€ 55,855	€ 63,949	€ 72,219	€ 77,049
% Variance from Target Start	-11%	-11%	-11%	-11%	-11%	-11%	-11%
% Variance from Target End	45%	49%	52%	53%	61%	69%	73%

Table 4.59. Stock values at the start and at the end of the experiment.

All participants started with a stock value that was 89% of the target stock value. After the 12 - month experiment this percentage had increased for all the participants (ranging from 145% to 173%). The target stock value at the end of the experiment had changed to €104,999 due to a reduction in overall SKU targets.

The inventory performance by each participant was compared by SKU group to the statistical forecast. Where the % variance from the target stock was higher than the statistical forecast variance then from an inventory perspective the participant performed worse against target stock.

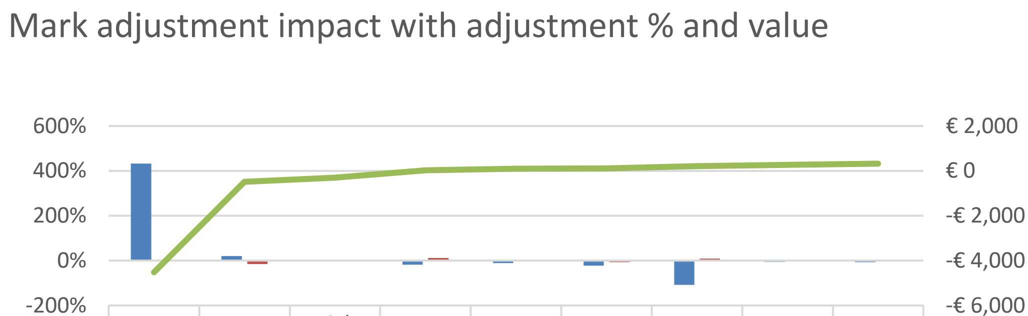
For Dominic for example, the target stock was €89,851 at the end period of the experiment. Dominic's stock total was €159,928 (Table 4.59) compared to the statistical forecast result of €160,855. The results versus target were 52% and 53% more than target. That meant that Dominic's adjustments for the SKU groups had performed 1% better than the statistical forecast and improved the variance from target stock by €927.



	Increasing demand	Lower frequency - Highest Value	Decreasing demand	Longer leadtime	Lower frequency - Lowest Value	Very low frequency	New items	High frequency - Low Value	High frequency - High Value
Dominic impact	449%	72%	20%	3%	74%	83%	8%	-1%	-7%
Average adjustment	-19.63%	0.24%	-2.00%	-0.13%	1.49%	-1.78%	9.80%	-0.01%	0.74%
Value of impact	-€ 4,705	-€ 1,705	-€ 156	-€ 154	-€ 141	-€ 96	-€ 36	€ 29	€ 6,036

Table 4.60. Dominic inventory impact to target stock by SKU group.

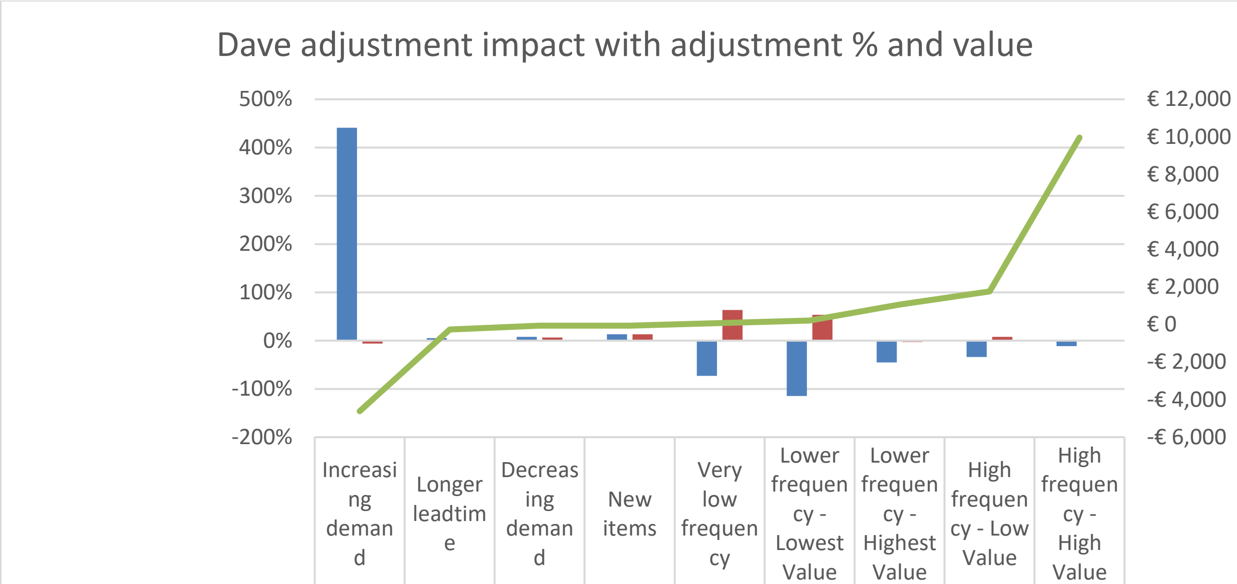
Table 4.60. shows the impact that Dominic had on the inventory value for each group. Dominic impact is the % improvement over the statistic forecast. In the case of Increasing demand, for example, Dominic's forecast was 449% closer to target stock than the statistical forecast. The average adjustment is the average adjustment made expressed as a % by Dominic to the increasing demand SKU group. The value of impact is the value of the difference between Dominic's forecast and the statistical forecast. The -€4,705 is the amount closer to the target value Dominic's adjusted forecast was. In the following graphs any negative figure is an improvement to the statistical forecast inventory position and positive is a worsening of the statistical inventory position.



	Increasing demand	Lower frequency - Highest Value	High frequency - High Value	Very low frequency	Decreasing demand	New items	Lower frequency - Lowest Value	Longer leadtime	High frequency - Low Value
Mark impact	432%	20%	0%	-18%	-11%	-23%	-108%	-5%	-6%
Average adjustment	-0.18%	-15.25%	-3.10%	12.23%	-0.53%	-7.50%	9.17%	-0.07%	-0.50%
Value of impact	-€ 4,523	-€ 487	-€ 306	€ 21	€ 89	€ 104	€ 204	€ 268	€ 321

Table 4.61. Mark inventory impact to target stock by SKU group.

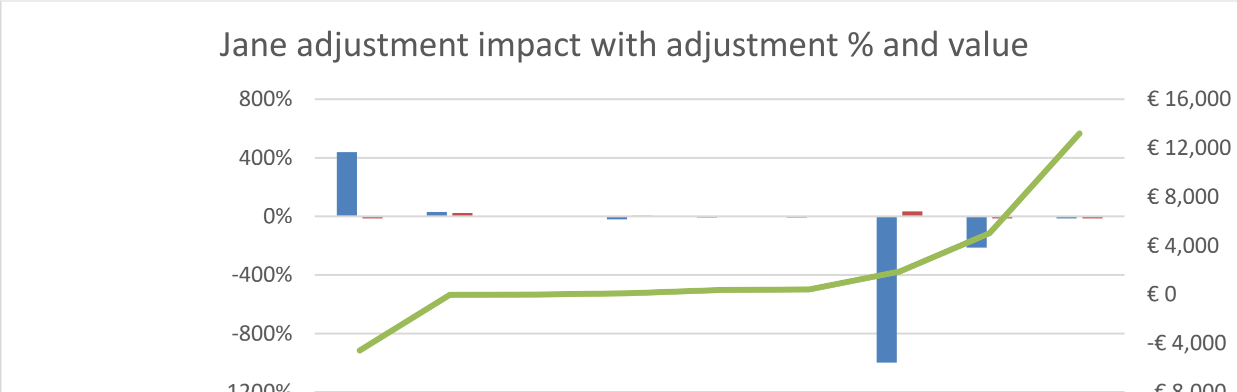
There are three SKU groups where the inventory position is improved compared to the statistical forecast. The value of the improvement of the increasing demand SKU group is much larger than the other figures. The high frequency – High Value group is improved very slightly unlike most of the other participants where this groups is significantly negatively impacted. Marks impact to most of the groups is relatively small apart from the increasing demand group which was positive.



	Increasing demand	Longer leadtime	Decreasing demand	New items	Very low frequency	Lower frequency - Lowest Value	Lower frequency - Highest Value	High frequency - Low Value	High frequency - High Value
Dave impact	441%	5%	8%	13%	-73%	-115%	-45%	-34%	-11%
Average adjustment	-6.27%	1.86%	6.45%	12.82%	63.25%	53.88%	-2.93%	7.59%	1.77%
Value of impact	-€ 4,618	-€ 262	-€ 63	-€ 58	€ 85	€ 217	€ 1,072	€ 1,758	€ 9,963

Table 4.62. Dave inventory impact to stock by SKU group.

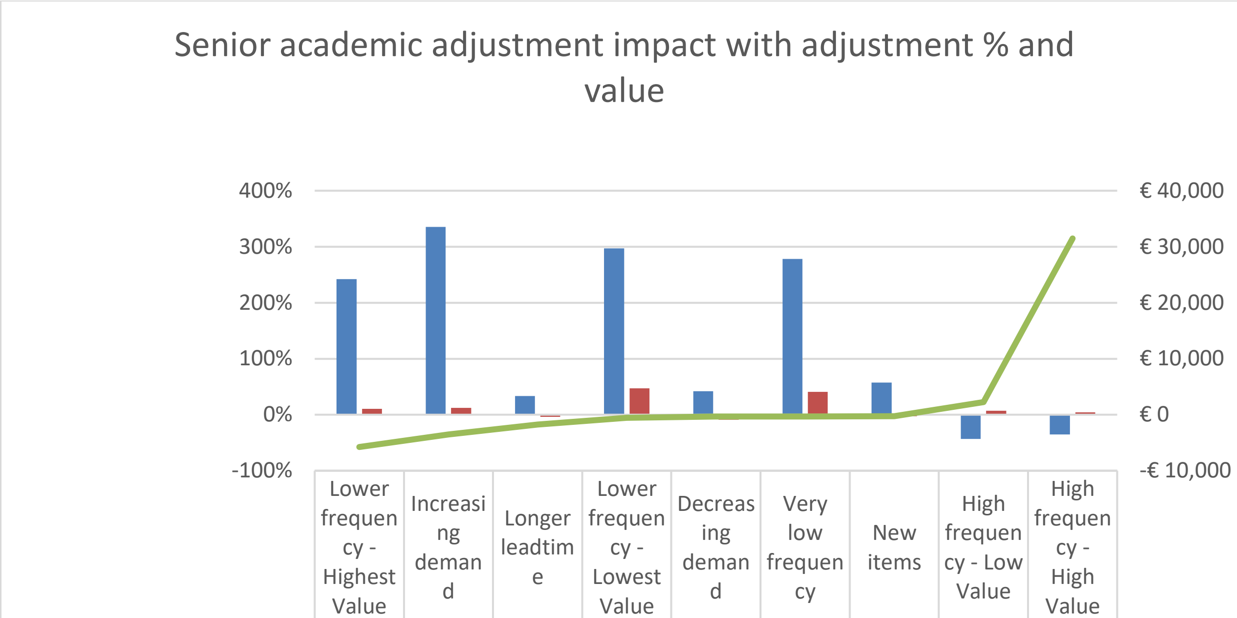
There are four SKU groups where the inventory position is improved compared to the statistical forecast. Although the Increasing demand group is improved it is shown that the High frequency – High value is made much worse which outweighs the positive impact from the Increasing demand. Comparatively Dave has a large impact on 4 of the groups with three being negative.



	Increasing demand	Very low frequency	Decreasing demand	New items	High frequency - Low Value	Longer leadtime	Lower frequency - Lowest Value	Lower frequency - Highest Value	High frequency - High Value
Jane impact	438%	31%	1%	-21%	-7%	-8%	-998%	-212%	-15%
Average adjustment	-14.47%	23.59%	0.68%	5.59%	1.19%	-0.38%	34.51%	-13.93%	-14.76%
Value of impact	-€ 4,591	-€ 35	-€ 9	€ 95	€ 361	€ 420	€ 1,887	€ 5,038	€ 13,199

Table 4.63. Jane inventory impact to stock by SKU group.

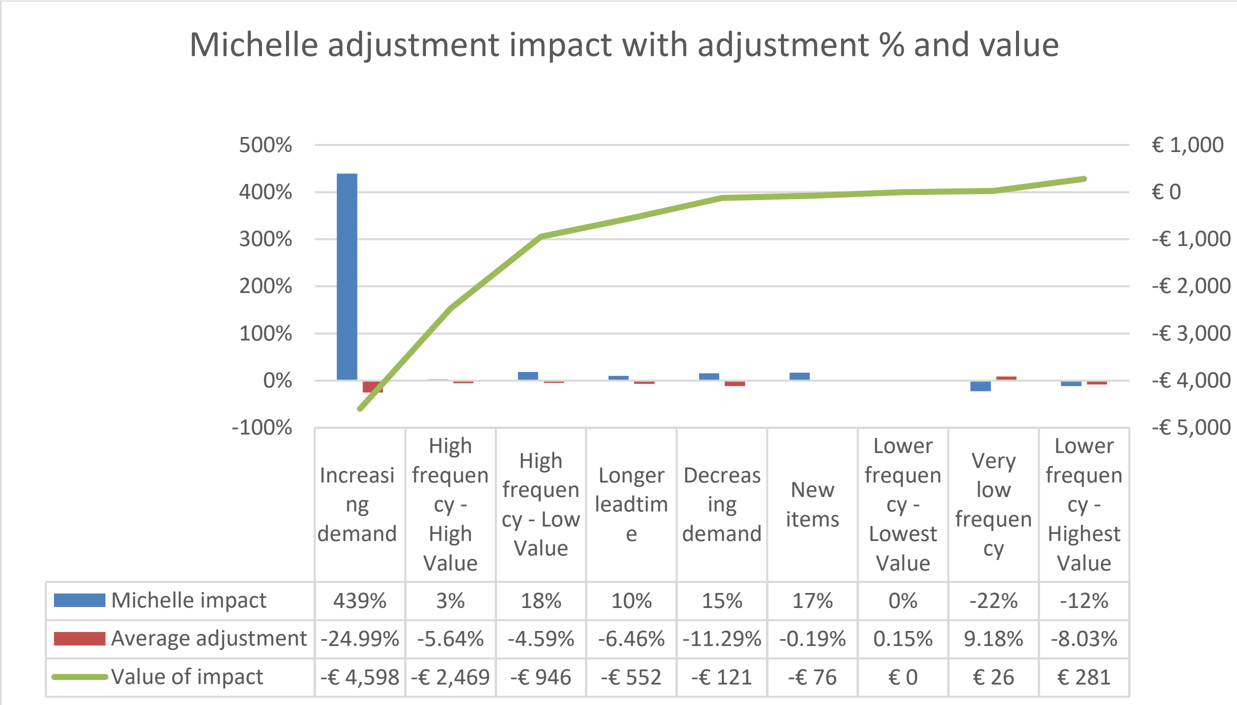
There are three SKU groups where the inventory position is improved compared to the statistical forecast. The same result here with the Increasing demand group positively impacted but the High frequency – High value group being much more negatively affected. Jane has a high impact to three of the groups increasing demand positively and the High value groups negatively. Overall a negative impact.



	Lower frequency - Highest Value	Increasing demand	Longer leadtime	Lower frequency - Lowest Value	Decreasing demand	Very low frequency	New items	High frequency - Low Value	High frequency - High Value
Professor impact	242%	335%	33%	297%	42%	278%	57%	-44%	-35%
Average adjustment	10.42%	12.37%	-3.97%	47.02%	-8.91%	41.06%	-2.46%	6.82%	4.49%
Value of impact	-€ 5,763	-€ 3,511	-€ 1,773	-€ 562	-€ 327	-€ 322	-€ 260	€ 2,237	€ 31,475

Table 4.64. Senior academic inventory impact to stock by SKU group.

There are seven SKU groups where the inventory position is improved compared to the statistical forecast. The overall result of the Senior academic is the worse compared to the other participants, but we can see that for all but two groups the results were a positive one. The High frequency – High value groups was so negatively affected it rendered the previous good results irrelevant when overall inventory value was compared.



	Increasing demand	High frequency - High Value	High frequency - Low Value	Longer leadtime	Decreasing demand	New items	Lower frequency - Lowest Value	Very low frequency	Lower frequency - Highest Value
Michelle impact	439%	3%	18%	10%	15%	17%	0%	-22%	-12%
Average adjustment	-24.99%	-5.64%	-4.59%	-6.46%	-11.29%	-0.19%	0.15%	9.18%	-8.03%
Value of impact	-€ 4,598	-€ 2,469	-€ 946	-€ 552	-€ 121	-€ 76	€ 0	€ 26	€ 281

Table 4.65. Michelle academic inventory impact to stock by SKU group.

There are 6 SKU groups where the inventory position is improved compared to the statistical forecast. Michelle performed the best out of all the participants for the High frequency - High value group which coupled with the positive impact to the increasing demand group made her the nest overall from an inventory perspective.

Typically, the value of the high frequency – high value SKU group adhered to the Pareto rule and accounted for 85% of the inventory total. If performance was good in this group, then the overall inventory result was significantly influenced.

Using the same ranking system as Table 4.23. showing the SKU groups in order of positive impact from the adjustments to inventory target Table 4.66. shows the Increasing demand as the clearly the group most beneficially affected.

SKU group	Dom	Mark	Dave	Jane	Professor	Michelle	Sum
Increasing demand	1	1	1	1	2	1	7
Decreasing demand	3	5	3	3	5	5	24
Longer leadtime	4	8	2	6	3	4	27
Lower frequency - Highest Value	2	2	7	8	1	9	29
Very low frequency	6	4	5	2	6	8	31
New items	7	6	4	4	7	6	34
Lower frequency - Lowest Value	5	7	6	7	4	7	36
High frequency - Low Value	8	9	8	5	8	3	41
High frequency - High Value	9	3	9	9	9	2	41

Table 4.66 SKU group ranking by positive inventory impact versus target.

The Increasing demand group is the most positively affected group for all the participants apart from one (where it was 2nd). This is to be expected due to the large improvements made by the judgmental adjustments shown as an average in Table 4.52. All the participants could identify that the statistical forecast did not produce an optimum suggestion. The Senior academic's alternative statistical forecast was also an improvement from an inventory target perspective.

The two worse performers were the high frequency groups where the statistical forecast performed reasonably well. Table 4.14. shows the statistical forecast ranked 2nd for High frequency – Low value group and Table 4.15. shows the statistical forecast ranked 5th for the High frequency – High value group (although the margins of separation were very close for this group).

Intuitively the above result would suggest that the high frequency groups should have produced a better result regarding the impact on inventory target however as the forecast were MAE/MD across all the SKU's in the specific group per period this can mask some forecast errors individually per month. For example, a higher value SKU from the group could be continually the worse forecast but as we are looking at the average errors over all the groups SKU's this can produce a different result when looking at forecast accuracy as oppose to inventory performance.

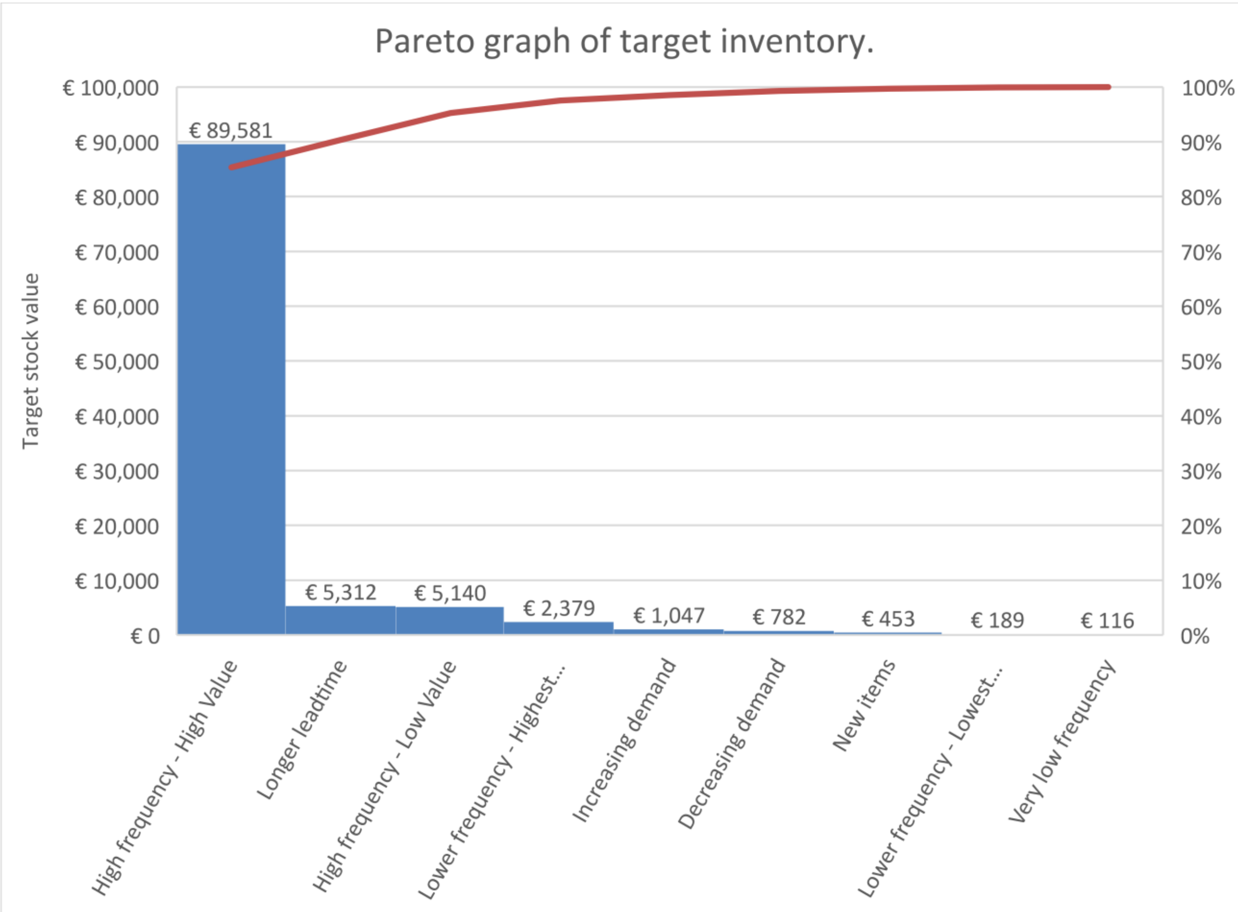
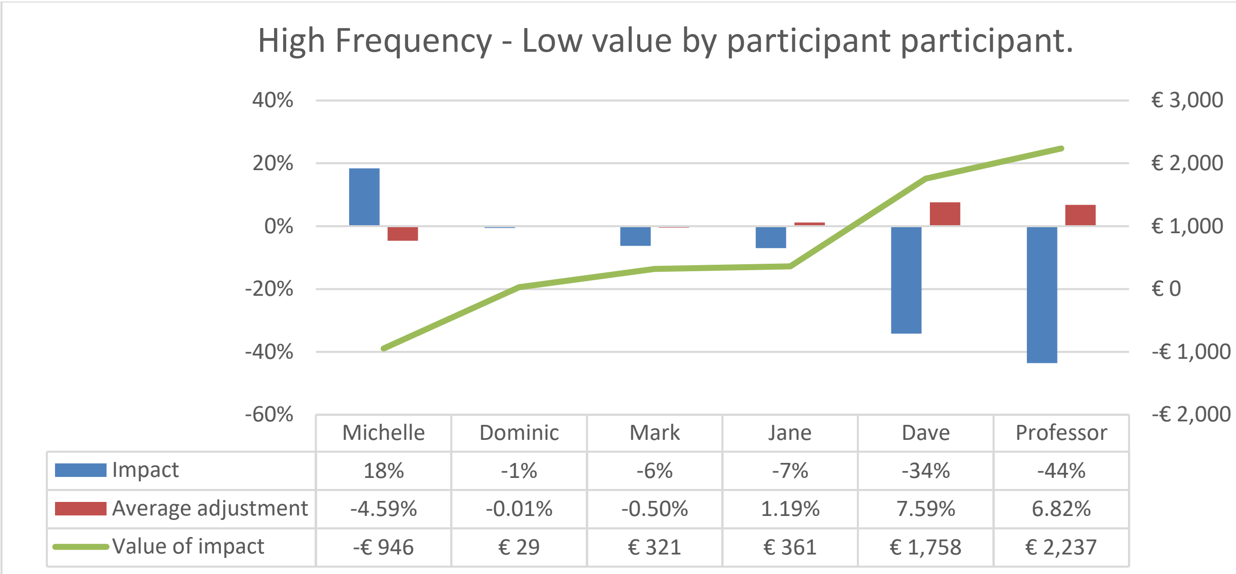


Table 4.67. Pareto graph of SKU group inventory targets.

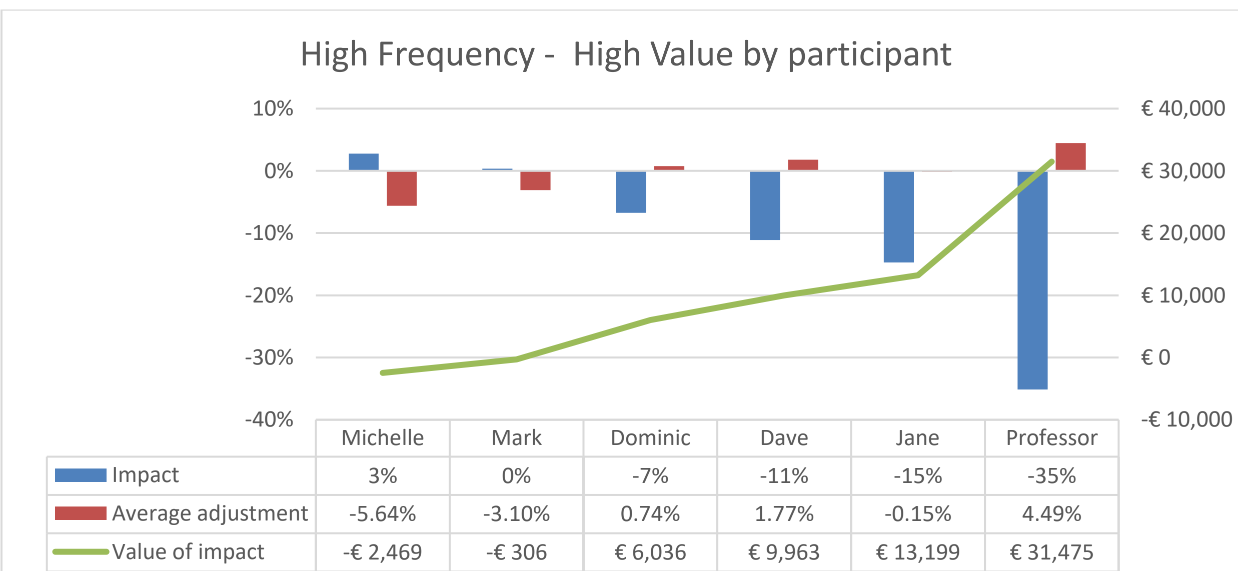
The impact to inventory was also plotted by SKU group.



	Michelle	Dominic	Mark	Jane	Dave	Professor
Impact	18%	-1%	-6%	-7%	-34%	-44%
Average adjustment	-4.59%	-0.01%	-0.50%	1.19%	7.59%	6.82%
Value of impact	-€ 946	€ 29	€ 321	€ 361	€ 1,758	€ 2,237

The aggregated effects of judgment on the high frequency – low value SKU group show a worsening from the statistical forecast of €3,760.

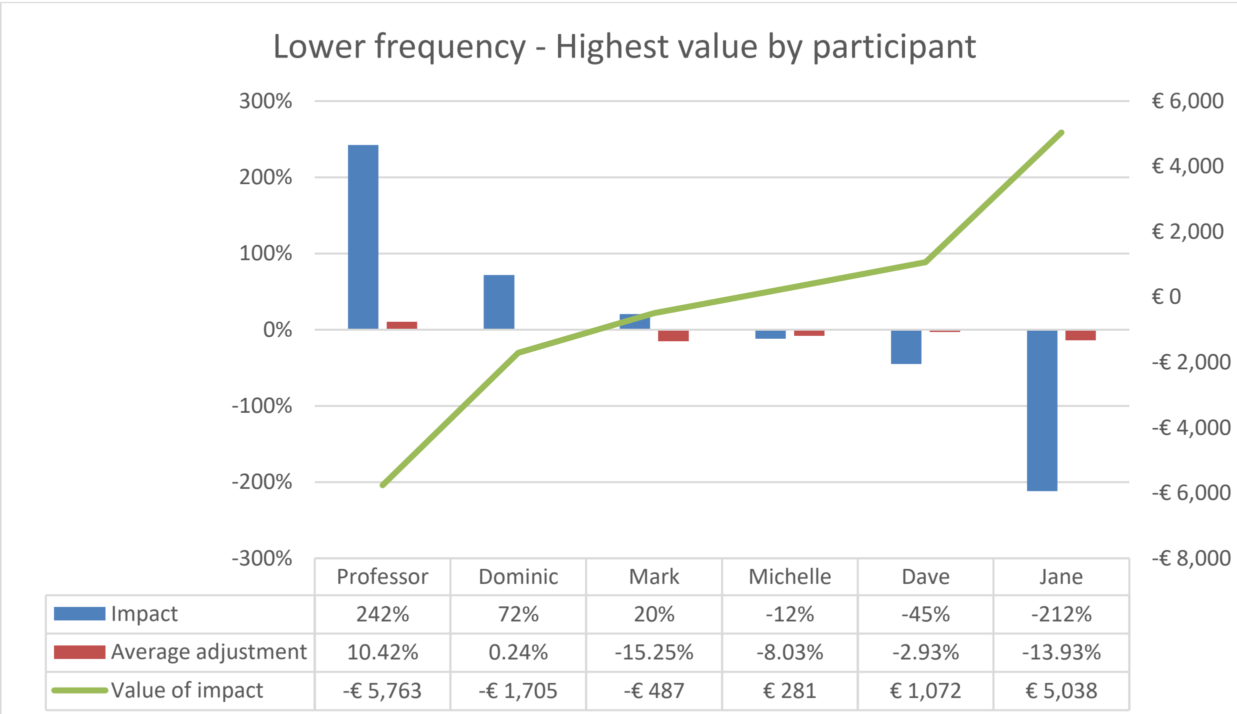
Table 4.68. Participant impact to target stock High frequency – Low Value.



	Michelle	Mark	Dominic	Dave	Jane	Professor
Impact	3%	0%	-7%	-11%	-15%	-35%
Average adjustment	-5.64%	-3.10%	0.74%	1.77%	-0.15%	4.49%
Value of impact	-€ 2,469	-€ 306	€ 6,036	€ 9,963	€ 13,199	€ 31,475

Table 4.69. Participant impact to target stock High frequency – High value.

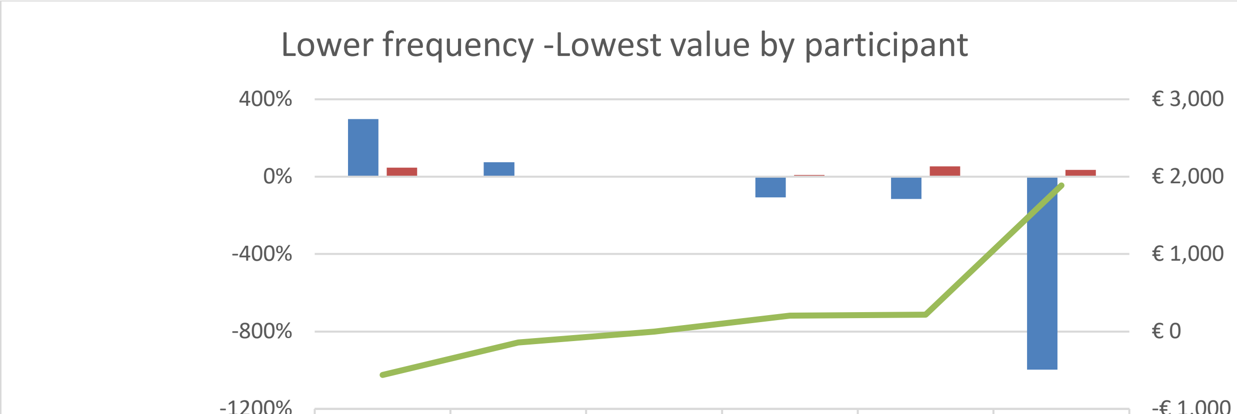
The aggregated effects of judgment on the high frequency – high value SKU group show a worsening from the statistical forecast of €57,898. Only Mark and Michelle made an improvement to this group.



	Professor	Dominic	Mark	Michelle	Dave	Jane
Impact	242%	72%	20%	-12%	-45%	-212%
Average adjustment	10.42%	0.24%	-15.25%	-8.03%	-2.93%	-13.93%
Value of impact	-€ 5,763	-€ 1,705	-€ 487	€ 281	€ 1,072	€ 5,038

Table 4.70. Participant impact to target stock Lower frequency – Highest value.

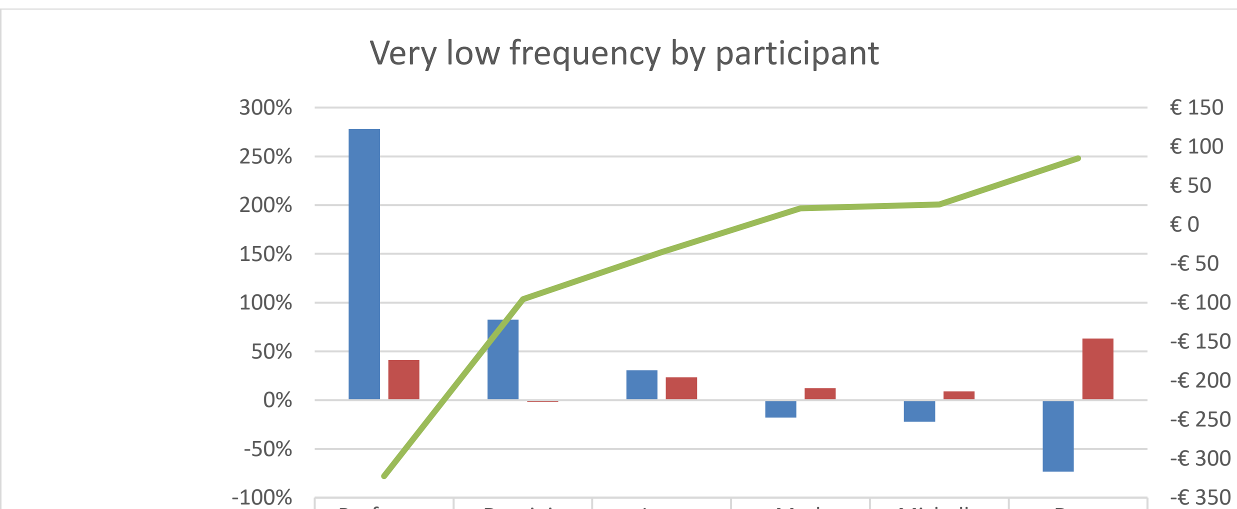
The aggregated effects of judgment on the Lower frequency – Highest value SKU group show an improvement to the statistical forecast of €1,564.



	Professor	Dominic	Michelle	Mark	Dave	Jane
Impact	297%	74%	0%	-108%	-115%	-998%
Average adjustment	47.02%	1.49%	0.15%	9.17%	53.88%	34.51%
Value of impact	-€ 562	-€ 141	€ 0	€ 204	€ 217	€ 1,887

Table 4.71. Participant impact to target stock Lower frequency – Lowest value.

The aggregated effects of judgment on the Lower frequency – lowest value SKU group show a worsening from the statistical forecast of €1,605.



	Professor	Dominic	Jane	Mark	Michelle	Dave
Impact	278%	83%	31%	-18%	-22%	-73%
Average adjustment	41.06%	-1.78%	23.59%	12.23%	9.18%	63.25%
Value of impact	-€ 322	-€ 96	-€ 35	€ 21	€ 26	€ 85

Table 4.72. Participant impact to target stock Very low frequency.

The aggregated effects of judgment on the Very low frequency SKU group show an improvement to the statistical forecast of €322.

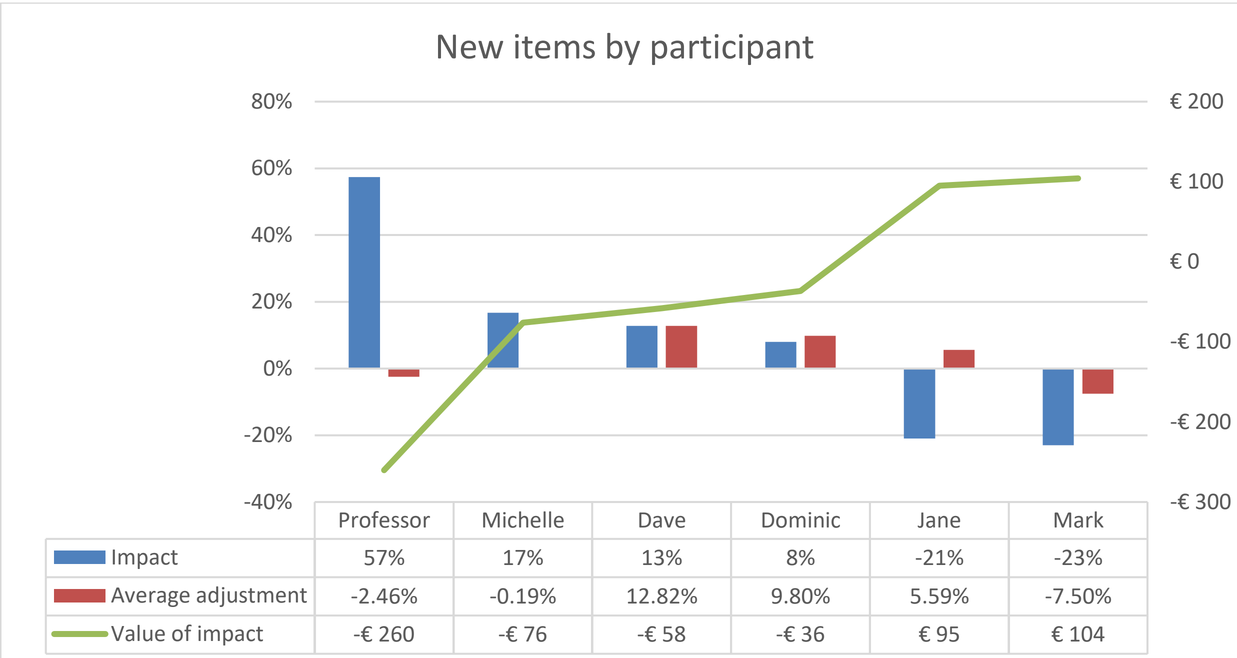
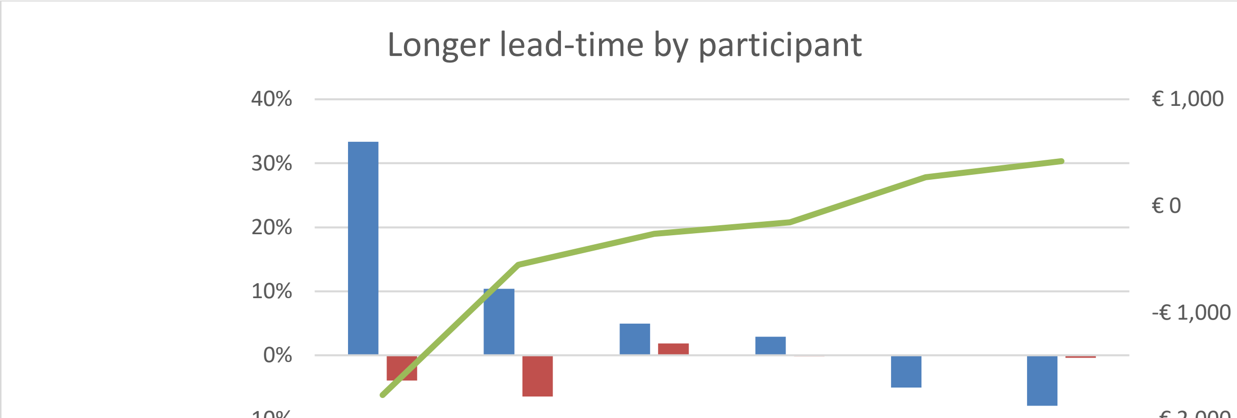


Table 4.73 Participant impact to target stock New Items.

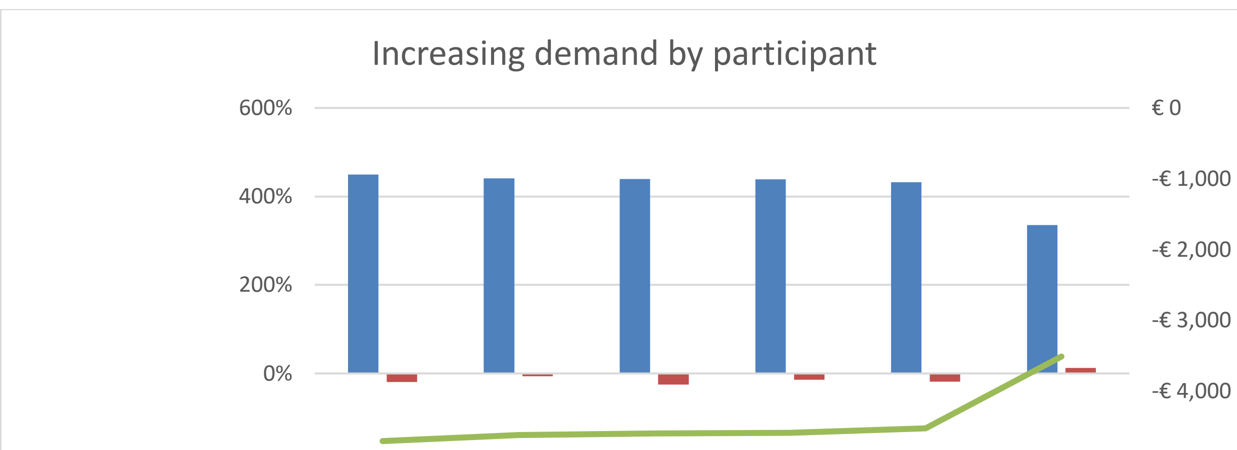
The aggregated effects of judgment on the New item's SKU group show an improvement to the statistical forecast of €232.



	Professor	Michelle	Dave	Dominic	Mark	Jane
Impact	33%	10%	5%	3%	-5%	-8%
Average adjustment	-3.97%	-6.46%	1.86%	-0.13%	-0.07%	-0.38%
Value of impact	-€ 1,773	-€ 552	-€ 262	-€ 154	€ 268	€ 420

Table 4.74. Participant impact to target stock Longer lead-time.

The aggregated effects of judgment on the Longer lead-time SKU group show an improvement to the statistical forecast of €2,053.



	Dominic	Dave	Michelle	Jane	Mark	Professor
Impact	449%	441%	439%	438%	432%	335%
Average adjustment	-19.63%	-6.27%	-24.99%	-14.47%	-18.40%	12.37%
Value of impact	-€ 4,705	-€ 4,618	-€ 4,598	-€ 4,591	-€ 4,523	-€ 3,511

Table 4.75. Participant impact to target stock Increasing demand.

The aggregated effects of judgment on the Increasing demand SKU group show an improvement to the statistical forecast of €26,546.

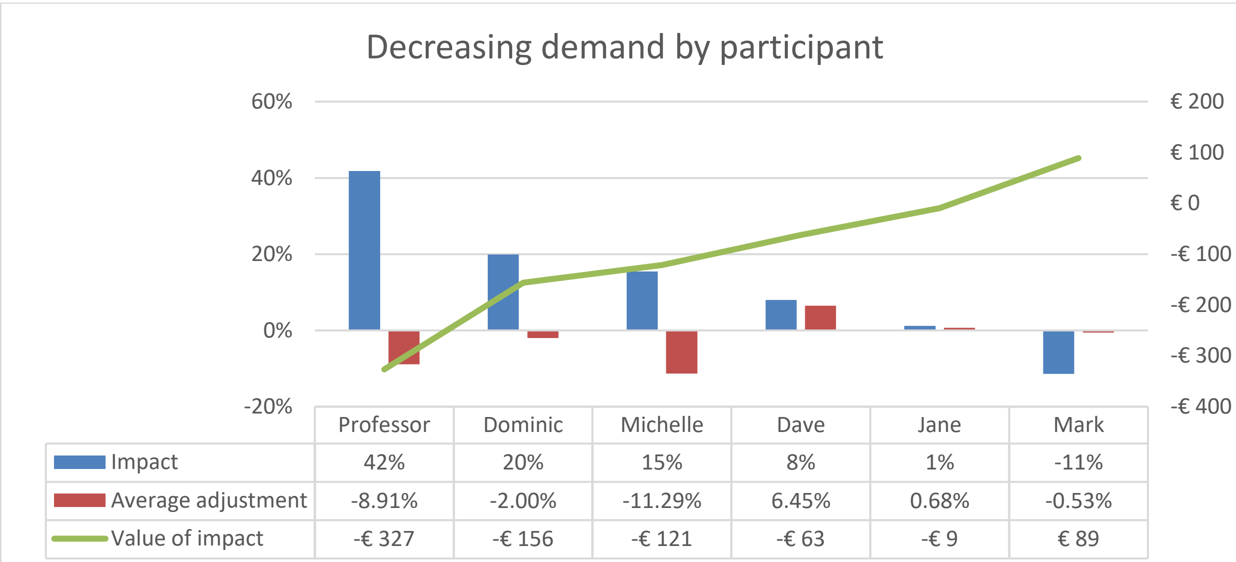


Table 4.76. Participant impact to target stock Decreasing demand.

The aggregated effects of judgment on the Decreasing demand SKU group show an improvement to the statistical forecast of €587.

The ranking of participant performance by SKU group was plotted as below:

SKU group	Increasing demand	Decreasing demand	Longer leadtime	Lower frequency - Highest Value	Very low frequency	New items	Lower frequency - Lowest Value	High frequency - Low Value	High frequency - High Value	Sum
Dominic	1	2	4	2	2	4	2	2	3	22
Professor	6	1	1	1	1	1	1	6	6	24
Michelle	3	3	2	4	5	2	3	1	1	24
Dave	2	4	3	5	6	3	5	5	4	37
Mark	5	6	5	3	4	6	4	3	2	38
Jane	4	5	6	6	3	5	6	4	5	44

Table 4.77. Participant ranking by positive inventory impact versus target

The table 4.77. shows a different ranking to the overall effect versus inventory performance versus target. Michelle and Mark who were the two best performers are 3rd and 5th respectively. This can be explained by the fact that they performed well (1st and 2nd) in the two high value SKU groups particularly the High value – High frequency groups which had a much larger effect on the overall outcome than the other SKU groups (see table 4.67. Pareto graph of SKU group inventory targets). The impact on group inventory target by the Senior academic was by group very good (best in six groups) but as the performance in the high value groups was not as good, overall performance was affected.

4.8. Conclusion

The experiment scrutinised results over 12 months, the impact of judgmental adjustment to the statistical forecast, the impact of a final adjustment, the direction and size of the adjustments and the subsequent impact to forecast accuracy and the effect of all the above on the inventory control of the SKU groups involved. The implications of these results will be compared to the current literature in the next chapter.

5. Discussion

This Chapter will consider the results of the experiment and will contextualize these with reference to the objectives, the literature and the expected contributions as described in the introduction.

5.1. What effect did expertise have on the accuracy of forecasts?

The aggregated result shown in Table 4.1 shows that there was a correlation between expertise and aggregated forecast accuracy. The ranking indicates direct relationship with the three most expert participants producing the least errors in order of expertise.

This was conducted to see if expertise improved the participants judgmental adjustments when applied to a statistical forecast.

At the aggregated level this result seems conclusive, supporting the research posited by many studies (Carbonne et al., 1983; Armstrong, 2006; Franses, 2011; Goodwin and Wright, 1993; Lawrence and O'Connor, 1996) that, incorporating expert knowledge into statistical based forecasts would be beneficial. The fact that the order of participant accuracy reflects the expertise level and that all the company participants performed better than the off-site senior academic shows that causal information and judgment were an improvement to the statistical only forecast. This was the case for both the company statistical forecast and the more sophisticated algorithm used by the senior academic.

The results showed that at aggregate level, judgmental adjustments did improve accuracy' supporting Syntetos et al. (2009) and Matthews and Diamantopolous (1986). However, improvements were not consistent across all the SKU groups.

For the company involved, the forecasting of spare parts integrated a statistical forecast and a possible judgment for all SKU's involved if the forecasters felt compelled. If the participant did agree with the statistical forecast this can still be deemed a judgment as the outcome was considered for judgmental adjustment. The fact that this was the case supports Goodwin (2000) regarding the integration of judgment and statistical methods and the complementary strengths and weaknesses. Within the company there was no constraint on the application of judgment, indeed it was encouraged for all SKU's across all SKU groups as a perceived positive influencer of the forecast accuracy.

It should be noted that the company's statistical forecast was kept relatively simple in order that the people undertaking the forecasting task could understand what the forecast was doing mathematically. This was discussed by Mentzer and Cox (1984) and Mentzer and Kahn (1995) as a reason why little improvement had been seen over the period they were reporting on (termed 'black-box forecasting'). This was specifically to help the company forecasters, by not overcomplicating the statistical forecast they were able to see why it was producing the forecasts it was. By trying to alleviate the issue discussed by Fildes and Hastings (1994) who commented on the lack of skills to understand a statistical forecast being a reason for increased judgmental adjustments, it was hoped that any intervention would be justified.

The spread of knowledge both technical and contextual produced results supporting Carbone et al. (1993) who, when comparing results for accuracy of judgmental adjustments of novices with time series models found that the simpler the model, the better the forecast accuracy.

When comparing the results to Sanders and Ritzman (1995) regarding their investigation into the question of whether technical or contextual knowledge improved accuracy, contextual knowledge produced better accuracy. This is the case as all the company participants understood the statistical forecast to a large degree (some may have understood the implications a little better than others, but this would only give a marginal benefit). The main differentiator was experience (or contextual knowledge). The senior academic who had no contextual knowledge did not perform as well on average as the participants who did.

5.2. How do time series types effect judgmental adjustments?

There is a general acceptance that spare parts are susceptible to a high degree of volatility and can require judgement due to this. The SKU groups themselves focused on issues discussed in the literature review such as low frequency (Syntetos et al., 2009, called for more research), new items (Hughes, 2001, highlighted this as reason managers do not trust the statistical forecast) and trending demand (could be a step change as discussed by Blattberg and Hoch, 1990, or promotions and marketing which Lawrence, 1977 showed is very difficult and time consuming to reflect in time series data). Webb O'Connor (1996) commented that "the major contribution of judgmental approaches lies in the ability to integrate this non- time series information into the forecasts".

The justification for selecting different time series was to analyse whether the participants were better at forecasting some time series more than others and to report if the level of expertise also had an impact on the accuracy of judgmental adjustments to the different time series (discussed by De Baets and Harvey, 2018).

The order of accuracy of each participant, relative to SKU group, was shown in table 4.9. This order was relatively similar for all participants and reflected the frequency and the values of the SKU's involved. The order reflected the inherent difficulty in forecasting time series with a high degree of intermittence as explained by Syntetos et al. (2008), with higher frequency SKU's having a lower MAE/MD ratio than the lower frequency SKU's for all participants. The SKU groups with a more constant demand were placed 1st and 3rd in forecast accuracy, indicating that they were easier to forecast (although not necessarily easier to improve by judgment).

When the average of all participants MAE/MD is compared to the Statistical MAE/MD by SKU group in table 4.10, the result indicated that the statistical forecast had a lower ratio for six of the 9 SKU groups. What this shows is that for some of the SKU groups when considered singly by participant, the statistical forecast was the better forecast and the judgmental adjustment was not an improvement.

The groups that had a MAE/MD lower ratio as an aggregated average using the statistical forecast were:

- High frequency – Low Value
- High frequency – High value
- Longer Lead-time
- Lower Frequency – Lowest Value

- New items
- Very Low frequency

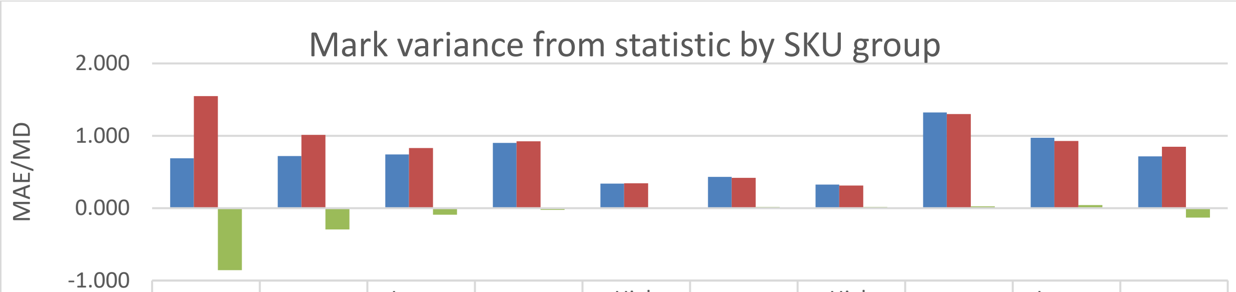
The groups that had a MAE/MD higher ratio as an aggregated average using the statistical forecast were:

- Increasing demand
- Lower frequency – Highest value
- Decreasing demand

It is notable that the two trending data series (increasing and decreasing demand) appear in the groups where the statistical forecast was not as accurate as the participants on average. The company's statistical forecast was average based typically, average based forecasts lag a trend due to their nature. Here a weighted average would possibly be more effective (as described in 2.2.2. of the Literature Review). When Sanders and Ritzman (1992) examined the effect of time series on forecast adjustments they commented that series with more volatility were judgmentally adjusted better than series with a more constant character. This is borne out in the results of the experiment. The trending series (more volatile) were better adjusted whereas the more constant series (less volatile) such as the high frequency groups were less well affected by adjustment.

When this is analysed by participant there are differences when compared to the average. The result by SKU group is reviewed in order of improvement to the statistical forecast. Where the variance is negative the forecast error was reduced. The results will be shown and then commented on as a group due to the general points and synergy that exists between the participants.

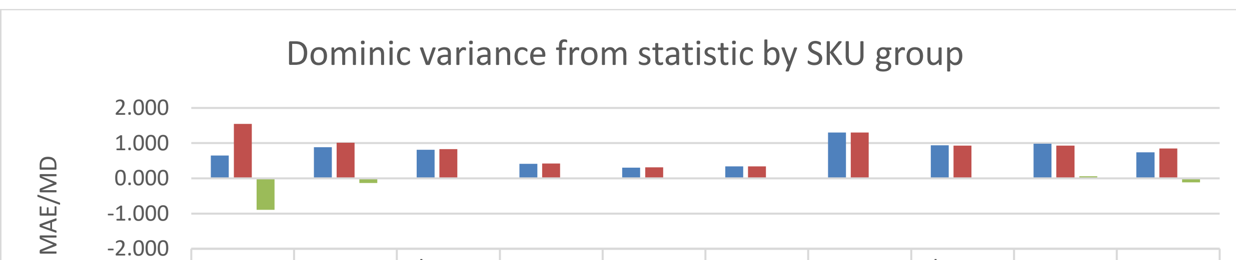
For Mark, the most experienced participant the results by SKU group in comparison to the statistical forecast were as below:



	Increasing demand	Decreasing demand	Lower frequency - Highest Value	New items	High frequency - High Value	Longer leadtime	High frequency - Low Value	Very low frequency	Lower frequency - Lowest Value	Average
Mark	0.690	0.720	0.742	0.904	0.340	0.433	0.327	1.324	0.971	0.717
Statistic	1.547	1.013	0.833	0.926	0.344	0.418	0.311	1.299	0.929	0.847
Variance	-0.857	-0.293	-0.091	-0.022	-0.004	0.015	0.015	0.025	0.042	-0.130

Table 5.1. Mark Variance from statistic by SKU group.

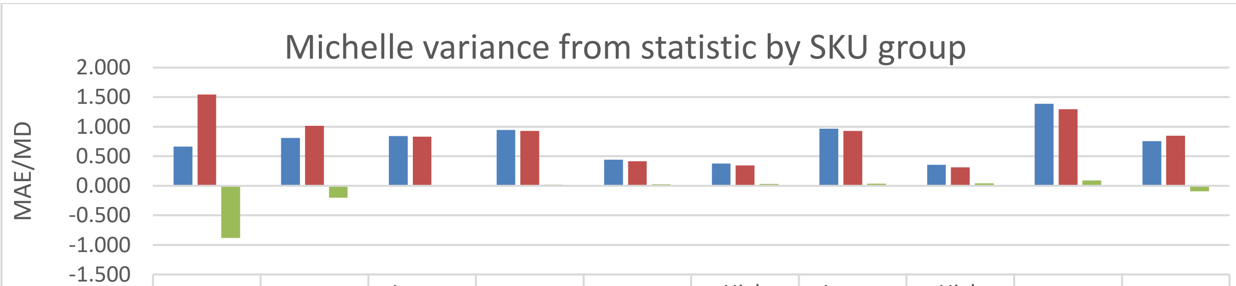
Mark's results (the most experienced of the company practitioners) showed the best average improvement to accuracy. 5 groups were improved when compared to the statistical average based forecast (supporting Fildes et al, 2009). Note when the results were made worse, the variance was low (little impact to overall result).



	Increasing demand	Decreasing demand	Lower frequency - Highest Value	Longer leadtime	High frequency - Low Value	High frequency - High Value	Very low frequency	Lower frequency - Lowest Value	New items	Average
Dominic	0.649	0.880	0.814	0.409	0.304	0.341	1.299	0.935	0.983	0.735
Statistic	1.547	1.013	0.833	0.418	0.311	0.344	1.299	0.929	0.926	0.847
Variance	-0.898	-0.133	-0.019	-0.009	-0.007	-0.003	0.000	0.006	0.057	-0.112

Table 5.2. Dominic Variance from statistic by SKU group.

6 SKU groups were improved by Dominic (the second most experienced company practitioner). The forecasts that were worse than the statistical forecast similarly to Mark produced a small increase error magnitude.



	Increasing demand	Decreasing demand	Lower frequency - Highest Value	New items	Longer leadtime	High frequency - High Value	Lower frequency - Lowest Value	High frequency - Low Value	Very low frequency	Average
Michelle	0.663	0.812	0.840	0.947	0.444	0.376	0.969	0.355	1.391	0.755
Statistic	1.547	1.013	0.833	0.926	0.418	0.344	0.929	0.311	1.299	0.847
Variance	-0.884	-0.201	0.007	0.021	0.026	0.032	0.040	0.043	0.092	-0.092

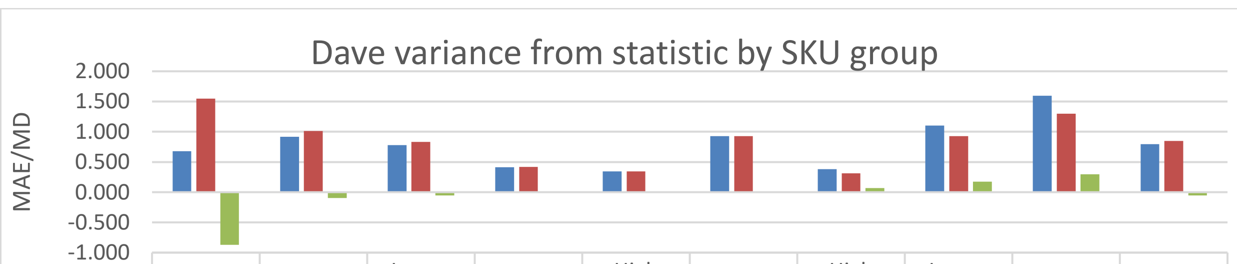
Table 5.3. Michelle Variance from statistic by SKU group.

2 SKU groups were improved by Michelle. This was lower than the other participants, but she was still the third best performer (indicating that the worsening of the statistical forecast was at a lower level for the other participants other than the more experienced Mark and Dominic).

	Increasing demand	Decreasing demand	Lower frequency - Highest Value	Longer leadtime	New items	High frequency - High Value	High frequency - Low Value	Very low frequency	Lower frequency - Lowest Value	Average
Jane	0.647	0.903	0.772	0.415	0.924	0.350	0.329	1.402	1.090	0.759
Statistic	1.547	1.013	0.833	0.418	0.926	0.344	0.311	1.299	0.929	0.847
Variance	-0.900	-0.110	-0.061	-0.003	-0.002	0.006	0.018	0.103	0.161	-0.088

Table 5.4. Jane Variance from statistic by SKU group.

5 SKU groups were improved by Jane (the 5th most experienced company participant). The groups where the forecast was made worse contained some larger variances for very low frequency and lower frequency – Lowest value.

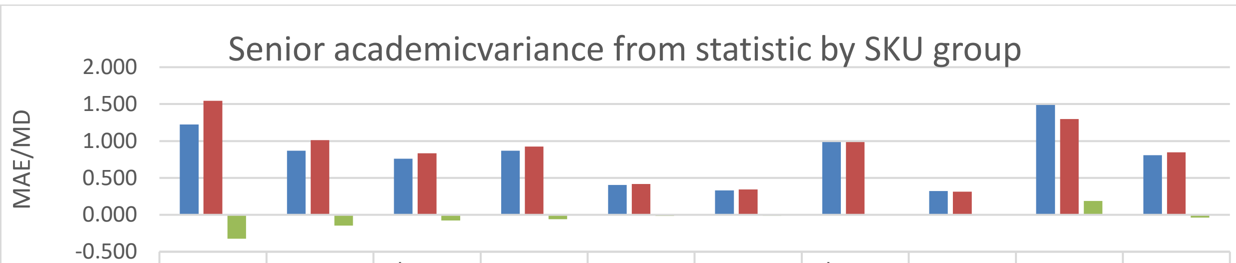


	Increasing demand	Decreasing demand	Lower frequency - Highest Value	Longer leadtime	High frequency - High Value	New items	High frequency - Low Value	Lower frequency - Lowest Value	Very low frequency	Average
Dave	0.678	0.918	0.781	0.415	0.342	0.926	0.381	1.103	1.595	0.793
Statistic	1.547	1.013	0.833	0.418	0.344	0.926	0.311	0.929	1.299	0.847
Variance	-0.869	-0.095	-0.052	-0.003	-0.002	0.000	0.070	0.174	0.296	-0.054

Table 5.5. Dave Variance from statistic by SKU group.

Dave (the 4th most experienced company participant) improved the statistical for forecast for 5 SKU groups. There were large variances for the Very low frequency and Lower frequency – Lowest value where the forecasts were significantly worsened. Again, as with Jane the tendency to over forecast the intermittent groups exists.

The two least experienced company participants performed the worse on the low frequency groups. They did not have the insight of the more experienced company forecasters and decided to over order and guarantee that stock as available.

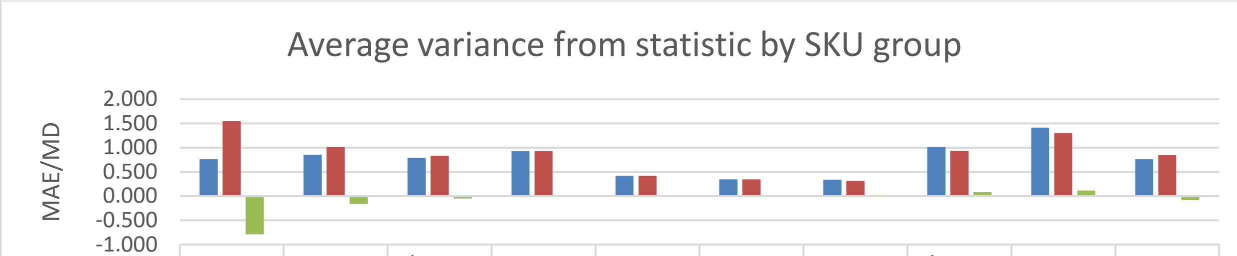


	Increasing demand	Decreasing demand	Lower frequency - Highest Value	New items	Longer leadtime	High frequency - High Value	Lower frequency - Lowest Value	High frequency - Low Value	Very low frequency	Average
Professor	1.225	0.868	0.758	0.868	0.403	0.333	0.984	0.322	1.487	0.807
Statistic	1.547	1.013	0.833	0.926	0.418	0.344	0.984	0.311	1.299	0.847
Variance	-0.322	-0.145	-0.075	-0.058	-0.015	-0.012	0.000	0.011	0.188	-0.039

Table 5.6. Senior academic Variance from statistic by SKU group.

6 SKU groups were improved by the senior academic with one group with no variance. Despite this good result by SKU group the reason that the senior academic was the least good performer of all the participants was that the improvement to the Increasing demand group was less that the others (all the others were < -0.800).

This reflects the improvement provided by the chosen statistical forecast over the company statistical forecast.



	Increasing demand	Decreasing demand	Lower frequency - Highest Value	New items	Longer leadtime	High frequency - High Value	High frequency - Low Value	Lower frequency - Lowest Value	Very low frequency	Average
Average	0.759	0.850	0.785	0.925	0.420	0.347	0.336	1.009	1.416	0.761
Statistical	1.547	1.013	0.833	0.926	0.418	0.344	0.311	0.929	1.299	0.847
Variance	-0.788	-0.163	-0.048	-0.001	0.002	0.003	0.025	0.079	0.118	-0.086

Table 5.7. Average variance from statistic by SKU group.

If we take the average of all the participants across groups there are 4 SKU groups that are positively affected by the judgmental forecasts. Overall the impact of applying judgment to the statistical forecast has been positive by 0.086 of MAE/MD.

The voluntary integration of the judgment (as suggested by Goodwin, 2000) has, in this experiment, proven to produce reduce errors. The similarity between the company participants shows that the judgments were not random as the same groups were usually improved by the highest margin (Increasing demand, Decreasing demand and Lower frequency – Highest value). The research by Sanders and Ritzman (1995) discussed the two types of judgmental forecasts: contextual and technical. Within the group we can say that the senior academic had technical information (regarding arguably the best statistical approach) and the company participants had the contextual information. The result of the experiment supports their reported outcome that contextual expertise was more advantageous than technical expertise. Whilst all participants were better than the original statistical forecast the participants with contextual information (and some technical information regarding the statistical forecast itself) did better.

All the participants judgmental adjustments made the Very Low frequency and the Lower frequency – Lowest value groups worse. This is commented on by Boylan and Syntetos (2010) when commenting on the difficulty of forecasting when dealing with lumpy or intermittent demand. In this case the company's statistical forecast did better than all the participants.

Increasing demand and decreasing demand were improved with judgment. Blattberg and Hoch (1990) discussed the incapacity of statistical methods when reacting to step changes. In this experiment it showed the participants adjustments did add value to the statistical forecast. These groups were also probably beneficiaries from the knowledge the company participants had regarding the SKU groups themselves and the fact that they could apply additional insight into the SKU's whereas the Senior academic who applied a specific statistical forecast may not have this additional input.

5.3. Does extending the horizon influence the judgmental adjustment in a positive or negative way?

The question of whether adjustments improved over longer horizons was posed by Theocharis and Harvey (2016). They proposed that forecasters considered the forecast period furthest away and move backwards to the most recent forecast period. This was not the case in this experiment however it will shed some light on horizons further away such as Lawrence et al. (1986) who stated that the furthest away the forecast was the more the uncertainty increases producing a worse outcome.

From table 4.4.1. none of the participants forecasts improve the further away the horizon. When we look at the SKU groups there is some evidence that forecast gets progressively worse the further away the horizon is for four of the nine SKU groups. The other groups did not improve progressively over the three months with more random fluctuations of accuracy.

For this experiment the participants were asked to forecast month 1 to month 3 and not the opposite way around as Theocharis and Harvey (2016) explored ("end anchoring"). The conclusion that forecasts generally get worse over the horizon was an indication for the experiment; it does not provide any insight into end anchoring.

The length of the experiment horizon (3 months) was driven by the SKU lead-time (sea freight from the Far East plus order period). The study by McCarthy et al. (2006) discusses the level of qualitative intervention over smaller lead-time horizons and the predominance of average based forecasts for the shorter period. The experiment supports their study in that the level of judgmental adjustments were large and the company used an average based lead-time.

Franses and Legestree (2011) discussed the overweighting of the horizon most relevant to the manager when applying judgment and the subsequent accuracy being less than the statistical forecast. This is supported by the experiment as it is seen that the less overweighting (negative adjustments) the more accurate the judgmental adjustment was and vice versa (large positive adjustments affected accuracy in a negative way).

5.4. How does judgmental adjustment size and direction effect accuracy of forecast?

This section looks at the number of adjustments made as a proportion of the total forecasts available. First by judgmental adjustment to the forecast (of which there were 19,440, 90 SKU's x 3 horizons x 9 SKU groups x 6 participants) and then by final adjustment (4860, 90 SKU's x 9 SKU groups x 6 participants). The data will be considered first by direction of adjustment only and then by direction / size and impact to accuracy secondly for both the judgmental adjustments and the final adjustments.

5.4.1. Forecast adjustment direction

From the research of Syntetos et al. (2009) and Fildes et al. (2009) the literature states that negative adjustments are more accurate than positive ones and that larger adjustments are more accurate than smaller ones.

In the experiment there were more positive judgmental adjustments that negative ones (32% to 25%). There was therefore a total of 57% of total adjustments made to the statistical forecast. From table 4.37 the senior academic made more adjustments to the statistical forecast than the rest of the participants. The variance of adjustment level was large with the senior academic making 83% and Michelle making 40%. The senior academic level of adjustment can be explained by the application of a new statistical forecasting method meaning the companies statistical forecast was not considered at all and where there was no adjustment that was just because the two forecasts were the same. The variance between the company participants was 20%. There were two participants that made more negative adjustments than positive – Mark and Michelle. Dominic made 2% more positive adjustments. The two least experienced participants Dave and Jane made significantly more positive adjustments than negative (22% and 18% respectively). This could show that more inexperienced forecasters have a positive bias when forecasting as found by Syntetos et al. (2009). The difference between the less experienced company participants and the more experienced was significant and may have some implications regarding whether judgmental adjustments should be made by forecasters with a certain level of experience to negate this positive bias which did have a negative impact to the forecast accuracy. The results off the least experienced forecasters were in line with Fildes et al. (2006).

There was also a significant variance in the way each SKU groups was adjusted. The range shown on table 4.36. was from 26% to 95%. The items which were adjusted the most were the high value groups. This may be because as there was also an inventory target the participants were scrutinizing these two groups more. The two groups with the least adjustments were the lowest frequency groups (one of which was the lowest value). Boylan and Syntetos (2010) discussed the level of intermittent and lumpy demand that can be seen in spare parts demand arrays. Given the company forecast was an adjusted average it is possible that the participants did not have any reason to change the forecast (that is they thought the average was as good as possible and had no additional input). This may also have been coupled with the low impact to inventory which these groups had.

The list of SKU groups showed a decrease in adjusted numbers from high value to low value and low frequency groups where justification and reason to adjust probably reduced from the company participants perspective. The middle groups in relation to adjustment percentage of trending demand, new items and longer lead-time, all could have induced the participants to add their experience or knowledge to the forecast.

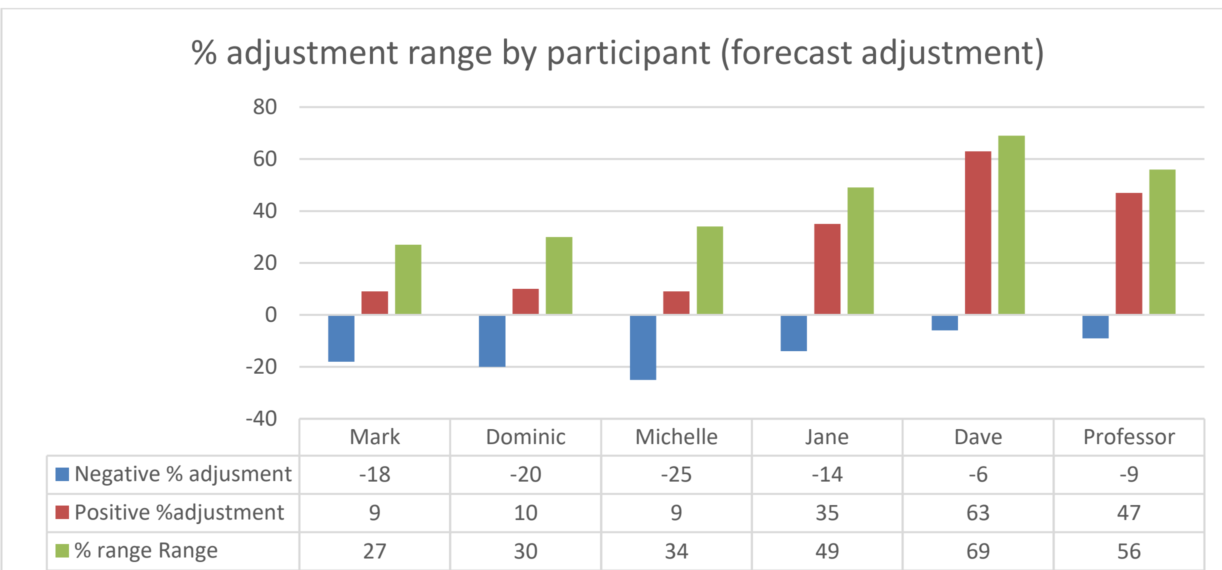
5.4.2. Forecast adjustments made in relation to accuracy

Table 4.52 Aggregate % change to forecast and impact on accuracy reports that on average relatively larger positive adjustments have a negative effect to the forecast. This is seen for the Very low frequency and Lower frequency – lowest value groups who are negatively affected by large positive adjustments. In contrast the relatively larger negative adjustments have a positive effect on the forecast with the groups increasing demand and decreasing demand show an improvement to forecast accuracy. The rest of the groups show small positive or negative effects to the forecast accuracy along with relatively small adjustments sizes and directions. The fact that all the groups were adjusted may show the inclination of forecasters to adjust where there is no real causal reason to so. This was found by Lim and O’Connor (1995) who noted that even when given excellent statistical forecasts the human desire to adjust can make forecasters adjust when there is no reason to so.

When the reports are considered for the participants results, the outcomes are very similar. The larger the positive percentage adjustment the worse the effect on the forecast accuracy and oppositely the larger the negative percentage adjustment the better the effect on the forecast accuracy. The only real outlier to this result was the senior academics results shown in table 4.51. Here two large positive changes to the average % of adjustments showed the first and third largest improvements. This was based on the alternative statistical forecast used.

This result partly supports the findings Diamantopoulos and Mathews (1989) found. They found that the larger the adjustment the more effective the effect on accuracy. The experiment found the larger the negative adjustment the more likely the positive affect on the forecast accuracy it does not support the finding the other way where larger positive adjustments had a negative effect on forecast accuracy.

The range of % adjustments is as shown in table 5.8. below:



	Mark	Dominic	Michelle	Jane	Dave	Professor
Negative % adjusment	-18	-20	-25	-14	-6	-9
Positive %adjustment	9	10	9	35	63	47
% range Range	27	30	34	49	69	56

Table 5.8. % adjustment range by participant (forecast adjustment).

The table shows that the range of % adjustment increases with lack of expertise (for the company participants). The more experienced participants have a larger negative minimum (more likely to reduce the statistical forecast) and they also have a smaller positive maximum (less likely to increase the statistical forecast by a large margin). The importance of understanding where to utilize expertise and where not to was covered by Fildes and Goodwin (2013) and the results shed some insight into where it could be positive for accuracy to constrain adjustments for less experienced participants both totally but also by direction.

The three most experienced company participants all have greater negative maximums than positive ones. The two least experienced participants have far higher positive maximums and lower negative ones. The experiment shows that the positive bias of less experienced forecasters is significant. The subject of 'big losses' pertaining to judgmental adjustments was discussed by Petropoulis, Fildes, Goodwin (2015) and this sheds light into possibly why less experienced forecasters err on the side of positive adjustments.

The level of experience in this experiment also has an influence on the average size of the judgmental adjustment (range) with more experienced participants having a much smaller range. The requirement for an FSS to feedback this sort of information was discussed by Moritz, Siemsen, and Kremer (2014) as a way if improvement via cognitive reflection.

5.4.3. Final adjustments made in relation to the number of forecasts

Evidence that orders were adjusted post the demand forecast was highlighted by Syntetos et al, (2011) as an area that has been neglected in academic literature. This is where the original statistical forecast is adjusted then it is adjusted a second time in relation to the placing of for example replenishment orders (as was the case in the experiment).

In the experiment there were more positive final adjustments the forecast (changes made to the judgmental adjustment) than negative ones (9% to 5%). There was a total of 14% final adjustments made to the forecasts. From table 4.44. Dominic made more final adjustments than anyone. The variance in final adjustments made was significant with Dominic making 36% adjustments compared to Michelle, Jane and Dave who made 8%, 7% and 1% respectively. The senior academic who had no reason to make any further adjustments to the adjusted forecast (or the alternative statistical forecast) made no final adjustments.

The disparity between Dominic and Mark (who both made over 30% final adjustments) and the other company participants could be down any number of factors: knowledge regarding the specific SKU, conviction to change greater with expertise or rounding of the final forecast. The subsequent impact to the forecast accuracy may allow further insight into this disparity.

The number of final adjustments is shown by SKU groups in table 4.45. It shows that the highest number of final adjustments was made to the high frequency groups (29% and 27% respectively). The rest of the SKU groups were below 9% and were not in the same order to the forecast adjustments.

5.4.4 Final adjustments made in relation to accuracy

Syntetos et al, (2016) noted that inventory related decisions are judgmentally adjusted more frequently than forecasts. This implies that where inventory targets are combined with demand forecasts the judgmental adjustment frequency increases. We can infer that one of the reasons for a final adjustment was inventory related (as all participants were working within a stock index parameter target), however it was not the sole reason (packing quantities, round up etc. are also drivers).

From table 4.53. only Dominic improved his forecasts by making a final adjustment in this experiment. The impact of the final adjustment was linked to the expertise of the participant. Only Dominic improved the forecast. Mark worsened it slightly, at the other end of the scale Dave and Jane (the participants with the least expertise) worsened the judgmentally adjusted forecast by the largest error. When the final judgment was broken down into SKU groups there was a different picture.

For the Lower frequency – high value group and the Decreasing demand group the final adjustment was positive for all participants the average adjustments being significantly negative (21% and 19% respectively). For all the other groups the final adjustments made a negative impact to forecast accuracy. The two high frequency groups were marginally worsened with the Longer Lead-time and Very Low Frequency being the most negatively affected.

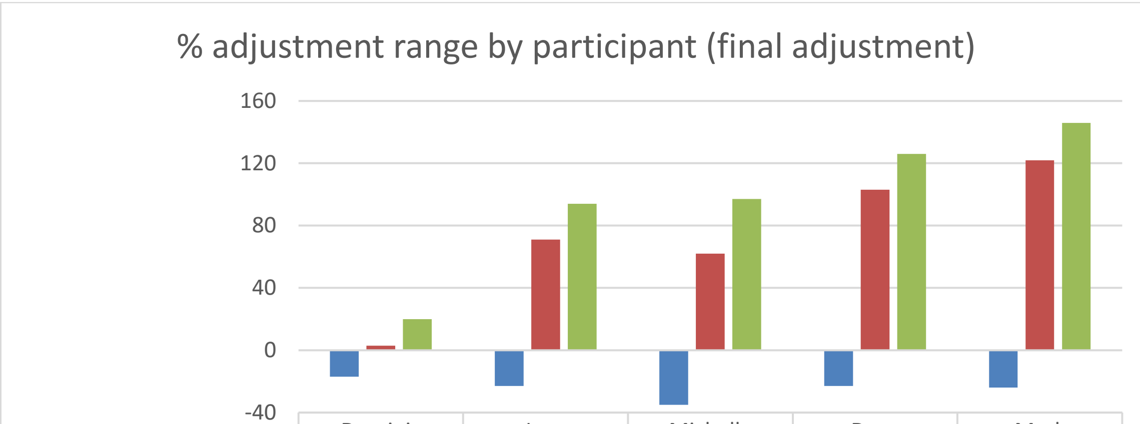
The aggregated poor impact on the forecast supports the literature that found judgmental adjustments to have a negative effect on statistical forecast (Armstrong 1986; Hogarth and Makridakis 1981; Makridakis 1988) however, it must be pointed out that this was the final adjustment and not the forecast adjustment.

Operationally within the company there could be a reason for the final adjustment of the judgmental adjustment having a negative effect on forecast accuracy. The company to help suppliers, and to reduce goods-in costs, often rounded up or down the judgmental adjustment. For example, if a forecast was 777 pcs then the order could be rounded to 800 pcs to receive the spare parts in boxes of 100pcs. This contextual knowledge does not necessarily impact the judgmental adjustment in any direction, but it is not based on forecast accuracy as it is done to improve operational efficiency in the supply chain.

When investigating the percentage change of the final forecast and the impact on accuracy (table 4.58.) the results show similar correlation as the forecast adjustments. Large positive final adjustments had a negative effect to forecast accuracy and large negative final adjustments has a positive effect. The number of groups that were positively affected in the aggregated table were less that the adjusted forecasts with only 2 being positive (as oppose to 4). The number of average negative percentage adjustments is also smaller at just 3 SKU groups (as oppose to 4). This could indicate a more positive bias when addressing the adjusted forecasts than the statistical forecasts.

The SKU groups which were positively affected were Decreasing demand, and Lower frequency – highest value. Both these groups were in the four that were positive affected in the forecast adjustment. The Increasing demand for the forecast adjustment was the most positively affected in the forecast adjustments but it was negatively affected by the final forecast adjustment.

The range of percentage final adjustments is as shown in table 5.8 below:



	Dominic	Jane	Michelle	Dave	Mark
Negative % adjusment	-17	-23	-35	-23	-24
Positive %adjustment	3	71	62	103	122
% range	20	94	97	126	146

Table 5.9. % adjustment range by participant (final adjustment).

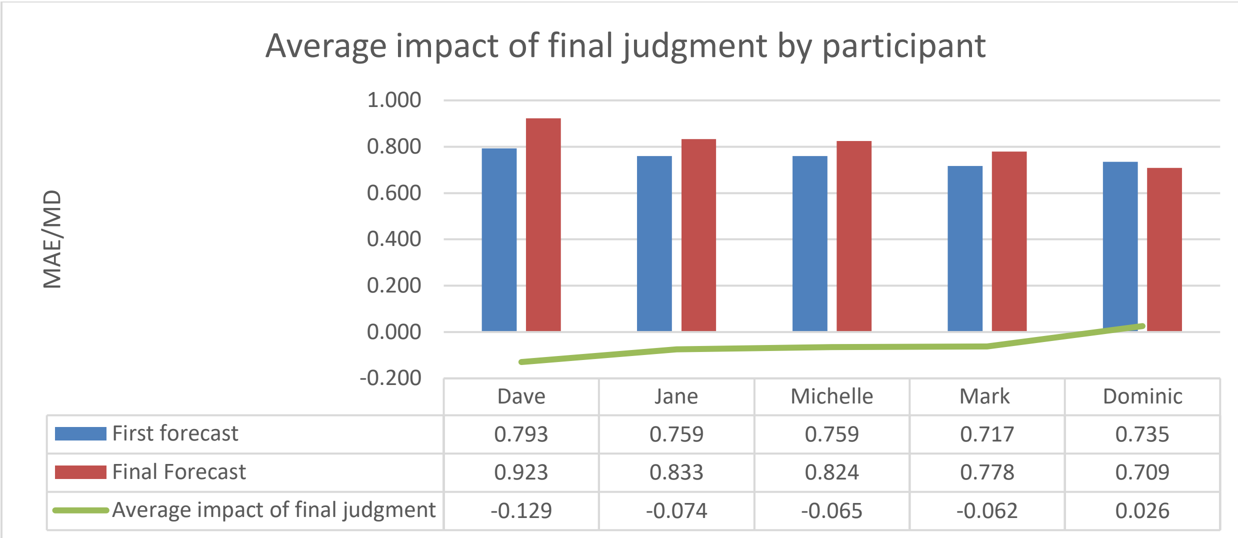
The final adjustments range show a different result from the forecast adjustment range. The ranges themselves are much larger (the largest forecast adjustment range was 69), whereas for the final adjustments it is 146. There is only one participant who had more negative adjustments than positive ones (Dominic) and the size of the positive adjustments were much larger than the negative adjustments.

The reasons for the final adjustments as already explained were different to the forecast adjustment (rounding of quantities based on contextual information) so the primary impact was not necessarily accuracy. The level of positive adjustments points to a bias towards increasing the forecast (whether judgmentally adjusted or not) which may be influenced by other factors such as the cost of the part and the knowledge that overstocking of a low value part will not impact the inventory value significantly and could also save costs in the future by meaning there would be less purchase orders placed for the part (less freight costs and inbound warehouse costs). The driver for adjustment is not always accuracy of forecasts particularly for the final adjustment.

Another possible reason for the increase is the unknown. New items can suddenly increase in demand and this may have been a tactic to negate the impact of such sudden increases. Hughes (2001) discussed this issue where no previous data exists and highlighted the support needed for forecasters when judgement was the only method available pointing out that cross functional support from sales managers would improve the forecasters knowledge base.

When we consider the SKU groups which show the largest average percentage adjustment in table 4.58. it is shown that the two largest positive percentage increases are the New items and Lower frequency – lowest value SKU groups. This gives some credence to the argument above. Both groups having no previous data or data that is low in frequency forcing the forecaster to use judgment to a higher degree (rather than a combination of statistical forecast and judgment).

When we consider the impact of the final adjustment on the forecast adjustment and whether that improved accuracy, the results are shown in the table below:



	Dave	Jane	Michelle	Mark	Dominic
First forecast	0.793	0.759	0.759	0.717	0.735
Final Forecast	0.923	0.833	0.824	0.778	0.709
Average impact of final judgment	-0.129	-0.074	-0.065	-0.062	0.026

Table 5.10. Average impact of final judgment by participant.

For the five participant that made final judgments there was only one that improved the accuracy of the forecast (Dominic the 2nd most experienced company participant).

Jane, Michelle and Mark all made the average forecast accuracy worse by approximately 0.07 MAE/MD however Dave made his forecasts worse by 0.129 MAE/MD. This was primarily due to the negative effects his final adjustments to the SKU groups Longer lead-time and Increasing demand.

The company allowed a final judgment based on supply chain efficiency. It was not known whether this improved the forecast accuracy or not. Considering the results from the experiment it would be worthwhile reflecting on the impact and possible creating some rules regarding the size and direction of the final forecast. It would be useful to understand empirically if this final judgment was positive or not for the company in line with suggestions by Spithourakis et al., (2015) who propose learning effects via a forecasting and foresight support.

The level of changes made for the final adjustment for 3 of the company participants was at a negligible level compared to the judgmental adjustment. Given that the impact was generally negative it would be worthwhile understanding what was the reason for their intervention. For Mark and Dominic (who made 36% and 30% final adjustments) it may be of use to reflect on the negative impacted groups with a view to limiting the allowance of a final adjustment to groups where there was a positive impact.

Overall the impact of adjustment can be shown as below by using the level of contextual information available per SKU group and plotting that against the impact of the judgment whether it be positive or negative.

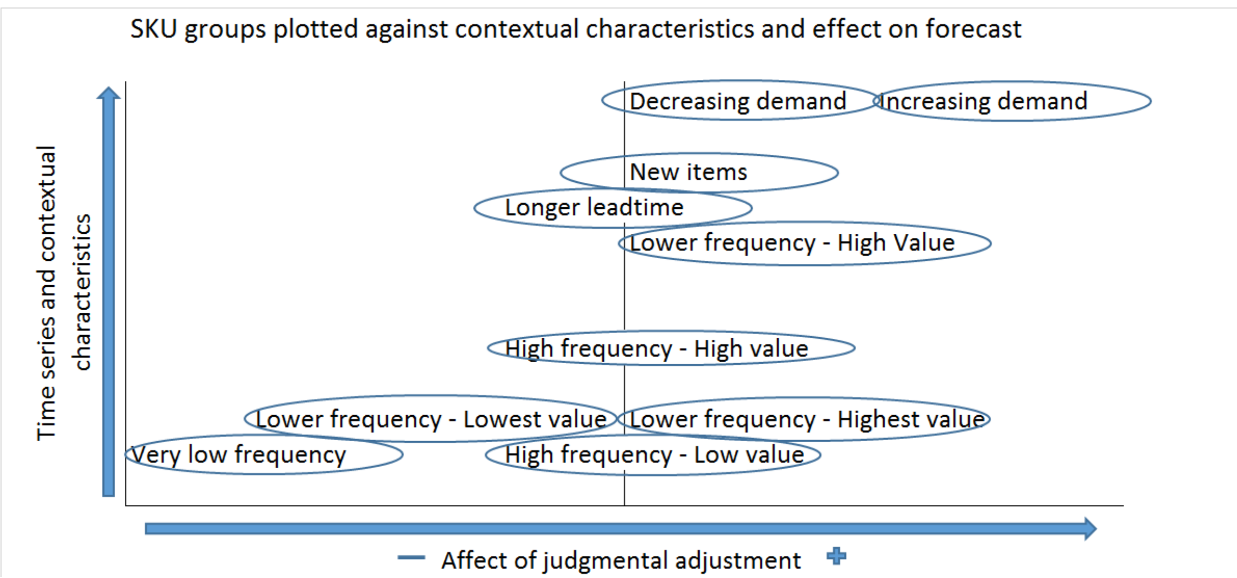


Table 5.11. SKU groups plotted against contextual characteristics and effect on forecast.

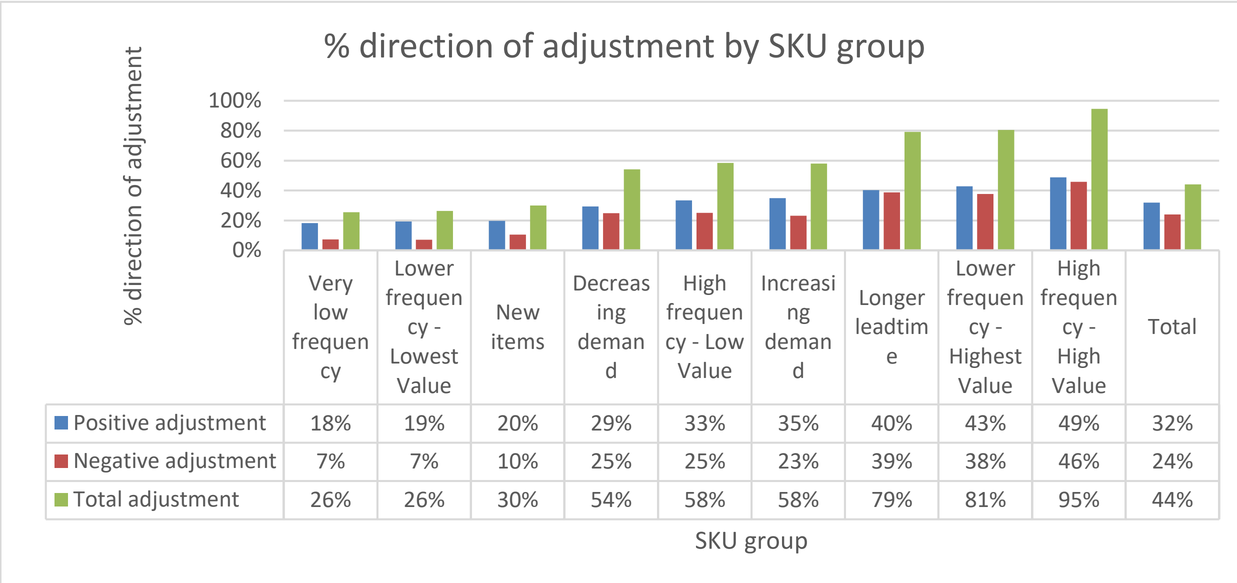
The graph 5.11. provides a visual picture of where the most positive results and the most negative results were seen from the SKU groups. SKU groups with characteristics such as trending or containing contextual time series when a form of average is used for the statistical forecast are where the focus of judgmental adjustment should be. This is not reflected when we look at the % of SKU's adjusted by group (as shown in table 5.12. repeated below). This supports partly Franses (2013) conclusion that experts should choose where they insert judgement to a statistical forecast (in this experiment the positive instances of judgmental adjustment were greater than described by Franses).

The participants focused the most of their resource on the High Frequency – High value group when the results of the adjustments brought a very small negative impact to the statistical forecast on average (table 4.52.). 95% of all SKU's on average were adjusted. Whereas for the group “increasing demand” which was the most improved group from judgmental adjustment only 58% of SKU's were adjusted.

The point here is not that 58% may be too low (although it may be) but 95% was too high. The groups New Items and Longer lead-time were negligibly affected by judgmental adjustment (with 30% and 79% of SKU's adjusted respectively) meaning that the contextual information here was on average not meaningful.

The results are different by participant as explained earlier and this could mean more groups were positively affected by judgment by participant.

It is clear from the experiment that certain groups were not improved such as Very Low frequency. Although this group showed the lowest number of adjustments at 26% it was to a negative impact.



	Very low frequency	Lower frequency - Lowest Value	New items	Decreasing demand	High frequency - Low Value	Increasing demand	Longer leadtime	Lower frequency - Highest Value	High frequency - High Value	Total
Positive adjustment	18%	19%	20%	29%	33%	35%	40%	43%	49%	32%
Negative adjustment	7%	7%	10%	25%	25%	23%	39%	38%	46%	24%
Total adjustment	26%	26%	30%	54%	58%	58%	79%	81%	95%	44%

Table 5.12. % direction of adjustment by SKU group.

From the results of the experiment we can say that there are SKU groups that lead themselves to positive judgmental adjustments more than others. It seems that the instinct to adjust a statistical forecast is a factor (as stated by Seifert et al., 2015). Specifically, for perceived important groups where forecasters are perhaps too eager to alter the forecast without good reason.

It would therefore be useful for time series characteristics to be known before forecasters are asked (allowed) to apply judgmental adjustments as the effect of their intervention could be somewhat guided by this (this was discussed by Adya and Lusk, 2016).

It is worth noting that this company used a simple weighted average when extrapolating the results and that there was some difference between the participants based on experience.

5.5. The impact of adjustment on inventory

Rekik, Glock and Syntetos (2017) confirmed a previous study by Syntetos et al, (2016) that showed judgmental adjustments to replenishment orders were beneficial to stock control.

The combined impact of both the forecast adjustment and the final adjustment was reported by its impact to inventory to see whether there was any correlation between participant, SKU group and value of effect to the inventory. The judgmental adjustment and the final adjustment were not separated as the final forecast was a single adjustment unlike the judgmental adjustment which was spread over the 3-month horizon.

Table 4.59. shows that there were three participants that were closer to the inventory target than the statistical forecast. These were Michelle, Mark and Dominic. These participants were the most experienced of the company participants. The other three participants: Dave, Jane and the senior academic were further away from the inventory target than the statistical forecast. All the participants and the statistical forecast were over the target inventory. This result contrasts the result regarding forecast accuracy where all participants performed better than the statistical forecast. It is not surprising that all the participants were not able to meet the target inventory as spare parts are inherently difficult to forecast. As pointed out by Bacchetti and Saccani (2012) the pressure not to have possible downtime and stock outs were always an issue that could lead to the risk of obsolescence.

We can see from the overall results that Michelle was the best performer and the senior academic was the worse. This, however, masks the fact that the senior academic was in fact the best performer for 6 of the 9 groups. The statistical forecast used by was actually very good from an inventory perspective. This supports the suggestion that different methods should be used per classification of time series (Syntetos et al., 2005). For smoother items exponential smoothing and for intermittent items Croston's method or a derivative was the suggestion. The three groups where the senior academic did not perform well were the Increasing demand and high frequency groups. The reason for this could be two-fold – the company participants could apply causal information for the groups and the fact that one of the groups was high value meant that they would also scrutinize this group closer with a view to inventory impact from month to month when forecasting.

From table 4.66. we can see that the biggest improvement to inventory were made in the trending SKU groups and the worst to the High frequency groups (apart from Mark and Michelle). This result would infer that the best improvements to inventory via judgmental adjustments are received when the time series are not constant but also not intermittent either. The participants could use their knowledge of the parts to improve the inventory outcome.

The effect of intermittent and lumpy demand as described by Boylan and Syntetos (2010) as requiring special attention (such as specific forecasting techniques) and the contrary requirement to have a simple company statistical forecast which all the forecasters could easily understand mean that the resulting forecast may be sub-optimal. One pressure is to utilize algorithms that the forecaster may not understand but would produce better forecasts (specifically for lower frequency time series) and the other conflicting pressure is to use simple methods such as averages to enable the forecasters (often limited in their understanding of mathematics) to come to a best solution.

There was an acknowledgment that for low value items the forecasters should err on the side of caution regarding unavailability by the company. This inevitably resulted in a positive bias for the groups which were low frequency and (or) low value. From an inventory perspective small increases in forecasts for these groups were outweighed significantly by any over forecast for the High value – high frequency group. Baccheti and Saccani (2012) discuss the practical applicability of the different methods for forecasting spare parts and the user skill and support systems that are available. This was certainly the case for the company involved in the experiment. Whilst service level was important the subsequent impact to inventory was not investigated and reflected upon. This type of attitude in sales led manufacturer often means that for spare parts inventory is a secondary thought. When the inventory levels of product and accessories are compared to spare parts it can often mean inventory levels of the spare parts are hidden in the overall inventory picture. This view can be somewhat short sighted when we consider findings from Glueck et al., (2011) who reported 50% of total profits can come from after sales service to many companies.

In summary the experiment found:

1) Expertise did have an impact to judgmental adjusted accuracy. In this case the results correlated nearly exactly to the level of expertise of the participants. The conclusion is that forecasters who are to apply judgment

to statistical forecasts should have a higher level of expertise which would increase the likelihood of improvements to the statistical forecast. This is a clear indication that experts do improve statistical forecasts.

2) Time series types do influence judgmental adjustments outcomes. Constant and low frequency time series produced negative judgmental adjustments. Trending and groups where contextual information was important produced positive judgmental adjustments. The experiment showed that judgmental adjustment should not be applied on certain types of time series where the statistical forecast can provide a better adequate forecast. A matrix is shown to reflect the real-life empirical results to provide insight into where judgmental intervention is most effective at improving a statistical forecast.

3) Final adjustments generally did not improve the forecast. The reason for the final adjustment is important. Is it to improve the forecast (which in the experiment was only the case for 2/9 SKU groups) or to make supply chain efficiencies? The issue is whether adjustments made to replenishment final orders should be judged by purely forecast accuracy, there are other drivers that are included, and they can themselves provide positive impacts to the supply chain.

4) Extending the horizon did not improve the forecast. Note this was not end anchored but based on a longitudinal basis of 3 months. Overwhelmingly, the forecast was less accurate the further away the horizon moved.

5) There were more positive adjustments than negative ones. The level of adjustment increased with expertise. The level of negative adjustment also increased with expertise. For the most experienced experts some of the adjustments were not adding value to the statistical forecast, specifically for certain time series types.

6) Large negative adjustments added the most value to accuracy. Large positive adjustments effected accuracy the most negatively. This was the case for judgmental adjustments to the statistical forecast and final adjustments to the adjusted forecast. This could reflect a positive bias that needs to be monitored for future judgmental strategy constraints.

7) Judgmental adjustments and its effect on inventory show that expertise does reduce the inventory error and that groups where causal / contextual information exists are the most positively affected by intervention. This result was an extension to the findings in accuracy by time series to an extent. However, high value groups were positively affected by adjustments implying forecasters scrutinized these groups more with inventory implications also in their mind.

8) The table 5.10. shows that the level of contextual information is correlated to whether the judgmental adjustment has a positive effect on accuracy or not. It is also used to highlight the level of judgment that is focused on SKU groups that were not positively affected by judgmental adjustment. The

logical conclusion is that the time series groups that were not positively affected should not be judgmentally adjusted as a rule.

5.6. Conclusion

The impact of judgment to a statistical forecast and its accuracy is linked to the direction and size of the adjustment (including both the judgmental forecast and to the final judgmental adjustment) it is also affected by the expertise of the forecaster.

The experiment provides hitherto unique empirical evidence and insight, of which, there is very little academic real-world examples (no longitudinal real-life experiments) into judgmental adjustments (both to the statistical forecast and the purchase order quantity) and how they affect the accuracy from both a demand forecast and inventory perspective.

The implications of the variance of the forecast accuracy results can be used to provide insight into which direction, size, type of SKU and time series can be positively affected by judgmental adjustments. Table 5.11. shows an overall matrix from the participants of where the most accurate forecast adjustments were made. This could be re-produced for the individual components listed above.

This chapter has provided an analysis of the results of the experiment and compare against the aim and objectives stated in the introduction.

6 Limitations

In this chapter the possible improvements and constraints in the PHD will be discussed and reflect on.

The prime constraint regarding the experiment was deciding the scope of the experiment. The company participants were providing their time whist at work doing the forecasting task in a real-world environment. The level of extra work which was required to provide insight via the experiment was limited and had to be reasonable for the forecasts provided to reflect the quality of their day to day tasks.

Because of this firstly, the number of SKU's required to be forecast was limited (although the SKU's involved were able to reflect the required times series characteristics as required). Secondly, the ability to gain further insight from the experiment from a questionnaire which may have shed further insights into FSS and bias for example was not able to be conducted due to these constraints.

The issue of end anchoring was not able to be investigated. It was confirmed that for a horizon of three months the accuracy of the judgmental forecasts reduced the further from the current period. This was to be expected but had not been reported on before empirically from a real-life experiment but notwithstanding did not address the end anchoring question.

If the experiment was run with more time available, then it would have been possible to focus on bias and FSS. Interviews and a questionnaire to establish more insight into how the judgments were made regarding external information and inventory targets.

7. Further research

Whilst this experiment can be contrasted with other laboratory type studies it would be most relevant to compare against experts in other organisations in a similar real word environment for the spare parts area and for products.

7.1 Origin of statistical forecasts

The methodology for the experiment was shaped to enable data to be extracted from the ERP system and to then forecast within an excel spreadsheet before uploading the forecast back into the system. It would be interesting to compare with forecasts done entirely within an ERP system where the forecaster had the chance to judgmentally adjust within the system rather than outside to see if this had any impact to the levels of activity. It could be that the manipulation of the data outside the ERP system induces a different level of judgmental adjustments due to a reduction of the "black box" type scenario. That is, the forecaster may feel they have more "ownership" outside a system which could induce a different level of judgmental adjustment.

7.2 Forecast techniques

The forecast used in the company was a single average-based forecast. There may be cases where more complicated forecasts are used in other companies. The level of judgment applied could be affected by this. If they were to have been presented with more options, it would be interesting to see if the judgment required to choose a statistical forecast reduced any inclination to judgmentally adjust after this and whether they chose the same statistical forecast to use. This was discussed as an area to research by Chase (2013) and Lawrence et al., (2006). This single forecast also meant there were limited “black box” type responses and possibly more judgment adjustments because of this (as discussed by Mentzer and Khan, 1995). That is because the participants knew what the forecast was doing, they were more willing to adjust with confidence.

In the study the participants were given the statistical forecast as a fait accompli with the opportunity to adjust that forecast. Further research offering a more diverse spread of statistical forecast would be interesting to investigate. For example, would more experienced participants have chosen techniques that were more complicated and less experienced have chosen simpler options? The importance of technical knowledge was not a specific area of interest in the PHD (contextual knowledge was more the focus in relation to expertise). This would expand upon work done by Carbone et al., (1993) and Sanders and Ritzman (1995).

A next step to check the explanatory power of the causes (forecaster attributes such as experience and academic qualifications) for the judgmental adjustments would be to look at explanatory models. By utilizing regression this could be done using dummy variables (an artificial variable created to represent an attribute with two of more distinct levels). It would be a natural next step of this research to do so.

7.3 Combination of forecasts

The company believed in the positive impact of judgment post the statistical forecast (supported by Sanders and Ritzman, 1995). There was historical top level (overall hit ratio) evidence to confirm this belief, but no investigation had been undertaken to prove the fact at a SKU level. Other companies may have investigated the role of judgment and made some constraints to where judgmental adjustment was allowed. The company involved had taken some steps to scrutinize the reasons for judgmental adjustments by requiring forecasters to give a reason for any adjustment made (for example, increasing demand trending above the average based statistical forecast). Anecdotally, this reduced the number of forecasts made (there had been a record of adjustments prior to this additional reason requirement showing a higher level). This is supported by Fildes et al., (2009), where the compulsion to judgmentally adjust a statistical forecast with little evidence a negative result was discussed.

This may not always be the case for other companies where the statistical forecast may be enforced or where only a chosen number (possibly based on expertise) are able to influence judgment post the statistical forecast. The level of judgmental forecasts was highlighted as too high by Sanders and Manrodt (1994) during their investigation who supported some sort of limitation.

It would be interesting to follow compare overall accuracy and inventory results per time period where some experts were constrained, and some were not as to the level of adjustments allowed to the statistical forecast.

7.4 Spare parts

The experiment was set in a Pan European Spare Parts environment. This meant that the cost of most of the SKU's was relatively low when compared to some Product forecasts for example. It is difficult to state if this would have an impact on the level of judgmental adjustments when compared to more expensive SKU's. It would be interesting to compare a similar study where SKU's were more expensive and possibly products.

The fact that spare parts can show a different demand pattern to other SKU's may limit the comparability across SKU types. For example, product can often have a shorter lifecycle (not extended by warranty such as spare parts) and be less likely to become obsolete due to options such as sell through activity (price promotion and marketing). This point was made by Chase (2013) when discussing different forecast-ability with reference to lifecycle, and priority of SKU involved. It may be that some of the time series involved would benefit from different types of statistical forecasts that were not available to the participants in the experiment.

The company makes predominantly business machines namely printers and multi-functional machines (MFC's) which can also scan. The lifecycle for this industry is typically 5-7 years. There are industries where spare parts can be required for more than this period (motor / airline industry for example) where this could influence the judgment of forecasters. It could be that obsolescence would be less of a factor if the spare part lifecycle was much longer as the sell through period would be longer.

7.5 Expertise/ Causal influences / Opinion

Although the level of expertise was reported it was not made clear what type of extra information this allowed the participants to use and whether this was shared amongst the participants.

The level of expertise was assumed to mean some participants had access to extra knowledge (tacit knowledge). The nature of this knowledge was not recorded. It was known that the more expert participants of the company participants were more aware of knowledge available cross functionally within the company, for example, service information such as warranty lengths in Europe but this was not recorded.

Carbonne et al., (1983) reported that the simpler the statistical technique, the better the forecasts were, specifically for novices. In the experiment this was the case and should have resulted in all company participants understanding the statistical forecast fully. This should have removed the impact of technical knowledge and left contextual knowledge as the main factor for the participants input as described by Sanders and Ritzman (1995).

It would be interesting to understand the importance of different knowledge basis. For example, was academic background more of a factor than knowledge of the spare parts themselves or if cross functional knowledge was more important than understanding the statistical forecast. If a list of positive factors could be produced a set of skills required for accuracy could be ranked and focused on when companies required forecasting expertise.

7.6 The use of FSS

Following on from 7.5 which is more individually focused it would be insightful to look at the use of a forecasting support system, to test if participants using an FSS performed better than those forced to forecast on their own without any external support. It would also be interesting to understand the process of any support system and to examine the usefulness, or weight, each participant gave to the different inputs they solicited. Expanding on research by Fildes et al., (2006) it would be useful to understand what areas experts were most influenced by when applying judgmental adjustments.

7.7 Forecast Horizon

The forecast horizon for this experiment was 3 months. There are industries where the horizon is much longer for example, mining and aeronautical industries. This may have implications for the level of adjustments made and allow for end anchoring to be properly investigated as suggested by Chase (2013) and Theocharis and Harvey (2016). It would be useful to relate findings from this experiment with similar insights from an experiment with SKU's with much longer time series horizons and where end anchoring was used. Would the same directions and sizes be found for the same time series characteristics?

7.8 Inventory

The performance against inventory was reported but without any deeper investigation into factors that could have influenced the decision to judgmentally adjust such as model importance (was the model an important model that required 100% spare part availability) or part criticality (was this part particularly critical for repairs in the field). There could have been reasons to stock a SKU other than those relating the demand levels. Indeed, some of these motivations may have influenced the forecast accuracy but were not investigated other than a known level of education, position in the company, years of employment etc...

Teunter and Sani (2009) suggested that inventory levels or service levels were better accuracy measurements this was not investigated other than that each SKU did have a stock index target. This index was used to produce the statistical forecast which the participants made judgmental adjustments to. It was noted that where SKU groups were low value and or intermittent some participants made decisions to forecast higher than the statistical forecast to place stock ‘just in case’ one shot demands did occur.

A more focused study on the reasons related to forecast adjustments in relation to inventory performance specifically may highlight some tendencies regarding adjustments size and direction.

7.9 General discussion points

The fact that the participants were involved in an experiment may have influenced their behaviour when producing the forecast. Knowing that the results would be investigated (via the PHD) more than usual may have affected their judgment. To mitigate any “experimental impacts” the experiment used exactly the methods that the company utilized and as explained one of the forecasts for all SKU’s was the actual forecast applied in real world. The participants knew which SKU forecasts were owned by themselves and so would be the actual forecast used when demand planning so it is unknown if when forecasting a SKU which was not a real-world responsibility if it would produce a less focused forecast.

8. Concluding remarks

The thesis aims to assess the impact of judgment on forecasts for spares parts and its' impact on inventory management by running a 12-month long experiment utilising real word experts plus a senior academic.

The experiment was completed, and all participants finished all forecasts.

The constraints to the breadth of the experiment meant that any further insight in the form of questionnaires and interviews was not viable due to time available which would have been interesting regarding some aspects of the thesis (bias and FSS for example). Using real world experts doing their job simultaneously is unique (positive for insight versus rest of the literature) but it does limit the opportunity to cover all areas due to reasonable requests for time. This is a trade-off between laboratory experiments and real-world ones. However, the insight given using real world forecasters and not undergraduates or experts using non real data is hugely important when reflecting on the previous literature and its real-world validity. Many previous papers which were not tested in a real-world experiment have been validated by this research. The clear evidence provided here over a long time period confirms many reported insights that were hitherto not borne out by evidence via experimental research using live data and experts on the job.

The academic objectives, to produce a longitudinal experiment whilst noting inventory implications of judgmentally adjusted forecasts, were successfully achieved.

From a practical, the impact of judgment was recorded, and the data was able to produce fresh insights into size and directions of adjustments.

The type of demand series and the expertise of the forecaster was also reported on giving some important insights regarding possible focus for managers in the future when applying judgmental adjustments to a statistical forecast. The matrix produced is a visual help to forecasters certainly in the spare parts environment but also possibly to many other areas where time series are similar.

Reflecting on the objectives, during the experiment some sort of feedback / FSS system could have been implemented for the participants. This may have given the opportunity to measure improvement as the experiment progressed. Participants did not have the motivation nor information, which may have caused some change to their adjustments which could have then been measured longitudinally.

The results regarding size and direction of adjustments and their merits did broadly support most of the literature existing.

The recommendations for which time series are best suited to judgmental adjustments can be used to stimulate further discussion and application from both an academic and practical perspective. Specifically, it gives an insight as to what time series are best judgmentally adjusted and rates the importance of expertise alongside those time series. This is useful for both academics in order to focus on both where forecast accuracy was improved and how to improve further and where accuracy was not improved in order to understand why not. Practically these results can be used to focus resource where most added value can be found and to reflect on areas where the accuracy was made worse by either stopping adjustments or at least curtailing them (possibly based on size and direction of the intervention).

The thesis provided a unique opportunity to work with experts who when setting the parameters of the experiment were able to interact and be interested in the possible implications it could have on their forecasting techniques. On completion of the thesis the next developmental phase would be to solicit their observations of the reported results and recommendations.

The quote by Gertrude Stein from Blattberg and Hoch (1990) that “Everybody’s got so much information all day long that they lose their common sense” is something that reverberates strongly with the researcher. It is imperative to empirically test which pieces of information make a positive impact to judgmental adjustment and focus any activity on those areas to which accuracy is most improved.

References:

Alvarado-Valencia, J. A., and Barrero, L. H., Onkal, D., and Dennerlein, J. T. (2017). Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. *International Journal of Forecasting.* **33(1),** 298-313.

Amabile, T. M., and Kramer, S. J. (2007). Inner work life. *Harvard business review*, ***85***, 72-83.

Amabile, T. M., Hill, K. G., Hennessey, B. A., and Tighe, E. M. (1995). " The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations": Correction.

Andreassen, P. B. (1991). Causal prediction versus extrapolation: Effects on information source on judgmental forecasting accuracy. In *Sloan School of Management Working paper*.

Armstrong, J. S. (1985). *Long range forecasting: From crystal ball to computer (*second edition). NY: Wiley.

Armstrong, J. S. (1988). Research needs in forecasting. *International Journal of Forecasting.* **4** 449-465

Armstrong, J. S. (2001). The forecasting dictionary. In J.S. Armstrong (Ed.), *Principles of forecasting* (pp. 495-515). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting,* **22(3),** 583-598.

Armstrong, J.S. and Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. *Forecasting with judgment*. John Wiley and Sons Ltd., 269-293.

Baccetti, A., and N. Saccani. "Spare parts classification and demand forecasting for stock control." *Omega* **40** (2012): 722-737.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., and Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? *Organizational behavior and human decision processes*, ***53***, 35-54.

Bendoly, E., Croson, R., Goncalves., P., and Schultz, K. L. (2006). Bodies of knowledge for research in behavioral operations. *Productions and Operations Management,* **19(4),** 434-452.

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

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

Bolger, F., and Wright, G. (2011). Improving the Delphi process: lessons from social psychological research. *Technological Forecasting and Social Change*, ***78***, 1500-1513.

Bolton, G. E., Ockenfels, A., and Thonemann, U. W. (2012). Managers and students as newsvendors. *Management science,* **50(11),** *2225-2233.*

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

Boulaksil, Y. and Franses, P.H. (2009). Experts stated behaviour. *Interfaces*, **39**, 168-171.

Boylan, J. E., and Syntetos, A. A. (2010). Spare parts management: a review of forecasting research and extensions. *IMA journal of management mathematics*, ***21***, 227-237.

Budescu, D. V., and Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta psychologica*, ***104***, 371-398.

Budescu, D. V., Rantilla, A. K., Yu, H. T., and Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, ***90***, 178-194.

Bunn, D. (1987). Expert use of forecasts: Bootstrapping and linear models. In G. Wright, and P. Ayton (Eds.), *Judgmental forecasting* (pp. 229-241). Chichester: Wiley.

Bunn, D.W. and Taylor, J.W. (2001). Setting accuracy targets for short term judgmental sales forecasting. *International Journal of Forecasting*, **17**, 159-169.

Carbone, R., Andersen, A., Corriveau, Y. and Corson, P.P., (1983). Comparing for different time series methods the value of technical expertise individualized analysis, and judgmental adjustment. *Management Science*, ***29***, pp.559-566.

Cheung, C. K., Rudowicz, E., Lang, G., Yue, X. D., and Kwan, A. S. (2001). Critical thinking among university students: Does the family background matter? *College Student Journal*, ***35***.

Chase, C. W. (2013). *Demand-driven forecasting: a structured approach to forecasting*. New Jersey: John Wiley and Sons.

Christopher, M. (1992). *Logistics and Supply Chain Management.* Harlow: Pearson.

Clemen, R. T. (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting,* **5(4),** *559-583.*

Collopy, F. and Armstrong, J.S. (1992). Expert opinion about extrapolation and the mystery of the overlooked discontinuities*. International Journal of Forecasting*, **8**, 575-582.

Corsten, D., and Gruen, T. W. (2004). Stock-outs cause walkouts. *Harvard Business Review*, ***82***, 26-28.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society*, ***23***, 289-303.

Dalkey, N.C. and Helmer-Hirschberg, O. (1962). An experimental application of the Delphi method to the use of experts. *RAND Report RM-727*: Chicago.

Dalrymple, D. J. (1975). Sales forecasting methods and accuracy. *Business Horizons*, ***18***, 69-73.

Dalrymple, D. J. (1987). Sales forecasting practices: Results from a United States survey. *International Journal of Forecasting*, ***3***, 379-391.

Dawes, R. (1975). Graduate admission variables and future success. *Science,* **187**, 721-743.

De Baets, S., and Harvey, N. (2018). Forecasting from time series subject to sporadic perturbations: Effectiveness of different types of forecasting support. *International Journal of Forecasting.* **34(2)**, 463-487.

Diamantopoulos, A., and Mathews, B. (1989). Factors affecting the nature and effectiveness of subjective revision in sales forecasting: An empirical study. *Managerial and Decision Economics*, ***10***, 51-59.

Donnelly, J. M. (2013). The case for managing MRO inventory. *Supply Chain Management Review*, ***17***, 18-24.

Durach, C F., Kembro, J., and Wieland, A. (2017). A new paradigm for systematic literature reviews in supply chain management. *Journal of Supply Chain Management, **53(4),*** 67-85.

Ebrahim-Khanjari, N., Hopp, W. and Iravani, S. M. R. (2012). Trust and information sharing in supply chains. *Production and Operations Management,* **21(3),** 444-464.

Edmundson, R. H. (1990). Decomposition; a strategy for judgmental forecasting. *Journal of Forecasting*, ***9***, 305-314.

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

Erez, A., and Judge, T. A. (2001). Relationship of core self-evaluations to goal setting, motivation, and performance. *Journal of applied psychology*, ***86***, 1270.

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

Fildes, R., and Beard, C. (1992). Forecasting systems for production and inventory control. *International Journal of Operations and Production Management*, **12**(5), 4-27.

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

Fildes, R., Goodwin, P. and Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems*, **42**, 351-361.

Fildes, R. and Goodwin, P. (2007). Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting. *Interfaces, **36,*** 570-576.

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

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

Fischer, I., and Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International journal of forecasting*, ***15***, 227-246.

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

Franses, P.H. and Legestree, R. (2010). Do experts' adjustments on model-based SKU-Level forecasts improve forecast quality? *Journal of forecasting*, **29**, 331-340.

GAO (United States General Accounting Office). 2011. DOD's inventory management improvement plan, GAO-11-240R, Washington, D.C., Jan.07

GAO (United States General Accounting Office). 2012. Defense inventory: actions underway to implement improvement plan, but steps needed to enhance effort, GAO-12-493, Washington, D.C., May 03.

Gibbons, A. M., Sniezek, J. A., and Dalal, R. S. (2003, November). Antecedents and consequences of unsolicited versus explicitly solicited advice. In *D. Budescu (Chair), Symposium in Honor of Janet Sniezek. Symposium presented at the annual meeting of the society for judgment and decision making, Vancouver, BC*.

Goodwin, P. (2000a). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting,* **16**, 261-275.

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

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

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

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

Green, K.C. and Armstrong, J.S. (2007b). Value of expertise for forecasting decisions in conflicts. *Interfaces, **37,*** 287-299.

Green, K.C. and Armstrong, J.S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business research,* **68,** 1678-1985.

Griffin, D., and Bremner, L. (2004). Perspectives on probability judgment calibration. In D. J. Koelher and N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 177-199). Oxford: Blackwell.

Grushka-Cockayne, Y., Jose, V. R. R. and Lichtendahljr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science,* **63(4),** 1110-1130.

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

Harvey, N., Harries, C., and Fischer, I. (2000). Using advice and assessing its quality. *Organizational behavior and human decision processes*, ***81***, 252-273.

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

Harvey, N., and 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.

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

Holt, C. C. (2004). Forecasting seasonal and trends by exponentially weighted moving averages. *International journal of forecasting*, ***20***, 5-10.

Humphrey, S. E. , Hollenbeck, J.R., Meyer, C.J., Ilgen, D,R. (2002), Hierarchical team decision making. *Research in Personnel and Human Resources Management Research in Personnel and Human Resources Management,* ***21****,* 175 – 213

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

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

Judge, T. A. and Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*, **87**, 797-807.

Jungermann, H., and Fischer, K. (2005). Using expertise and experience for giving and taking advice. *The routines of decision making*, 157-173.

Katok, E. (2011). Using laboratory experiments to build better operations management models. *Foundations and Trends in Technology, Information and Operations Management,* **5(1),** 1-86.

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A., and Van Bruggen, G. H. (2009). How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research*, ***20***, 527-546.

Kim, S. H., Cohen, M. A., and Netessine, S. (2007). Performance contracting in after-sales service supply chains. *Management Science*, ***53***, 1843-1858.

Klassen, R.D. and Flores, B.E. (2001). Forecasting practises of Canadian firms: Survey results and comparisons. *International Journal of Production Economics*, **70**, 163-174.

Klayman, J. (1988). On the how and why (not) of learning from outcomes. *Human judgment: The SJT view*, 115-162.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: toward an integrative approach. *Psychological bulletin*, ***107***, 296.

Koudal, P. (2006). The service revolution in global manufacturing industries. *Deloitte Research*, ***2***, 1-22.

Kray, L. J. (2000). Contingent weighting in self-other decision making. *Organizational behavior and human decision processes*, ***83***, 82-106.

Kray, L., and Gonzalez, R. (1999). Differential weighting in choice versus advice: I'll do this, you do that. *Journal of Behavioral Decision Making*, ***12***, 207-218.

Krueger, J. I. (2003). Return of the ego--Self-referent information as a filter for social prediction: Comment on Karniol (2003). *Psychological Review,* ***110***, 585-590.

Lalonde, B.J. and Zinszer, P.H. (1976) Customer Service: Meaning and Measurement. *National Council of Physical Distribution Management*, Chicago, IL, 156-159.

Lawrence, M., and O'Connor, M. (2005). Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting*, ***21***, 3-14.

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

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

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

Li, B., Oliva, R., and Watson, N. (2018). Do retail managers rock or paddle the boat? Working paper. Texas A and M University, Collage Station, TX.

LI, M., Petruzzi, N. C., and Zhang, J (2017). Overconfident competing newsvendors. *Management science,* **62(9),** 2705-2721.

Limm, J.S., and O'Connor, M. (1995). Judgmental adjustments of initial forecasts: Its effectiveness and biases. *Journal of Behavioural Decision making,* **8**, 149-168.

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

Lim, K. H., O'Connor, M. J., and Remus, W. E. (2005). The impact of presentation media on decision making: does multimedia improve the effectiveness of feedback? *Information and Management*, ***42***, 305-316.

Litsiou, K., Polychronakis, Y., Karami, A., and Nikolopoulos, K. (2019). Relative performance of judgmental methods for the forecasting of megaprojects. Accepted for publication in *International Journal of Forecasting (*May 2019).

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1993). Forecasting. *Methods and Applications* (3rd edition). John Wiley and Sons, Inc., New York.

Maio, G. R., Pakizeh, A., Cheung, W. Y., and Rees, K. J. (2009). Changing, priming, and acting on values: effects via motivational relations in a circular model. *Journal of personality and social psychology*, ***97***, 699.

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

McCarthy, T. M., Davis, D. F., Golicic, S. L. and Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practises*. Journal of Forecasting*, **25**, 303-324.

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

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

Moritz, B., Siemsen, E., and Kremer., M. (2014). Judgmental Forecasting: Cognitive reflection and decision speed. *Production and Operations Management,* **23(7),** 1146-1160.

Morse, A. 2012. Ministry of defence: managing the defence inventory, National Audit Office (NAO), UK.

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

Nikolopoulis, K., Fildes, R., Goodwin, P. and Lawrence, M. (2005). On the accuracy of judgmental interventions on forecasting support systems.

Nikolopoulos, K., Litsa A., Petropoulos, F., Bougioukos, V., and Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of business research,* ***68(8),*** 1785-1791.

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

Oliva, R., and Watson, N. (2009). Managing functional bias in organisational forecasts: A case study of consensus forecasting in supply chain planning. *Production and operations management,* **18(2),** 138-151.

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

Önkal, D., Gönül, M.S. and Lawrence, M., (2008). Judgmental adjustments of previously adjusted forecasts. *Decision sciences*, **39**, 2.

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

Ozer, O., Zheng, Y., Chen, K. (2011). Trust in forecast information sharing. *Management Science,* **57(6),** 1111-1137.

Parikh, M., Fazlollahi, B., and Verma, S. (2001). The effectiveness of decisional guidance: an empirical evaluation. *Decision Sciences*, ***32***, 303-332.

Petropoulis, F., Fildes, R., and Goodwin. P. (2015). Do 'big losses' in judgmental adjustments to statistical forecasts affect experts' behaviour? *European Journal of Operational Research*

Petropoulis, F., Goodwin, P. and Fildes, R. (2017). Using a rolling training approach to improve judgmental extrapolations elicited from forecasters with technical knowledge. *International Journal of Forecasting*, **33**, 314-317.

Porter, M. (1985). *Competitive advantage*. New York: Macmillan.

Rekik, Y., Glock, C. H., and Syntetos, A. A. (2017). Enriching demand forecasts with managerial information to improve inventory replenishment decisions: Exploiting judgment and fostering learning. *European Journal of Operational Research,* **0,** 1-13.

Ren, Y., and Croson, R. (2013). Overconfidence in newsvendor orders: An experimental study. *Management science,* **59(11),** 2502-2517.

Rotter, J. B. (1954). Social learning and clinical psychology.

Rotter, J. B., and Mulry, R. C. (1965). Internal versus external control of reinforcement and decision time. *Journal of personality and social psychology*, ***2***, 598.

Rowe, G., and Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, ***15***, 353-375.

Sanders, N.R. (1992). Accuracy of judgmental forecasts: A comparison. *Omega,* **20**, 353.

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

Sanders, N. R., and Graman, G, A. (2016). Impact of bias magnification on supply chain: The mitigating role of forecast sharing. *Decision sciences,* **47(5),** 881-906.

Sanders, N.R., and Manrodt, K. B. (1994). Forecasting practices in US corporation: Survey results. *Interfaces*, **24**, 92-100.

Sanders, N. R., and Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, ***5***, 39-52.

Sanders, N. R., and Ritzman, L. (1995). Bringing judgment into combination forecasts. *Journal of Operations Management*, **13**, 311-321.

Sanders, N. R., and Ritzman, L. P. (2001). Judgmental adjustment of statistical forecasts. In *Principles of forecasting*, 405-416. Springer, Boston, MA.

Scheele, L. M., Thonemann, U. W., and Slikker, M. (2018). Designing incentive systems for truthful forecast information sharing within a firm. *Management Science,* **64(8),** 3690-3713.

Schweizer, M. E., and Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science,* ***46(3),*** 404-420.

Seifert, M., Siemsen, E., Hadida, A. L., and Eisingerich, A. B. (2015). Effective judgment forecasting in the context of fashion products. *Journal of Operations Management,* **36,** 33-45.

Silver, M.S. (1991). Decisional guidance for computer-based support. *MIS Quarterly,* **15,**105-133.

Singh, D. T. (1998). Incorporating cognitive aids into decision support systems: the case of the strategy execution process. *Decision Support Systems*, ***24***, 145-163.

Spiliotopoulou, E., Donohue, K., and Gurbuz, M. C. (2016) Information reliability in supply chains: The case of multiple retailers. *Production of Operations Management,* **25(3),** 548-567.

Spithourakis, G. P., Petropoulos, F., Nikolopoulos, K., and Assimakopoulos, V. (2015). Amplifying the learning effects via a forecasting and foresight support system. *International Journal of Forecasting* **31,** 20-32.

Stone, E. R., and Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational behavior and human decision processes*, ***83***, 282-309.

Surowiecki, J. (2004). *The Wisdom of crowds.* London: Abacus.

Syntetos, A. A., Kholidasari, and Naim, M. (2015). The effects of integrating management judgement into OUT levels: in or out of context? *European Journal of Operational Research*

Syntetos, A. A., Nikolopoulos, K., Boylan, J.E., Fildes, R. and Goodwin, P. (2008). Integrating judgement into intermittent demand forecasts. *International Journal of Production Economics*, **118**, 78-81.

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

Syntetos, A.A., Keyes, M.A. and Babai, M.Z. (2008). Demand categorisation in a European spare parts logistics network. *International Journal of Operations and Production Management,* Vol. **29**, 292 – 316

Syntetos, A. A., Babai, M. Z., Davies, J., and Stephenson, D. (2010). Forecasting and stock control: A study in a wholesaling context. *International Journal of Production Economics*, ***127***, 103-111.

Syntetos, A. A., and Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of forecasting*, ***21***, 303-314.

Syntetos, A.A., and Boylan, J.E. (2008). Demand forecasting adjustments for service-level achievement. *Journal of Management Mathematics,* **19,** 175−192

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

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

Teunter, R., and Sani, B. (2009). Calculating order-up-to levels for products with intermittent demand. *International Journal of Production Economics*, ***118***, 82-86.

Theocharis, Z., and Harvey, N. (2016). Order effects in judgmental forecasting. *International Journal of Forecasting*, **32**, 44-60.

Thompson, M.E., Pollock, C.A., Gonul, M.S. and Onkal, D. (2013). Effects of trend strength and on performance and consistency on judgmental exchange rate forecasting. *International Journal of Forecasting*, **29**, 337-353.

Tong, J., and Feiler, D., (2016). A behavioural model of forecasting: Naïve statistics on mental samples. *Management Science, **63(11),*** 3609-3627.

Trapero, R. J, Pedregal, J. D, Fildes, R., Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions *International Journal of Forecasting, **29,** 234-243.*

Wagner, S. M. (2006). Supplier development practices: an exploratory study. *European journal of marketing*, ***40***, 554-571.

Walker, C. O., Greene, B. A., and Mansell, R. A. (2006). Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement. *Learning and individual differences*, ***16*,** 1-12.

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

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

Wheelwright, S. C., and Clarke, D. G. (1976). Corporate Forecasting: Promise and Reality. *Harvard Business Review*, ***54***, 40-64.

Willemain, T. R., Smart, C. N., and Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of forecasting*, ***20***, 375-387.

Wilkie, M. E., Tuohy, A. P. and Pollock, A. C. (1993). Examining heuristics and biases in judgmental currency forecasting. *VBA journal,* **2**, 12-17.

Winklhofer, H., , A., and Witt, S. F. (1996). Forecasting practice: A review of the empirical literature and an agenda for future research. *International Journal of Forecasting*, ***12***, 193-221.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, ***6***, 324-342.

Worthen, B. (2003). Future Results Not Guaranteed; Contrary to what vendors tell you, computer systems alone are incapable of producing accurate forecasts. *CIO*, 1-1.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, ***93***, 1-13.

Yaniv, I. (2004). The benefits of additional opinions. *Organizational Behaviour and Human Decision Processes*, **13**, 2.

Yaniv, I., and Hogarth, R. M. (1993). Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psychological science,* **4**, 58-62.

Yaniv, I., and Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, ***83***, 260-281.

Yule, G. U. (1923). *An Introduction to the Theory of Statistics*. London, Griffen and Co., Ltd.