

Singapore Management University

Institutional Knowledge at Singapore Management University

Dissertations and Theses Collection (Open Access)

Dissertations and Theses

2-2023

Remediating system neglect in judgmental demand forecasting

Srikant VINAKOTA

Singapore Management University, srikantv.2017@dba.smu.edu.sg

Follow this and additional works at: https://ink.library.smu.edu.sg/etd_coll



Part of the [Business Administration, Management, and Operations Commons](#), and the [Organizational Behavior and Theory Commons](#)

Citation

VINAKOTA, Srikant. Remediating system neglect in judgmental demand forecasting. (2023). 1-124.
Available at: https://ink.library.smu.edu.sg/etd_coll/468

This PhD Dissertation is brought to you for free and open access by the Dissertations and Theses at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Dissertations and Theses Collection (Open Access) by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

**REMEDIATING SYSTEM NEGLECT
IN
JUDGMENTAL DEMAND FORECASTING**

VINAKOTA SRIKANT

SINGAPORE MANAGEMENT UNIVERSITY

2023

Remediating System Neglect In
Judgmental Demand Forecasting

Vinakota Srikant

Submitted to Lee Kong Chian School Of Business In Partial Fulfillment
Of The Requirements For The Degree Of Doctor Of Business
Administration

Dissertation Committee

Pascale Crama (Chair)

Professor Operations Management
Singapore Management University

Bhavani Shankar Uppari

Assistant Professor Operations Management
Singapore Management University

Wu Yaozhong

Associate Professor Analytics & Operations
National University of Singapore

SINGAPORE MANAGEMENT UNIVERSITY

2023

Copyright (2023) Vinakota Srikant

I hereby declare that this DBA dissertation is my original work, and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in this dissertation.

This DBA dissertation has also not been submitted for any degree in any university previously.

V. Srikant

Vinakota Srikant

15 February 2023

Remediating System Neglect In Judgmental Demand Forecasting

Vinakota Srikant

Abstract:

Prior research has shown that individuals tasked with judgmental forecasting of demand based on time-series data overreact in stable environments and underreact in unstable environments. Kremer et al. (2011) attributed this to the system neglect hypothesis, which claims that forecasters emphasize forecast errors over the system parameters.

The present research investigates interventions that mitigate system neglect and address the causal factors for overreaction and underreaction. Given the desire by organizations to move towards touchless planning and automated decision-making, minimizing human judgment and understanding its drivers is of significant practical importance.

We tested four different interventions on an online subject pool and found that the base treatment (simplest method in terms of cognitive load) outperforms all other interventions. In contrast to Kremer et al.'s original work we found a disconnect between subject's forecast adjustment scores and forecasting performance.

Table of Contents

| | |
|---|-----------|
| CHAPTER 1 INTRODUCTION AND PROBLEM STATEMENT | 1 |
| CHAPTER 2 LITERATURE REVIEW | 6 |
| 2.1 BIASES IN JUDGMENTAL FORECASTING | 7 |
| 2.2 IMPROVING JUDGMENTAL FORECASTING..... | 10 |
| CHAPTER 3 THEORY & HYPOTHESIS | 16 |
| 3.1 GRAPHICAL PRESENTATION OF INFORMATION..... | 20 |
| 3.2 DECOMPOSITION..... | 21 |
| 3.3 FORECAST BIAS FEEDBACK | 22 |
| 3.4 BIAS & ROLLING TRAINING | 24 |
| CHAPTER 4 METHODOLOY | 25 |
| 4.1 EXPERIMENTAL CONTEXT..... | 29 |
| 4.2 CONDUCTING THE EXPERIMENT AND DATA COLLECTION | 36 |
| CHAPTER 5 EMPIRICAL ANALYSIS | 37 |
| 5.1 VARIABLE DEFINITIONS..... | 39 |
| 5.1.1 <i>Dependent Variables</i> | 40 |
| 5.1.2 <i>Independent Variables</i> | 40 |
| 5.1.3 <i>Control Variables</i> | 40 |
| 5.2 INITIAL ANALYSIS | 44 |
| 5.2.1 <i>Base Treatment</i> | 45 |
| 5.2.2 <i>Adjustment Scores</i> | 46 |
| 5.3 MULTILEVEL NESTED MODELS | 50 |
| 5.4 HYPOTHESIS TESTING | 53 |
| 5.4.1 <i>Fan Charts</i> | 54 |
| 5.4.2 <i>Decomposition</i> | 59 |
| 5.4.3 <i>Bias</i> | 62 |
| 5.4.4 <i>Bias with Rolling Training</i> | 66 |
| 5.4.5 <i>Full Parameter Knowledge Treatment</i> | 68 |
| CHAPTER 6 FORECASTING PERFORMANCE IMPLICATIONS | 73 |
| CHAPTER 7 MANAGERIAL IMPLICATIONS | 77 |
| CHAPTER 8 CONCLUSION..... | 80 |
| REFERENCES..... | 84 |

| | |
|---|------------|
| APPENDIX 1 (RESPONDENT RECRUITMENT MATERIAL) | 89 |
| APPENDIX 2 (TREATMENT SPECIFIC INSTRUCTIONS) | 91 |
| LIST OF FIGURES..... | 115 |
| LIST OF TABLES | 116 |

ACKNOWLEDGEMENT

The Guru mantra is an homage to the Guru or the teacher, considered the dispeller of darkness.

Gurur Brahma Gurur Vishnu Gurur Devo Maheshwaraha

Guru Saakshaat ParaBrahma Tasmai Sri Gurave Namaha

The Guru is the one who creates and sustains knowledge while destroying ignorance, akin to the Trinity in Hinduism. This work acknowledges the Gurus who helped shape my thought process and character from primary school to graduate school. I am forever grateful to Prof. Pascale Crama, who has steadfastly supported me in this journey. She has challenged my thinking and helped me push the boundaries as a researcher. My committee members, Dr. Bhavani Shankar Uppari and Dr. Wu Yaozhang, have been generous with their time and suggestions to help shape this work.

I also would like to acknowledge the generous grant from ABRI (ASEAN Business Research Initiative), which supported the cost associated with the research.

Thank You!

Srikant Vinakota.

DEDICATION

This work is dedicated to my family. My wife, Suma, offered her unconditional support during this journey. My kids put up with their father during the weekends he was away. My mother, with her simplicity and generosity, never ceases to amaze me. My father instilled a sense of continuous learning and unfortunately passed away before the work came to fruition.

1. Introduction and Problem Statement:

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know." — John Kenneth Galbraith

The above quote by the renowned economist John Galbraith succinctly summarizes the challenges associated with the forecasting process. A forecast is usually the first step in the business cycle, which triggers subsequent supply chain activities such as distribution planning, production planning, raw material procurement, and manufacturing. Forecasts also form the basis of business plans, including setting financial targets and measuring the progress toward achieving those targets. Forecasting also assumes greater significance since the subsequent steps in supply chain planning are usually automated by planning systems based on pre-defined parameters such as order quantities, safety stock, and lead time.

Inaccurate forecasts can trigger disproportionate errors (both in internal functions and in external trading partners) in the subsequent processes ranging from product shortage/overage and resource utilization issues to missing financial targets. Some causes of forecast inaccuracies include manual adjustments to system-generated forecasts, poorly applied statistical models, incomplete or inaccurate historical data, incorrect assumptions, incentives, functional biases, and organizational politics. While it is a commonly accepted aphorism that forecasts are never accurate, it still makes sense from an organizational standpoint to invest in improving the forecasting process's accuracy, especially given

the extended supply chain lead times and multi-step production and distribution processes. Figure 1 below shows a typical forecasting process in an organization.

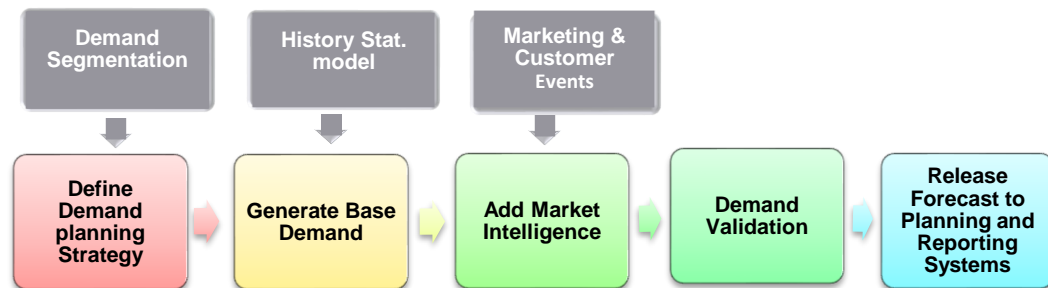


Figure 1 Typical Forecasting Process in a Firm

The process can be categorized into five major activities, which are briefly discussed as follows:

- i) Defining demand planning strategy - is mainly concerned with developing a demand planning approach for products/product families, considering demand segmentation, go-to-market approach, and performance metrics.
- ii) Generating base demand – This process involves applying analytical models that consider historical data to estimate the future demand for a product, e.g., time series forecasting models. For new products without any previous demand history and segmentation, the demand history of similar products, management judgment, or a combination of both is leveraged to create a demand forecast.
- iii) Add Market Intelligence – Analytical models may not accurately predict events such as promotions, competitor activity, etc.

Market intelligence requires the identification of external factors that may impact the forecast.

- iv) Demand Validation: Base demand and market intelligence together form the demand plan. The various organizational units review the demand plan in forums such as demand review meetings, S&OP meetings, etc.
- v) Forecast Release: The final step involves releasing the validated demand plan to supply planning and reporting systems

As figure 1 illustrates, the reviewers can adjust forecasts generated by the planning systems to consider factors not considered by the statistical models and add in market intelligence. The forecast resulting from the manual adjustment is usually termed in the literature as a **judgmentally adjusted forecast**. A pure judgmental forecast is, by contrast, derived manually without the help of any statistical models or tools.

A recent Fildes and Petropoulos (2015) study finds that only 29% of forecasters use statistical models exclusively. Of the balance, 71% use judgment or a combination of judgment and statistical forecasting to generate forecasts. Their results are comparable to the previous survey by Fildes and Goodwin (2007). Fildes et al. (2009) analyzed data on 600,000 forecasts generated across four companies and found that large adjustments generally improved accuracy while smaller adjustments negatively impacted accuracy.

An interesting phenomenon related to judgmental forecasting is the concept of system neglect. Massey & Wu (2005) originally developed the system neglect hypothesis, and Kremer et al. (2011) later applied it to time series forecasting. The system neglect hypothesis essentially states that individuals emphasize the recent signals, i.e., forecast errors, relative to the system's characteristics that generate the signals. Kremer et al. (2011) applied the system neglect hypothesis to time-series forecasting. From a time-series forecasting perspective, a stable environment is a time series with constant mean and variance. In contrast, an unstable environment is characterized by both noise and change in the time series mean.

Kremer et al. (2011) demonstrated that individuals overreact to forecast errors in stable environments and underreact in unstable environments. Their research suggests that human judgment is inherently more suited for unstable environments than stable but noisy environments. The system neglect findings, in conjunction with the earlier results about the use of judgmental forecast, have potentially significant implications for forecasters and management alike. It implies that judgmental changes should be minimized in stable environments (use of normative models maximized) and human judgment be reserved for unstable environments.

Given the desire by organizations to move towards touchless planning and automated decision-making, minimizing human judgment for stable environments assumes greater significance. Decision-makers need to be made aware of their tendency to overreact and underreact in the

various demand environments and factor in the system's characteristics. In other words, we need to remediate the system neglect and rectify both overreaction and underreaction. Our research intends to identify approaches that will remediate the impact of the system neglect and reduce the salience of the error, which are the key causal factors behind underreaction and overreaction, respectively.

2. Literature Review:

One of the most comprehensive reviews of Judgmental Forecasting was published by Lawrence et al. (2006). Key findings from the study include that results from judgmental forecasting can be as accurate as statistical methods, but this is not guaranteed. Lawrence et al. (2006) suggest that while the literature provides practical principles which can be applied, the results from the literature are contradictory. Individual performance is dependent on the characteristics of the time series, and small changes in the time series and its presentation can have significant changes in performance. As an example, findings from Bolger and Harvey (1993) indicate that individuals use different versions of anchor and adjustment heuristics for trended and untrended data, and their adjustments are suboptimal and biased. In the case of trended data, they found significant evidence of serial dependence in the time series.

Lawrence et al. (2006) also cite research that finds contradictory findings on the impact of noise on judgmental forecasting. These contradictions could be attributed to varying biases due to *"different beliefs about the nature of the time series being forecasted."* In a more recent review, Perera et al. (2019) adopt a supply chain lens to judgmental forecasting and update the Lawrence et al. (2006) review. They find increased use of judgmental forecasting in retail promotions, increased research into the behavioral elements of forecasting, and effectiveness of task feedback over performance feedback in the case of Judgmental Forecasting. Arvan et al. (2019) reviewed research involving the integration of human judgments with analytical models. They suggested

that rules-based forecasting (wherein condition – actions statements are used to integrate human judgment and statistical models) and Forecasting by Analogy (use of structured analogies to incorporate insights from similar past events) produce superior results.

As operations management evolves to incorporate various bodies of knowledge from cognitive psychology, social psychology, group dynamics, and system dynamics (Bendoly et al., 2009), research in Judgmental Forecasting has also evolved to integrate these bodies of knowledge. Donahue et al. (2019) and Fanhimnia et al. (2019) provide a good review of this area of research.

The preceding sections provided a broad overview of articles covering extant research in Judgmental Forecasting. However, from our research's scope, there are two streams of research on Judgmental Research that are relevant **i)** Biases associated with the Judgmental Forecasting process and **ii)** How to improve judgmental forecasting. We offer a brief review of the two streams of literature below.

2.1 Biases in Judgmental Forecasting:

Biases have been investigated extensively as causal variables impacting the accuracy of judgmental forecasts. Biases result from heuristics, which are simplified decision rules applied by individuals in varying scenarios. Harvey (2007) proposes that individuals choose one of availability, representativeness, or anchoring-adjustment heuristics based on the type of information available to them. E.g., anchoring refers to giving extra credence and basing the decision on a potentially

irrelevant piece of information or observation. The anchor on which the individual bases his / her decision varies. In the context of forecasting, two anchors are relevant, i) the statistical forecast (Eroglu & Croxton 2010) and ii) the financial sales target. In anchoring based on the statistical forecast, extra weight is placed on historical data, ignoring recent trends and actuals. In anchoring based on the financial sales target, forecasts are manipulated to hit a commercial target without adequately adjusting for current trends and actuals. Abundant anecdotal evidence suggests anchoring based on sales targets leads to inventory surpluses and write-offs.

Overconfidence is an often-studied bias in the case of interval forecasting. The overconfidence bias in forecasting can manifest itself as Overprecision, Overestimation, and Overplacement (Healy & Moore 2007, Bazerman 2013). Overprecision from a forecast standpoint means the forecaster is too sure of his judgment and has a tighter confidence interval (lower variance) surrounding his forecasts. Overestimation refers to excessive confidence in one's capability. From a forecasting standpoint, overestimation contributes to overly optimistic demand estimates. Overplacement refers to falsely ranking one ability higher than others. Overconfidence leads to systematic and predictable errors in the forecast (e.g., Gino & Pisano 2008, Bendoly et al. 2010).

Within individuals, varying traits have been found to mitigate biases. Moritz et al. (2014) studied one such trait called cognitive reflection. Cognitive reflection is distinct from intelligence and measures individuals' ability to defer their initial response and engage in deeper

analysis. Prior work has identified that individuals with high levels of cognitive reflection are less susceptible to errors associated with heuristics and biases (Toplak et al., 2011). Moritz et al. (2014) focused on individual factors such as engaging in cognitive reflection and the time taken to generate the forecast. They demonstrate that individuals with higher levels of cognitive reflection deliver better forecasting performance even while controlling for intelligence. Moritz et al. (2014) used a learning model to predict the time required for the forecasting task. They found that forecast errors increase when the time spent relative to the learning model is less or more (which they suggest is a measure of under or overthinking). Kremer et al. (2016) find that judgmental forecasting is more effective in a top-down process (distribution center to store) as compared to a bottom-up (store to distribution center).

Feiler et al. (2013) find evidence for judgment biases in censored environments (where individuals do not have access to all the true values when deciding) and individuals exhibit overly risk-averse behavior. Tong et al. (2018) propose a behavioral remedy to the censorship bias and find that asking individuals to estimate the missing values explicitly helps reduce the bias by creating a more representative sample. Tong and Feiler (2017) use the sample naivete theory to advance a behavioral model of forecasting. They argue that individuals generate a forecast based on a small (less than 7) and randomly generated sample of the series and naively assume that the sample represents the true population.

Motivational biases are usually introduced by incentives and can play a significant part in forecasting. While sales personnel may be motivated to inflate a forecast to ensure product availability and maximize sales, operations may be motivated to minimize inventory and obsolescence to achieve financial incentives. Oliva and Watson (2009) used a case study-based approach to identify functional biases and categorize them into intentional and unintentional biases. Incentives and power balance within the organizations are identified as the cause of intentional biases. They identified a consensus forecasting process driven by an independent group as a critical mechanism to address the shortcomings resulting from the biases. In a more recent study, Scheele et al. (2018) argued that rewarding salespeople based on forecast accuracy is not enough to enable truthful information sharing. They should also be penalized for over-forecasting more severely than under-forecasting.

2.2 Improving Judgmental Forecasting

There is extensive literature that has focussed on how to improve human judgment in forecasting. The research on improving Judgmental Forecasting has addressed two significant drivers of forecast inaccuracies, i) the inconsistency of the decisions and ii) biases. The inconsistency of the decisions is also termed noise or random error. Bias, as elaborated earlier, refers to a systematic deviation resulting from heuristics employed in the decision-making process. The key themes emerging from the literature which are relevant for improving judgmental time series forecasting are i) Presentation of the Data, ii) Decomposition

of the time series, iii) Feedback, iv) Using groups of individuals, and v) Training. We provide a brief overview of each of the key themes.

i) **Data Presentation:** Goodwin and Wright (1993) find qualified evidence that graphically presenting information helps short-term forecasting tasks, whereas tabular displays work better in longer-horizon tasks. Later work by Lawrence et al. (2006) claims that there is no conclusive evidence of the superiority of the graphical presentation mode over tabular presentation. Perhaps, Harvey (2001) provides the most pragmatic guidance. Harvey (2001) recommends presenting the time-series data in a graphical format as a preferred approach in a judgmental forecasting task when forecasters do not have any prior information about the series.

One of the most consistent results from the trended time series research is the trend-dampening phenomenon. Trend dampening results in lower / higher than the optimal forecast in case of an upward or downward trended forecast. Trend dampening is attributed to the anchor and adjustment heuristics and random error. Harvey (2001) suggests that fitting a line through the data series helps reduce the random error even though it does not do much to reduce the anchoring.

From a point forecasting approach, uncertainty associated with the forecast is not explicitly considered. However, as Kremer et al. (2011) demonstrate, the level of noise or change impacts the overreaction or underreaction and hence the forecast accuracy. Kreye et al. (2012) researched different approaches to display uncertainty in cost estimates (three-point trend forecast, bar-chart, and a fan chart) and found that fan

charts are the most effective means to increase awareness of uncertainty associated with the data.

ii) **Decomposition:** Decomposition is another tool used to aid judgemental forecasting. It involves breaking down the task into simpler components that are easier to estimate than the target variable. The component estimates are then aggregated to create a forecast. Decomposition may be accomplished graphically or numerically. Prevailing research suggests decomposition is suitable when time series have high uncertainty (McGregor 2001). Lee & Siemsen (2017) found that decomposition coupled with decision support improved the performance of the newsvendor problem. They decomposed the problem into point forecasts, uncertainty estimates, and service level projections.

iii) **Feedback:** The extant research distinguishes between outcome feedback, performance feedback, cognitive process feedback, and task properties feedback (Lawrence et al. 2006, Donahue et al. 2019). Outcome feedback simply means providing information about the accuracy of the last forecast. Performance feedback reports forecast accuracy over multiple periods. While outcome feedback is suitable for simple forecasting tasks, complex tasks require more highly processed feedback (e.g., forecast bias). This is because both outcome and performance feedback provides limited information on improving accuracy.

Furthermore, outcome and performance feedback make the individuals sensitive to recent errors, which may be simply a result of noise

(Donahue 2019). Petropoulos et al. (2017) employed a rolling training approach and provided bias feedback to forecasting experts resulting in improved performance. Harvey (2001) suggests that sharing past records of forecasts and feedback helps improve forecasting accuracy. Feedback also helps overcome hindsight and confirmation biases. Hindsight bias usually drives forecasters to overestimate their accuracy, while confirmation bias will cause subjects to look for evidence supporting their beliefs.

iv) **Combining Forecasts:** As noted earlier, individual forecasts are often characterized by both random error and biases. By averaging forecasts of groups of individuals whose forecast errors are negatively correlated with one another, the accuracy of the forecasts may be improved (Harvey 2001, Goodwin 2000). This approach has also been referred to as the "Wisdom of the Crowds" (Surowiecki 2005). Several variants to the combination approach have been studied judgmental weights instead of simple/weighted averages, combined with a statistical forecast, etc. Combining statistical and judgmental forecasts can reduce inconsistency in the forecasts and incorporate contextual factors (e.g., Sanders and Ritzman 1995). The popular Delphi Method is another approach to combining forecasts. The Delphi Method has improved accuracy beyond a simple average (e.g., Goodwin and Song 2014).

v) **Technical Knowledge & Training:** Whereas some research (Sanders and Ritzman 1992, Edmundson 1990) has found no impact of training on the accuracy of judgmental forecasting, other research (Lawrence 1985) has found training / technical knowledge helps improve

accuracy when information is presented in a tabular format. Legerstee and Franses (2014) used a natural experiment in a pharmaceutical company and provided experts with various training and feedback. They found that the accuracy increased over the year. Petropoulos et al. (2017) found that the rolling training approach combined with feedback on bias helped improve the forecast accuracy of forecasters with technical knowledge. A rolling training approach relies on providing forecasters with feedback on their performance (forecast bias) at regular intervals and including the complete records.

A neighboring area of research has been the types of decision support. Siefert et al. (2015) reviewed the impact of contextual and historical factors on the effectiveness of judgmental forecasting in the fashion industry. In a purely judgmental forecasting scenario, they find that providing historical and contextual factors is beneficial; however, when human judgment is combined with statistical models, providing only contextual factors is better. This is because statistical models are better suited to detecting patterns than humans. They also suggest that humans are more skilled at identifying the interaction between contextual factors.

The above summary indicates that there has been active research on judgmental forecasting and how to improve judgmental forecasting. However, there have not been any significant efforts to mitigate the system neglect found by Kremer et al. (2011), which leads to an overreaction in a stable environment and an underreaction in an unstable environment. The present research applies the learnings from

the rich history of judgmental forecasting to identify measures that could remediate the system neglect. Such measures could assist decision-makers in calibrating their response by making them more aware of their system parameters and reducing the salience of errors. The measures would also aid organizations in moving towards touchless planning, which relies on reducing human intervention.

3. Theory & Hypothesis

Applying the System Neglect theory of Massey and Wu (2005) to the task of time series forecasting, Kremer et al. (2011) tested decision makers' ability to distinguish change from noise. They posited that forecasters emphasize forecast errors more than the system parameters (noise and change levels). This is due to the salience of the forecast errors over the system parameters. The forecasters did not have a priori knowledge about the system parameters and estimated them based on the demand signal and the forecast errors. Comparing the subject behavior to the normative forecasting model (a single exponential smoothing model for a time series with noise and change), they find evidence for overreaction in stable environments and underreaction in unstable environments.

Kremer et al. (2011) model the demand process as follows:

$$D_t = \mu_t + \varepsilon_t \quad \mathbf{1(a)}$$

$$\mu_t = \mu_{t-1} + v_t \quad \mathbf{1(b)}$$

Where D_t represents the actual demand in time t , μ_t represents the true level of the time series. $\varepsilon_t \sim N(0, \sigma^2)$ and $v_t \sim N(0, \sigma^2)$ are independent normal random variables representing the noise and the change components. Noise represents a temporary disruption valid for only one period, whereas level change represents a permanent change. The single exponential smoothing model of the forecast can be expressed as:

$$F_{t+1} = \alpha D_t + (1 - \alpha)F_t \quad \mathbf{2(a)}$$

$$= F_t + \alpha(D_t - F_t) \quad \mathbf{2(b)}$$

The above equation indicates that the forecast for period t+1 produced in period t is a weighted average of the demand in period t and the forecast for period t. Equation 2(b) states that the forecast is a function of the error and the weight assigned to the error. Kremer et al. (2011) term the forecast error as the strength of the error and the factor α as the weight of the error. The weight is determined by the system parameters (c and n). The strength factor implies that ceteris paribus forecasts with higher errors should be adjusted to a greater degree. The optimal value of α per Kremer et al. (2011) is

$$\alpha^*(W) = \frac{2}{1 + \sqrt{1 + 4/W}} \quad \mathbf{(3)}$$

W represents the change to noise ratio in the above equation and is defined as $W = c^2/n^2$. Replacing the optimal value of α in equation 2(b) results in the following:

$$F_{t+1} = F_t + \alpha^*(W)(D_t - F_t) \quad \mathbf{(4)}$$

Reviewing equations 3 and 4 jointly implies that when $c=0$, α^* is 0, or in other words, when the time series is characterized by only random noise, the most recent period forecast is the optimal forecast. Based on the above, Kremer et al. (2011) claim that "forecast errors should be mostly discarded and should not influence the new forecast. In contrast, with high values of W (variations in demand mostly represent level changes), the forecast error should have a greater influence on a forecast". However, this is not the case. They demonstrate that individuals

overreact in stable environments and underreact in unstable environments, meaning the forecasters' α is greater than α^* in stable environments, and the forecasters' α is less than α^* in unstable environments.

Extending on Kremer et al. (2011), we argue that overreaction/underreaction due to system neglect can be reduced by interventions that trigger the subject to think about the system parameters and reduce the salience of forecast errors. Based on the survey of the literature presented in the preceding paragraphs, we investigate the impact of i) Graphical representation (Fan Charts), ii) Decomposition of the forecasting task, and iii) Feedback on Forecast Bias on the forecaster's performance

Before delving into details about the various treatments and how they impact overreaction and underreaction, a brief definition of the forecast performance metrics is in order.

- i. **Forecast error** is defined as the difference between the actual sales and the forecast.
- ii. **Absolute forecast error** refers to the absolute value of the difference between forecast and actuals.
- iii. **Absolute Percentage Error (APE)** is the Abs forecast error expressed as a percentage of the actual sales.
- iv. **MAPE** refers to the mean of the Absolute percentage error over the given time periods

- v. **Mean Absolute Error:** It is defined as the mean of the absolute deviation of the observed value from the mean
- vi. **Mean Squared Forecast Error:** It is defined as the mean of the squared forecast errors
- vii. **Variance of forecast errors** is defined as the sum of the squared forecast errors divided by one less than the total number of observations (Markridakis et al. 1998)
- viii. Information on **forecast bias** indicates the tendency to over-forecast or under-forecast. Bias is the Forecast Error divided by the Actual Sales.
- ix. **Forecast accuracy**, in turn, is defined as $(1 - \text{MAPE})$. Table 1 below illustrates the calculation of each of the metrics. This measure can be calculated for a particular time period or over all the time periods (cumulative forecast accuracy).

| Time Period | Actual Sales | Forecast | Error | APE (%) | Variance of Forecast Error | MAPE (%) | Forecast Accuracy Current Period | Mean Forecast Accuracy (100-MAPE) |
|-------------|--------------|----------|-------|---------|----------------------------|----------|----------------------------------|-----------------------------------|
| 37 | 100 | 90 | 10 | 10.0% | - | 10.0% | 90.0% | 90.0% |
| 38 | 90 | 100 | -10 | 11.1% | 200 | 10.5% | 88.9% | 89.5% |
| 39 | 80 | 60 | 20 | 25.0% | 233 | 14.8% | 75.0% | 85.2% |

Table 1 Calculation of Forecast Measures

3.1 Graphical Presentation of Information:

Harvey (2001) summarized the extant literature on the benefits of using graphical methods in the judgmental forecasting task. Harvey (2001) recommended presenting the time-series data in a graphical format as a preferred approach in a judgmental forecasting task when forecasters do not have any prior information about the series. In a time-series forecasting task involving noise and change, subjects should discern between noise and an actual level change. Bank of England first used fan charts in 1996 to communicate the uncertainty associated with point forecasts.

Fan charts can be set up based on prediction or confidence intervals. Kreye et al. (2012) found that a fan chart (compared to a three-point trend forecast or a bar chart) was most effective in making subjects aware of the uncertainty associated with cost forecasting. Visual fan charts which show the expected dispersion around the mean are an effective means to represent uncertainty. We argue that underscoring the uncertainty (noise) associated with the forecasting process will aid in reducing the salience of the forecast errors, which is desirable in stable scenarios.

In unstable scenarios, additional insights into the nature of the time series are desirable to illustrate the unstable nature of the demand. Since the demand function is devoid of any seasonality or trends, seasonal plots or trend charts are not suited. They may mislead the subjects to assume seasonality or trends when no such component exists in the demand function. Hence, in addition to the fan charts, we

propose presenting subjects with a line chart depicting the moving average of the past four time periods would also highlight the changing nature of the demand.

However, we cannot be sure if the reduction in the salience of errors also leads to a greater appreciation of the system parameters. Reducing the salience of errors is desirable in scenarios with low W values, where most errors are triggered by noise in the time series and should be ignored. In contrast, scenarios with high values of W forecast errors need to influence future forecasts more. The preceding arguments lead us to our first hypothesis:

H1a: *Use of visual fan charts in the judgmental forecasting process reduces the overreaction*

H1b-i): *Use of visual fan charts in the judgmental forecasting process reduces the underreaction*

H1b-ii) *Use of visual fan charts in the judgmental forecasting process increases the underreaction*

3.2 Decomposition:

In its simplest form, decomposition involves breaking down a variable into its various components and then aggregating the components to estimate the variable. Decomposition is recommended when estimating the components is easier than directly estimating the variable of interest. As applied to time series forecasting, traditional decomposition involves breaking down the time series into the trend, seasonal, and error components. They can take an additive or a multiplicative format

(Makridakis et al., 1998). Once the individual components are estimated, they can be aggregated mechanically, or the forecaster could aggregate manually based on the component estimates.

Interestingly Lee & Siemsen (2017) found that performance benefits were derived from the simplification associated with the decomposition task, and the form of aggregation (mechanical or manual) did not impact performance. We predict that eliciting the forecaster to estimate the mean demand, the change, and the error components will prompt the forecasters to consider the system parameter explicitly and reduce the salience of the errors by highlighting errors as an integral part of the forecasting process. Hence, we hypothesize:

H2: *Decomposition of the time series into individual components reduces overreaction and underreaction*

3.3 Forecast Bias Feedback:

The literature on feedback has argued the superiority of performance feedback over outcome feedback. In the case of outcome feedback, the forecaster only receives the actuals associated with the forecast. Feedback on forecast bias is a form of performance feedback. In contrast with accuracy (which does not provide a directional indication of the forecast error), bias is more processed information, which tells forecasters if they are over or under-forecasting. The bias results are not averaged across the time periods (as in the case of MAPE). In the case of time series with noise but no change, bias metrics are easily interpretable and actionable. Bias information is also helpful in series

with monotonic trends as it can highlight if the subject is under-forecasting or over-forecasting.

However, for unstable time series where the true level of the demand (μ_t) of the demand changes for each time period, bias information by itself may be difficult to action. Kremer et al. (2011) claim it is difficult to estimate "the extent that a variation in the demand signal D_t is evidence for a permanent change in the level rather than a random, transient shock." The overreaction and underreaction that Kremer et al. (2011) have demonstrated is another way bias is manifested.

Based on the above arguments, we predict that,

H3a: *Providing bias feedback reduces overreaction*

Effectively mitigating underreaction in unstable time series requires an appreciation of the system parameters. In addition, in scenarios with high values of W , forecast errors need to have a greater influence on future forecasts. As argued earlier, we cannot confidently predict forecasters will be able to incorporate bias feedback to help mitigate the underreaction; hence we state the impact of bias feedback as two-part hypotheses:

H3b-i): *Providing bias feedback in the judgmental forecasting process reduces the underreaction*

H3b-ii): *Providing bias feedback in the judgmental forecasting process increases the underreaction*

3.4 Bias and Rolling Training

Petropoulos et al. (2017) use an approach called "rolling training" to provide feedback on the bias to experts. They found that the rolling training approach reduced MAPE by 3.78%. The approach relies on giving forecasters feedback on their performance (bias) at regular intervals, including the complete records. The authors claim this approach enables the *"balance between the sensitivity and stability of the feedback."*

Similar findings are reported in a study related to the review of project abandonment review decisions by Long et al. (2020). The authors found that by limiting the number of reviews, reviews' "decision-making value" becomes more salient. The limited number of reviews makes the participants more cognitively attentive to information. We propose that the above findings can be extended to our current study domain of time-series forecasting.

The rolling approach to feedback underscores the demand generation process, and the magnitude of the bias provides processed feedback about the error. A forecast with higher bias levels will need to be adjusted by a greater amount in the opposite direction. While Petropoulos et al. (2017) studied the bias feedback and the rolling training approach in various time series (stationary, trended, seasonal both trended and seasonal), they did not study a non-stationary process. Therefore, we argue that a rolling approach paired with bias feedback will enable a reduction in overreaction and underreaction

H4: *A rolling approach to providing bias feedback reduces overreaction and underreaction*

4. Methodology:

The testing of the hypotheses requires an experimental setup. With the Covid-19 pandemic, scheduling in-person experiments in behavioral labs is a logistical challenge and not permitted in some cases. Online subject pools such as MTurk and Prolific have recently gained popularity and present a potential opportunity for researchers relying on experiments as a methodology. Lee et al. (2018) employed Amazon's MTurk subject pool to replicate the findings from prior studies conducted in a behavioral laboratory. They replicated the findings of the physical laboratory studies; however, they found that learning online occurs more slowly than in the physical laboratory.

Prolific is another service provider which has grown in popularity recently due to claims of a better-quality subject pool. A key difference between Prolific and MTurk is the cost: while Prolific requires subjects to be compensated based on minimum wage per UK standards, MTurk does not impose a minimum wage. MTurk also offers access to a pool of highly rated subjects at a premium price. Due to the lower cost structure and the large subject pool required for the experimental scenarios, MTurk was chosen as the platform to host the experiment. Participants were requested to sign up for the study via postings on the MTurk platform. The recruiting material used in MTurk for one of the treatments is included in Appendix 1. For other treatments, the recruitment material is similar, with minor modifications to describe treatment-specific nuances.

Pre-screen criteria of participant approval rate (greater than 95%) and the number of previous submissions (>100) are applied to ensure participants with reasonable quality and experience. Participants who participated once in an experimental scenario were not permitted to participate in other experimental scenarios to avoid learning effects. Upon enrolling in the study, participants were directed to a Qualtrics survey which provided the participants with a spreadsheet-based simulation tool.

The Qualtrics survey contained written instructions and detailed various performance measures associated with the forecast. Participants had the opportunity to revisit the instructions if they needed additional clarification. In addition, a short video instruction specific to each scenario was provided to the participants to help them familiarize themselves with the spreadsheet and the task. Participants also had the option to withdraw from the study if they chose to. All participants were paid US\$ 2 for completed responses. The top 3 respondents with the lowest MAPE within a given experimental scenario were rewarded with a bonus of US\$ 15 each. Participants' responses were automatically linked to their MTurk ID via piped text.

In their original study, Kremer et al. (2011) varied the change level to low, medium, and high, while noise was varied between low and high levels resulting in six distinct conditions. In our study, we test the extreme conditions, i.e., low and high change and low and high noise, resulting in four conditions. The four conditions are replicated for the base and additional treatments (Fan Chart, Decomposition, Bias, and Rolling

Feedback). This ensures that proposed treatments thoroughly address the critical low and high levels of change scenarios and all noise levels while limiting the number of experiments to a reasonable scope. The conditions are summarized in Table 2 below. Table 3 below shows the values of optimal α based equation 3 presented earlier.

| | <i>n= 10 (Low)</i> | <i>n =40 (High)</i> |
|------------------------|---------------------|---------------------|
| <i>c = 0 (Low)</i> | <i>Condition 1</i> | <i>Condition 2</i> |
| <i>c = 10 (Medium)</i> | <i>Not in Scope</i> | <i>Not in Scope</i> |
| <i>c = 40 (High)</i> | <i>Condition 3</i> | <i>Condition 4</i> |

Table 2: Listing of Experiment Conditions

| | <i>n= 10 (Low)</i> | <i>n =40 (High)</i> |
|----------------------|--------------------|---------------------|
| <i>c = 0 (Low)</i> | <i>0</i> | <i>0</i> |
| <i>c = 40 (High)</i> | <i>0.94</i> | <i>0.62</i> |

Table 3: Optimal Alpha by Condition

A time series with the same parameters (c and n) may have different demand realizations. We adopt Kremer et al.'s (2011) approach and, for each condition, generate four different demand realizations, in line with the original study using the demand function described in equations 1(a) and 1(b) prior to the experiment. The starting mean μ_0 is 517 (similar to the 500 from the original paper). The demand sets are the same across

all the treatments. Subjects are randomly assigned to the demand set within a treatment and condition.

The four conditions and the five treatments lead to a full factorial design of 20 scenarios.

4.1 Experimental Context:

The context is the same for all the treatment conditions. Participants are advised that the objective of the experiment is to study individual judgments. The experiment is not intended to test their knowledge but to understand the individual decision-making process. Participants take on the role of a demand planner for a super-market and are tasked to forecast the unconstrained demand for a product. Participants are provided the historical data for the last 36 periods. They are required to predict the subsequent 36 time periods, one step ahead for the next selling period, i.e., forecast period 37, using data available until period 36.

Actual sales for period 37 are then realized, and participants then forecast period 38. The process continues till period 72. The spreadsheet tool is designed to prevent changing previously submitted forecasts and submission of a negative forecast. Following the simulation, the participants complete a brief survey about their understanding of the forecasting process and the nature of the demand.

We now proceed to provide a brief description of each treatment below:

i). Base Treatment (Control Group): In the base condition or the control group, we replicate Kremer et al.'s (2011) study for four

conditions. The participants were provided with historical product sales and asked to forecast the future demand one period at a time. Performance measures such as Forecast Error, APE, and MAPE were provided. In addition, the actual sales and forecast information was displayed in a line chart for the participants to review. Screenshots of the input screens and the line charts for the base condition are shown below.

| Serial Number | Time Period | Actual Demand | Your Forecast | Forecast Error | APE (%) | MAPE (%) |
|---------------|-------------|---------------|---------------|----------------|---------|----------|
| 1 | T38 | 521 | 517 | 4 | 0.8% | 0.8% |
| 2 | T39 | 524 | 523 | 1 | 0.2% | 0.6% |
| 3 | T40 | 516 | 526 | 10 | 1.9% | 0.9% |
| 4 | T41 | 515 | 517 | 2 | 0.4% | 0.8% |
| 5 | T42 | 526 | 518 | 8 | 1.5% | 0.9% |
| 6 | T43 | 509 | 512 | 3 | 0.6% | 0.9% |
| 7 | T44 | 521 | 520 | 1 | 0.2% | 0.8% |
| 8 | T45 | 522 | 511 | 11 | 2.1% | 0.9% |
| 9 | T46 | 514 | 524 | 10 | 1.9% | 1.0% |
| 10 | T47 | 522 | 518 | 4 | 0.8% | 1.0% |
| 11 | T48 | 507 | 522 | 15 | 3.0% | 1.2% |
| 12 | T49 | 507 | 518 | 11 | 2.2% | 1.3% |
| 13 | T50 | 508 | 521 | 13 | 2.6% | 1.3% |
| 14 | T51 | 516 | 510 | 6 | 1.2% | 1.3% |
| 15 | T52 | 515 | 522 | 7 | 1.4% | 1.3% |
| 16 | T53 | 518 | 518 | 0 | 0.0% | 1.3% |
| 17 | T54 | 520 | 522 | 2 | 0.4% | 1.2% |
| 18 | T55 | 510 | 522 | 12 | 2.4% | 1.3% |
| 19 | T56 | 514 | 513 | 1 | 0.2% | 1.2% |
| 20 | T57 | 512 | 518 | 6 | 1.2% | 1.2% |
| 21 | T58 | 527 | 511 | 16 | 3.0% | 1.3% |
| 22 | T59 | 508 | 516 | 8 | 1.6% | 1.3% |
| 23 | T60 | 507 | 513 | 6 | 1.2% | 1.3% |
| 24 | T61 | 514 | 511 | 3 | 0.6% | 1.3% |
| 25 | T62 | 509 | 518 | 9 | 1.8% | 1.3% |
| 26 | T63 | 527 | 510 | 17 | 3.2% | 1.4% |
| 27 | T64 | 510 | 515 | 5 | 1.0% | 1.4% |
| 28 | T65 | 526 | 512 | 14 | 2.7% | 1.4% |
| 29 | T66 | 509 | 511 | 2 | 0.4% | 1.4% |
| 30 | T67 | 513 | 521 | 8 | 1.6% | 1.4% |
| 31 | T68 | 523 | 516 | 7 | 1.3% | 1.4% |
| 32 | T69 | 526 | 509 | 17 | 3.2% | 1.4% |
| 33 | T70 | 521 | 511 | 10 | 1.9% | 1.4% |
| 34 | T71 | 515 | 520 | 5 | 1.0% | 1.4% |
| 35 | T72 | 512 | 516 | 4 | 0.8% | 1.4% |

Figure 2 Base Treatment Respondent Forecast Input Screen

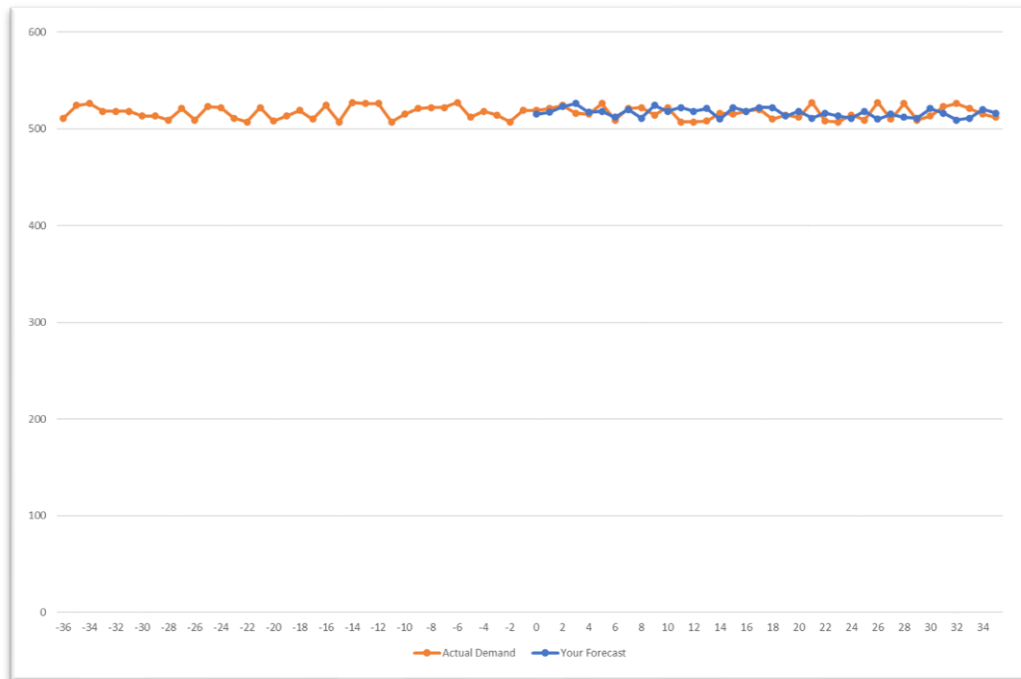


Figure 3 Line Graph of Actual Demand and Forecast

ii). Fan chart treatment: In this treatment, in addition to the standard line chart, participants were provided with a fan chart that provides a visual representation of the uncertainty associated with the forecast. The charts are updated after every time period. Fan charts typically depict the 90%, 95%, and 99% confidence intervals; we use the 95% and the 99% confidence intervals in the current research. Fan charts are a helpful tool to make the subjects aware of the uncertainty associated with the task. Kreye et al. (2012) found that depicting information in fan charts primes the subjects to think about the uncertainty associated with the task. The task remains the same, i.e., to forecast the next 36 weeks' demand one week at a time. The rolling average of the past 4-period demands is included as a reference point to indicate the evolution of the demand. A snapshot of the visual is shown below in figure 4.

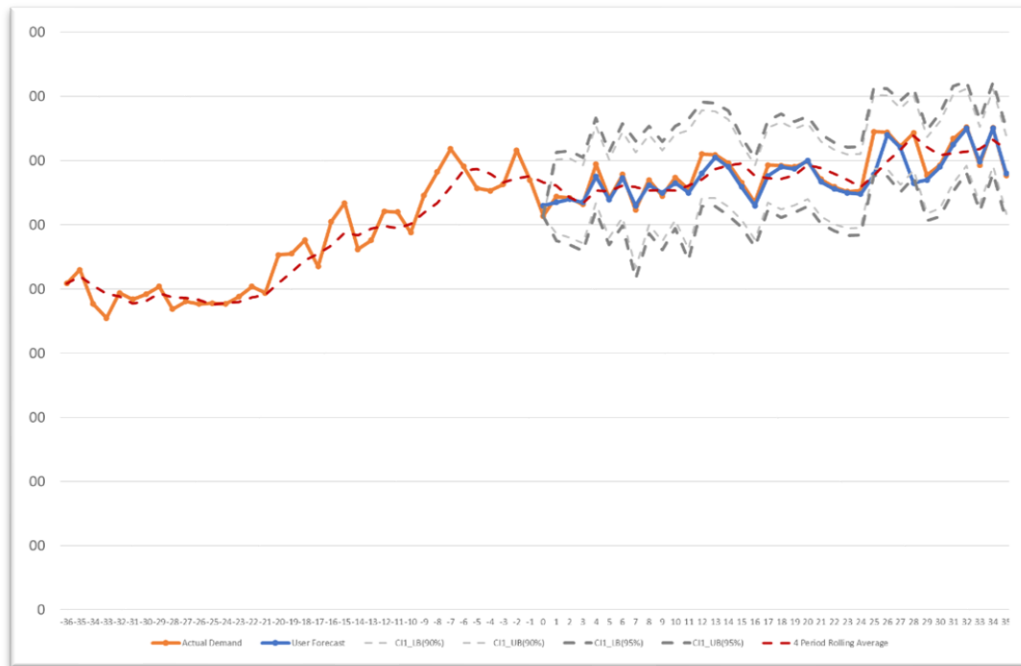


Figure 4 Fan Chart Graph

iii). Decomposition: Decomposition can take various forms (additive, multiplicative, graphical, and mathematical). Decomposition involves breaking down the time series into its components and then aggregating them. It relies on the simplification of the task. The time series components in our experiment are additive in nature and are made of the true mean μ , the change c , and the noise ϵ . By definition, the noise is to be ignored. The task for each time period effectively reduces to estimating the true level and the change. Hence for each time period, the forecaster is required to estimate the true level (past average demand) and the change in the level. The two components are then mechanically aggregated (via mathematical formula) to estimate the forecast. The formula used for aggregation is:

$$F_t = \mu_t + v_t \quad (5)$$

To familiarize the participants with the decomposition methodology, they simulate a decomposition calculation in the Qualtrics survey. In addition to the standard line charts, the graph includes a stacked bar graph that illustrates how the individual components are aggregated to make up the total demand.

| Time Period | Actual Demand | Past Average Demand | Level Change | Your Forecast (Mean + Change) | Forecast Error | APE (%) | MAPE (%) |
|-------------|---------------|---------------------|--------------|-------------------------------|----------------|---------|----------|
| T37 | 614 | 500 | 0 | 500 | 114 | 18.6% | 18.6% |
| T38 | 644 | 505 | 10 | 515 | 129 | 20.0% | 19.3% |
| T39 | 642 | 510 | 20 | 530 | 112 | 17.4% | 18.7% |
| T40 | 632 | 515 | 30 | 545 | 87 | 13.8% | 17.5% |
| T41 | 694 | 520 | 0 | 520 | 174 | 25.1% | 19.0% |
| T42 | 641 | 525 | -10 | 515 | 126 | 19.7% | 19.1% |
| T43 | 678 | 530 | -20 | 510 | 168 | 24.8% | 19.9% |
| T44 | 623 | 535 | -30 | 505 | 118 | 18.9% | 19.8% |
| T45 | 670 | 540 | 0 | 540 | 130 | 19.4% | 19.7% |
| T46 | 645 | 545 | 10 | 555 | 90 | 14.0% | 19.2% |
| T47 | 674 | 550 | 20 | 570 | 104 | 15.4% | 18.8% |
| T48 | 655 | 555 | 30 | 585 | 70 | 10.7% | 18.1% |
| T49 | 710 | 560 | 0 | 560 | 150 | 21.1% | 18.4% |
| T50 | 709 | 565 | -10 | 555 | 154 | 21.7% | 18.6% |
| T51 | 696 | 570 | -20 | 550 | 146 | 21.0% | 18.8% |
| T52 | 666 | 575 | -30 | 545 | 121 | 18.2% | 18.7% |
| T53 | 635 | 580 | 0 | 580 | 55 | 8.7% | 18.1% |
| T54 | 693 | 585 | 10 | 595 | 98 | 14.1% | 17.9% |
| T55 | 692 | 590 | 20 | 610 | 82 | 11.8% | 17.6% |
| T56 | 690 | 595 | 30 | 625 | 65 | 9.4% | 17.2% |
| T57 | 699 | 600 | 0 | 600 | 99 | 14.2% | 17.0% |
| T58 | 671 | 605 | -10 | 595 | 76 | 11.3% | 16.8% |
| T59 | 659 | 610 | -20 | 590 | 69 | 10.5% | 16.5% |
| T60 | 652 | 615 | -30 | 585 | 67 | 10.3% | 16.3% |
| T61 | 653 | 620 | 0 | 620 | 33 | 5.1% | 15.8% |
| T62 | 745 | 625 | 10 | 635 | 110 | 14.8% | 15.8% |
| T63 | 744 | 630 | 20 | 650 | 94 | 12.6% | 15.6% |
| T64 | 722 | 635 | 30 | 665 | 57 | 7.9% | 15.4% |
| T65 | 743 | 640 | 0 | 640 | 103 | 13.9% | 15.3% |
| T66 | 677 | 645 | -10 | 635 | 42 | 6.2% | 15.0% |
| T67 | 693 | 650 | -20 | 630 | 63 | 9.1% | 14.8% |
| T68 | 734 | 655 | -30 | 625 | 109 | 14.9% | 14.8% |
| T69 | 752 | 660 | 0 | 660 | 92 | 12.2% | 14.7% |
| T70 | 693 | 665 | 10 | 675 | 18 | 2.6% | 14.4% |
| T71 | 751 | 670 | 20 | 690 | 61 | 8.1% | 14.2% |
| T72 | 677 | 675 | 30 | 705 | 28 | 4.1% | 13.9% |

Figure 5 Respondent Input Screen for Decomposition

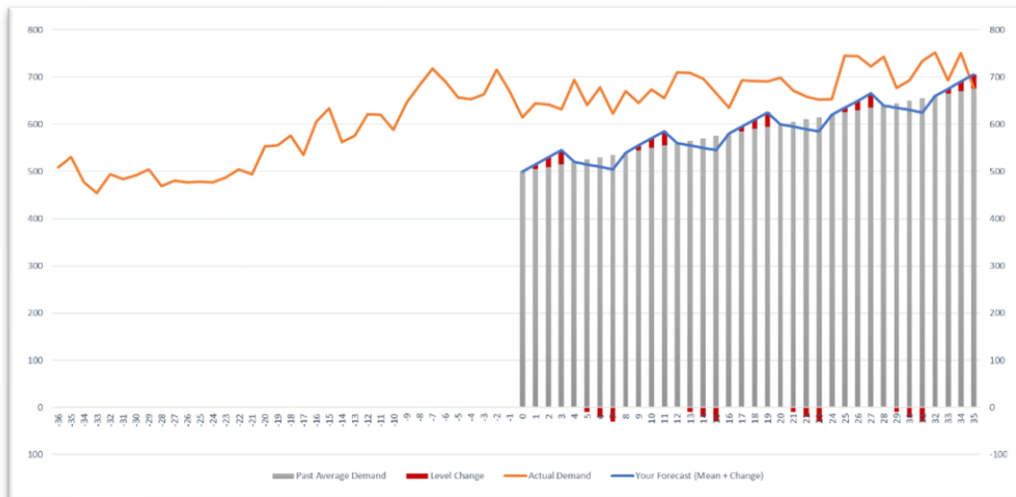


Figure 6 Graph for Decomposition Treatment

iv). Bias Feedback: Forecast bias is a directional measure of forecasting performance. The directional measure provides clear feedback if the participants are consistently under-forecasting (positive bias) or over-forecasting (negative bias). The participant forecast input screen includes a bias performance measure. In addition to the standard line graph, a bar graph that depicts the bias performance is included (Figure 7).

v). Rolling Training with Bias:

The rolling training approach provides a complete and updated record of forecasting performance at regular intervals. The experiment participants were asked to forecast for 36 periods by forecasting four periods at a time. Their actual performance for each time period was revealed after every four time periods, and they were provided the same performance metrics as the bias treatment (Figure 8).

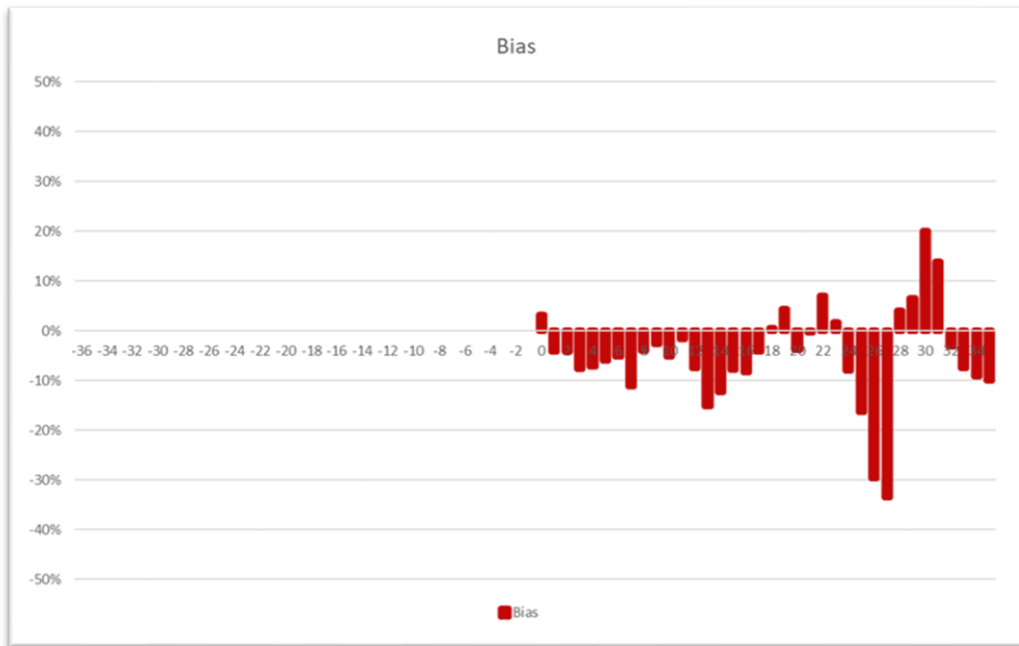


Figure 7: Bias Feedback

| Time Period | Actual Demand | User Forecast | Forecast Error | APE (%) | MAPE (%) | Bias |
|-------------|---------------|---------------|----------------|---------|----------|------|
| T37 | 715 | 692 | 23 | 3.2% | 3.2% | 3% |
| T38 | 663 | 690 | 27 | 4.1% | 3.6% | -4% |
| T39 | 672 | 700 | 28 | 4.2% | 3.8% | -4% |
| T40 | 661 | 712 | 51 | 7.7% | 4.8% | -8% |
| T41 | 658 | 705 | 47 | 7.1% | 5.3% | -7% |
| T42 | 663 | 703 | 40 | 6.0% | 5.4% | -6% |
| T43 | 664 | 698 | 34 | 5.1% | 5.4% | -5% |
| T44 | 629 | 699 | 70 | 11.1% | 6.1% | -11% |
| T45 | 598 | 622 | 24 | 4.0% | 5.8% | -4% |
| T46 | 602 | 618 | 16 | 2.7% | 5.5% | -3% |
| T47 | 585 | 615 | 30 | 5.1% | 5.5% | -5% |
| T48 | 595 | 605 | 10 | 1.7% | 5.2% | -2% |
| T49 | 556 | 598 | 42 | 7.6% | 5.4% | -8% |
| T50 | 523 | 602 | 79 | 15.1% | 6.1% | -15% |
| T51 | 525 | 590 | 65 | 12.4% | 6.5% | -12% |
| T52 | 543 | 585 | 42 | 7.7% | 6.6% | -8% |
| T53 | 508 | 550 | 42 | 8.3% | 6.7% | -8% |
| T54 | 499 | 520 | 21 | 4.2% | 6.5% | -4% |
| T55 | 508 | 505 | 3 | 0.6% | 6.2% | 1% |
| T56 | 538 | 515 | 23 | 4.3% | 6.1% | 4% |
| T57 | 525 | 545 | 20 | 3.8% | 6.0% | -4% |
| T58 | 553 | 555 | 2 | 0.4% | 5.7% | 0% |
| T59 | 591 | 550 | 41 | 6.9% | 5.8% | 7% |
| T60 | 570 | 560 | 10 | 1.8% | 5.6% | 2% |
| T61 | 537 | 580 | 43 | 8.0% | 5.7% | -8% |
| T62 | 503 | 585 | 82 | 16.3% | 6.1% | -16% |
| T63 | 463 | 600 | 137 | 29.6% | 7.0% | -30% |
| T64 | 451 | 602 | 151 | 33.5% | 7.9% | -33% |
| T65 | 469 | 450 | 19 | 4.1% | 7.8% | 4% |
| T66 | 503 | 470 | 33 | 6.6% | 7.8% | 7% |
| T67 | 538 | 430 | 108 | 20.1% | 8.2% | 20% |
| T68 | 493 | 425 | 68 | 13.8% | 8.3% | 14% |
| T69 | 485 | 500 | 15 | 3.1% | 8.2% | -3% |
| T70 | 479 | 515 | 36 | 7.5% | 8.2% | -8% |
| T71 | 481 | 525 | 44 | 9.1% | 8.2% | -9% |
| T72 | 482 | 530 | 48 | 10.0% | 8.2% | -10% |

Figure 8 – Subject Input Screen for Rolling Training Approach

4.2 Conducting the Experiment and Data Collation

As described earlier, the study was activated in MTurk. The average completion based on pre-tests was twenty-five minutes. The participant's actual average completion time varied between twenty and thirty-six minutes, with an overall average completion time of twenty-seven minutes. The participants in the decomposition treatments took the longest time, averaging thirty-three minutes. The completion times are reported based on tracking by Mturk and are calculated from when a participant signs up for the study to when he/she indicates the completion by keying a completion code in MTurk. Further breakdown, such as actual time spent working on the study versus idle time, is unavailable.

The completed responses are reviewed to discard blank responses, i.e., participants who signed up for the study but failed to complete the study and upload a blank spreadsheet. The individual spreadsheet responses are combined to create a complete data set using the extraction, transformation, and loading capabilities of the Alteryx software. We now proceed with the description of the empirical analysis.

5. Empirical Analysis

The consolidated forecasts are reviewed for typographical errors, primarily because participants cannot change the forecast once submitted. We replicate the methodology followed by Kremer et al. and review all forecasts with an absolute error greater than 300. Forecasts, where the typographical errors can be easily identified, are corrected (e.g., a forecast of 61 in a long series of forecasts in the range of 450 is codified as 461). However, when a determination cannot be made but the forecast is identified as a typographical error, it is coded as a missing value. Such adjustments are negligible, making up 0.3% of the observations. Additional outliers in the forecast and the forecast error data are winsorized based on each treatment, condition, and demand set within the condition.

Table 4 below shows the number of participants in each treatment, condition, and demand set within a condition. The average number of participants per study was 52. The number of participants ranges from 38 to 66. The variation in the number of participants is primarily due to the difference in the number signing up for a particular study and the number of qualified responses. The target for each treatment condition was to gather 45 responses per condition.

| Treatment Name | Change | Noise | D1 | D2 | D3 | D4 | Total |
|----------------|--------|-------|-------------|------------|------------|------------|--------------|
| Base | Low | Low | 12 (432) | 6 (216) | 10 (360) | 10 (360) | 38 (1368) |
| Base | Low | High | 22 (792) | 5 (180) | 15 (540) | 12 (432) | 54 (1944) |
| Base | High | Low | 15 (540) | 13 (468) | 15 (540) | 11 (396) | 54 (1944) |
| Base | High | High | 14 (504) | 13 (468) | 13 (468) | 17 (612) | 57 (2052) |
| Fan | Low | Low | 6 (216) | 10 (360) | 13 (468) | 13 (468) | 42 (1512) |
| Fan | Low | High | 10 (360) | 15 (540) | 8 (288) | 13 (468) | 46 (1656) |
| Fan | High | Low | 15 (540) | 12 (432) | 8 (288) | 10 (360) | 45 (1620) |
| Fan | High | High | 15 (540) | 22 (792) | 9 (324) | 12 (432) | 58 (2088) |
| Decomposition | Low | Low | 13 (468) | 16 (576) | 14 (504) | 12 (432) | 55 (1980) |
| Decomposition | Low | High | 20 (720) | 13 (468) | 13 (468) | 20 (720) | 66 (2376) |
| Decomposition | High | Low | 12 (432) | 16 (576) | 21 (756) | 6 (216) | 55 (1980) |
| Decomposition | High | High | 12 (432) | 18 (648) | 12 (432) | 18 (648) | 60 (2160) |
| Bias | Low | Low | 8 (288) | 15 (540) | 12 (432) | 11 (396) | 46 (1656) |
| Bias | Low | High | 21 (756) | 6 (216) | 25 (900) | 12 (432) | 64 (2304) |
| Bias | High | Low | 14 (504) | 15 (540) | 11 (396) | 15 (540) | 55 (1980) |
| Bias | High | High | 17 (612) | 17 (612) | 12 (432) | 15 (540) | 61 (2196) |
| BRT | Low | Low | 8 (288) | 10 (360) | 14 (504) | 13 (468) | 45 (1620) |
| BRT | Low | High | 12 (432) | 16 (576) | 17 (612) | 12 (432) | 57 (2052) |
| BRT | High | Low | 14 (504) | 12 (432) | 11 (396) | 6 (216) | 43 (1548) |
| BRT | High | High | 19 (684) | 17 (612) | 11 (396) | 10 (360) | 57 (2052) |
| Total | | | 279 (10044) | 267 (9612) | 264 (9504) | 248 (8928) | 1058 (38088) |

Table 4 Number of Participants Per Treatment Condition and Demand Set (number in parenthesis is the number of observations)

5.1 Variable Definition

Before we explain the empirical analysis, a description of the variables used in the analysis is in order. We estimate an exponential smoothing model to compare the observed and optimal forecasts. For clarity, we define the following:

- i) D_t denotes the actual demand for time period T , which is revealed after the subject submits the forecast
- ii) F_{sit} denotes the forecast made by an individual I for a time period T , for a given treatment condition and a given demand set S
- iii) F_t^* denotes the normative optimal forecast and is calculated as

$$F_t^* = F_{t-1}^* + \alpha^*(D_{t-1} - F_{t-1}^*) \quad (6)$$

- iv) $MAE(D_t, F_{it})$ denotes the mean absolute forecast error, which is the average forecast error of all subjects within a given treatment condition for all the time periods
- v) $MAE(D_t, F_t^*)$ is the optimal forecast error and is defined as

$$E_{st}^* = |F_{st}^* - D_{st}| \quad (7)$$

- vi) $MAPE$ refers to the mean of the Absolute percentage error over the given time periods.

5.1.1 Dependent Variables

- i) α_{its} is the adjustment score or the individual-level alpha and is calculated as

$$\alpha_{its} = (F_{sit} - F_{sit-1}) / (D_{st-1} - F_{sit-1}) \quad (8a)$$

Note that the calculation of the adjustment score is modified for the rolling training approach. In the case of the rolling treatment, since the participant observes errors only after four time periods, we modified the calculation of alpha based on the total demand and total forecast of the past four periods. Given that there are 36 time periods, there are 9 buckets.

$$\alpha_{its} = \sum_{t=1}^{t=4} (F_{sit} - F_{sit-1}) / \sum_{t=1}^{t=4} (D_{st-1} - F_{sit-1}) \quad (8b)$$

5.1.2 Independent Variables

- i) **Indicator** (dummy) variable representing each treatment condition
- ii) **Indicator** (dummy) variable representing the rolling training bucket to account for any clustering effect.

5.1.3 Control Variables

- i) E_{sit} denotes the absolute forecast error defined as per equation 9 below. In the regression analysis, we divided the forecast error by 100 to apply a scaling factor.

$$E_{it} = \frac{|F_{sit} - D_{st}|}{100} \quad (9a)$$

Note that for the rolling training approach, we average the forecast error across the four time periods of each bucket.

$$E_{it} = \sum_{t=1}^{t=4} \frac{|F_{sit} - D_{st}|}{100} \quad (9b)$$

- ii) **Graph usage:** As part of the treatments, we provided extra information in graphical format in each review period. We quizzed the participants on how frequently they used these charts through a 5-point Likert scale, ranging from never (1) to always (5). We code graph usage as a factor variable.
- iii) **Nature of Demand:** Participants are surveyed on their perception nature of the demand function, i.e., i) unpredictable, ii) Seasonal, iii) Stable, iv) Stable with noise, and v) Trended. In our analysis, we group stable and stable with noise and code each category as a dummy variable.
- iv) **Performance metrics** review is a 5-point Likert scale variable measure of how frequently the participants reviewed the performance metrics ranging from never (1) to always (5). The performance metrics shared with them include Forecast Errors, APE, and MAPE. Performance metrics are coded as a factor variable.
- v) **Attention check response:** Concerns exist around Mturk participants' attention to the task. We have two questions designed to verify that the participants read instructions carefully. The response is coded as a binary variable, taking 1 for a correct response and 0 for an incorrect response.
- vi) **Fan chart knowledge:** We query the participants' understanding of fan-chart charts. They are presented with a fac-chart and are asked two questions, one asking them to

identify the upper bound of the 95% confidence and the other asking them about the lower bound of the 95% confidence interval.

Stata allows us to evaluate the impact of each level of the factor variable on our prediction.

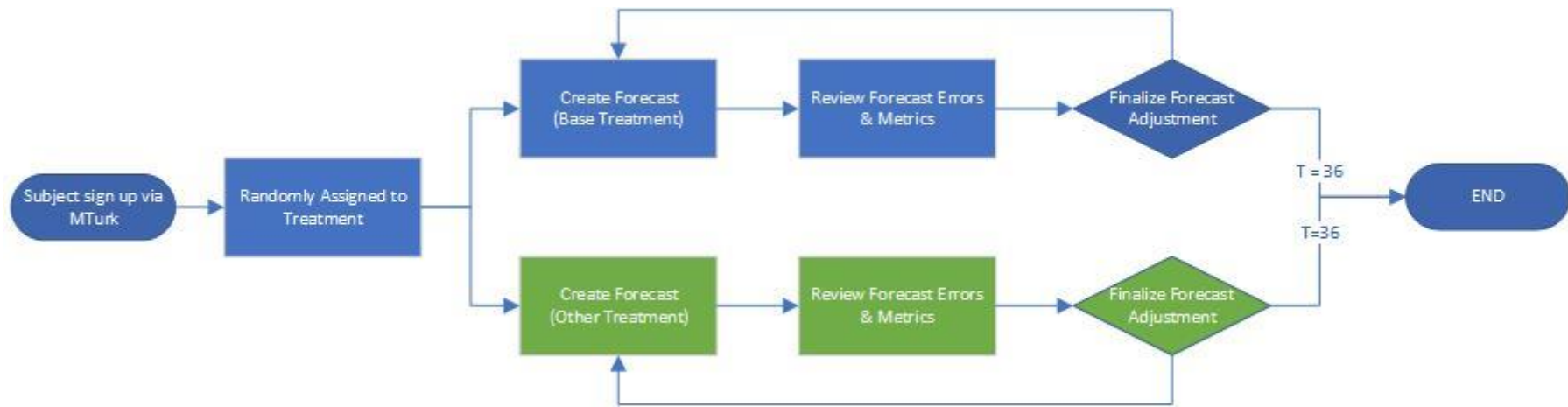


Figure 9 – Forecasting Process

5.2 Initial Analysis:

Our empirical analysis is centered around three critical steps of the forecasting process, i) review of the forecast errors, ii) decide the adjustment amount and direction, and iii) create a forecast which minimizes the MAPE. Depending on the treatment, the information made available to the participant varies, but the process never changes. Figure 9 illustrates the process. Except for the first-period forecast, the subjects perform all three steps as they receive information about their previous period's forecast performance in terms of forecast errors and MAPE. After this, they decide on the adjustment for the current period's forecast, following which they create and submit their forecast. The participants are incentivized and asked to minimize the MAPE associated with their forecast. This process continues for a total of 36 time periods.

We build on Kremer et al. (2011), which compares the forecasters' errors to the errors obtained under an exponential smoothing model with the optimal adjustment factor given the demand noise and change conditions. They attribute the lower performance of the subjects to either over- or underreaction to forecasting errors, as measured by the adjustment factor in their forecasts. The primary reason for such a response is that the forecast errors are more salient than the system parameters. We developed hypotheses intended to decrease or increase the forecasters' adjustment factor by reducing errors' salience and/or reducing system neglect by emphasizing the demand process.

The empirical analysis thus begins by replicating Kremer et al.'s (2011) findings. We subsequently test the hypotheses and discuss the forecasting performance in the various treatment conditions.

5.2.1 Base Treatment - Mean Absolute Errors

To replicate Kremer et al.'s findings about the forecast errors, we performed *t*-tests to compare the participant's observed MAE to the optimal MAE based on the normative forecast. The normative model (equation 2(b), reproduced below) requires subjects to consider both the forecast error's strength and weight. The strength is the magnitude of the forecast error, and the weight individuals assign to the forecast error is denoted by alpha. In the normative model, the optimal weight is uniquely determined by the system parameters *c* and *n*.

$$F_{t+1} = F_t + \alpha(D_t - F_t) \quad \mathbf{2(b)}$$

Table 5 shows the MAE levels for the base treatment compared to the optimal MAE. In line with Kremer et al. findings, forecasting performance deteriorates with increases in noise (*n*) and change (*c*) levels.

| | Base | |
|------------|------------------|------------------|
| | N10 | N40 |
| C0 | 8.83*** (6.08) | 24.25*** (28.87) |
| C40 | 44.82*** (19.06) | 48.16*** (26.94) |

Table 5 Two Tail T-Test Comparison of Optimal vs. Observed MAE

Note: The numbers in parenthesis show optimal MAE based on a single exponential smoothing model.

5.2.2 Adjustment Scores (α)

The adjustment score (α) or the weight assigned to the forecast error is undefined for the first period and coded as a missing value in our empirical analysis. We evaluate the adjustment scores for each forecast created by the subject. Table 6 below classifies the observed adjustment scores into different buckets.

| α_{it} | No of Obs | % of Observation |
|--------------------|-----------|------------------|
| $(-\infty, 0)$ | 8,957 | 24% |
| 0 | 2,434 | 6% |
| $(0, 1)$ | 14,526 | 38% |
| $(1, \infty)$ | 12,171 | 32% |
| Total Observations | 38,088 | |

Table 6 Subject Adjustment Scores Analysis by Observation

A negative adjustment score indicates that the participants adjusted their forecast in the opposite direction of their forecast error (24% of the observations). An adjustment score of zero (6% of the observation) indicated no reaction from the participants (i.e., the current forecast is the same as the previous forecast). Adjustment scores greater than zero and less than one demonstrate exponential smoothing behavior (38%). Finally, adjustment scores above one indicate that the participants are projecting a trend into the future.

Supplementary analysis of the answers to a survey question on the nature of demand reveals that about half of the participants (47%) identified the time series as either stable or stable with noise, whereas

21% of the respondents identified the time series as a trended one (Figure 10).

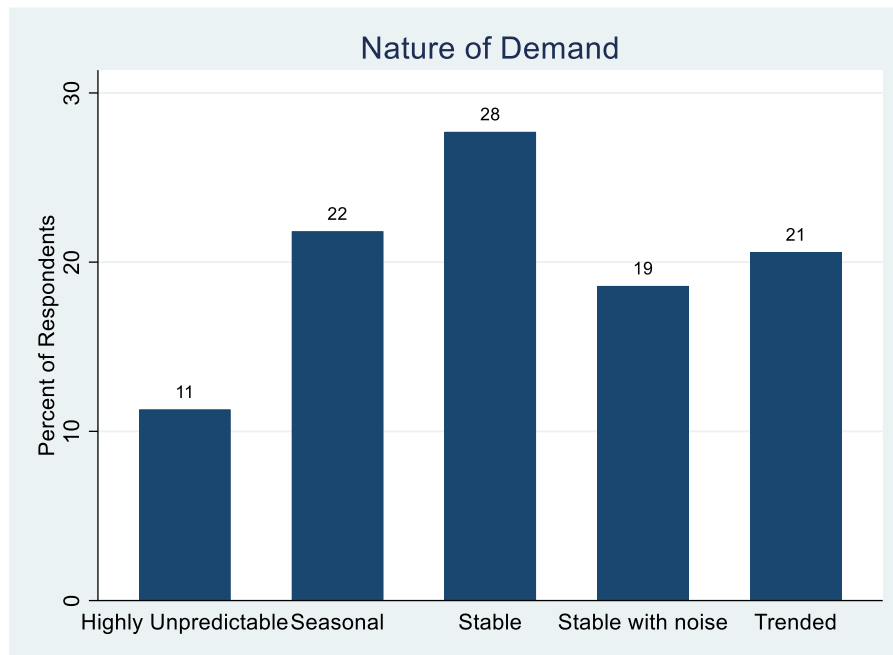


Figure 10 Participant Response to the Nature of Demand

To check if participants can effectively distinguish between noise and change, we reviewed how the participant's view of the nature of demand as stable or stable with noise varies based on the demand condition. Table 7 below shows the breakdown. We find that as noise and change level increase, the participant's perception of the demand environment as either stable or stable with noise decreases.

| | Nature of Demand (Stable + Stable With Noise) | |
|------------|--|-------|
| | N10 | N40 |
| C0 | 58.7% | 45.8% |
| C40 | 41.2% | 41.6% |

Table 7: Proportion of participants perceiving demand as stable based on demand condition.

We check whether the participants' perception of the demand impacts their adjustment scores. We use a nested regression to account for individual-level and demand set-level characteristics. Table 8a shows that the adjustment scores (α) of individuals who identified the time series as stable or stable with noise are significantly lower. In contrast, individuals who identified the time series as unpredictable have higher adjustment scores (α). The regression results are shown in Table 8b.

This preliminary analysis indicates that people find distinguishing between change and noise difficult. Hence, they only correctly identify stable time series in low noise conditions. Furthermore, we find that the participant's perception of the time series will affect their adjustment factors. If they find the series stable, they choose a lower adjustment factor than when they do not.

We included the absolute forecast error as a control. Note that the forecast error in the exponential smoothing model should not be a predictor of the adjustment score, as the optimal adjustment score is exclusively determined by the change and noise level. However, we observe that in our results, the absolute forecast errors are significant across all conditions and negatively associated with adjustment scores.

| | Model 1 Condition 1 (C0N10) | Model 2 Condition 2 (C0N40) | Model 3 Condition 3 (C40N10) | Model 4 Condition 4 (C40N10) |
|-------------------------|---|---|--|--|
| | α | α | α | α |
| Unpredictable | 0.40 | 0.35 | 0.45 | 0.52 |
| Stable & Stable w Noise | 0.26 | 0.34 | 0.31 | 0.29 |
| Seasonal & Trended | 0.38 | 0.48 | 0.52 | 0.48 |

Table 8a Adjustment Scores based on Nature of Demand

| | Model 1 Condition 1 (C0N10) | Model 2 Condition 2 (C0N40) | Model 3 Condition 3 (C40N10) | Model 4 Condition 4 (C40N10) |
|-------------------------|---|---|--|--|
| | Coefficient | Coefficient | Coefficient | Coefficient |
| Abs. Forecast Error | -0.42*** (0.03) | -0.26*** (0.02) | -0.33*** (0.01) | -0.27*** (0.01) |
| Unpredictable | 0.01 (0.06) | -0.11*** (0.05) | 0.1* (0.06) | -0.01 (0.05) |
| Stable & Stable w Noise | -0.07** (0.03) | -0.09*** (0.03) | -0.09*** (0.04) | -0.14*** (0.04) |
| Seasonal & Trended | 0 | 0 | 0 | 0 |
| Constant | 0.34*** (0.14) | 0.15 (0.42) | 0.35** (0.17) | 0.84*** (0.21) |

Table 8b Estimation Results Based on Nature of Demand

5.3 Multilevel Nested Models

We structured the experiment as a multilevel nested model. The individual participants are nested within one out of four possible demand sets, which are nested within a condition. There are four conditions nested within each treatment. Figure 11 below shows a pictorial representation of the nested structure.

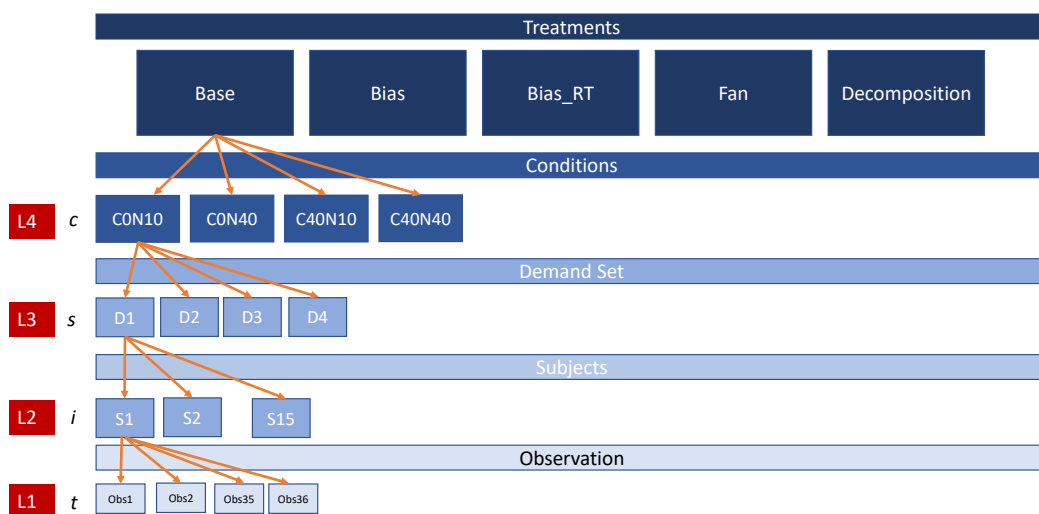


Figure 11: Nesting and Multilevel Structure of the Study

Standard ordinary least square regression can produce misleading results about the statistical significance of a relationship when it is used to analyze nested data sets. A nested structure leads to greater statistical dependency in the data (subjects within a demand set, within a condition, and in a treatment), which can cause the standard errors to be underestimated and inflate the statistical significance (O'Dwyer et al. 2014). Standard OLS modeling assumes each observation is independent and not co-related with other observations in the sample.

The clustering that naturally occurs in the case of nested models makes meeting this assumption challenging. Hence nested modeling techniques are needed to address the statistical dependency. In standard OLS regression, group characteristics are not considered effectively at the individual level. Nested models produce unbiased estimates of the standard errors of the regression coefficients. Nested modeling also allows group variables (i.e., Treatment, Demand Conditions, etc.) to explain individual-level outcomes better. The statistical test to check for the necessity to use a nested model is based on the interclass correlation coefficient (ICC), which measures the degree of statistical dependence in the data. The variance of a dependent variable Y_{ij} can be split into the within-the-group variance (σ^2) and between-group variance (τ_0^2).

$$ICC = \rho = \frac{\tau_0^2}{\sigma^2 + \tau_0^2} \quad (10)$$

To quote O'Dwyer et al. (2014), it is "*The intraclass correlation coefficient (ICC) in equation (10) is used to calculate the portion of the variance in the dependent variable that is explained at each level in subsequent models with the addition of individual and group measures*". The interclass correlation coefficient ranges from 0 to 1, with higher values signifying higher levels of statistical dependency.

To confirm whether our data exhibits a nested structure and requires a multilevel modeling approach, we fit a null model that only includes the dependent variable and the variables identifying the hierarchical structure. We used Stata's "mixed" command to estimate the model

using the maximum likelihood estimates option. Convention suggests ICC values greater than 0.05 indicate clustering, which needs to be controlled. Checking for the interclass correlation in our null model shows statistical dependency at the individual and demand set levels (Table 9). Consequently, all our analyses will be performed using the nested modeling approach.

| Interclass correlation | | | | |
|-------------------------------|-------|----------------|-------------------------|-------|
| Level | ICC | Standard Error | 95% Confidence Interval | |
| Treatment | 0.041 | 0.266 | 0.011 | 0.139 |
| Condition | 0.046 | 0.266 | 0.014 | 0.136 |
| Demand Set | 0.053 | 0.266 | 0.016 | 0.136 |
| Individual | 0.253 | 0.022 | 0.211 | 0.298 |

Table 9 Interclass Estimates for Null Model

Dummy Variables are generated to denote each treatment and condition combination uniquely. We estimate the model for a given condition across all treatments based on our nesting structure. Our approach yields a model each for the four different conditions. Within each model, the base treatment is defined as a reference category. The individual adjustment score is the dependent variable, and the treatments are the explanatory variables. The absolute forecast error, the usage levels of graphs, the performance metrics review frequency, and the attention check question responses are also included as control variables.

We fit a random intercept model for the four models. Models 1 and 2 represent stable conditions with no change, where the original Kremer et al. work observed overreaction, i.e., the observed adjustment factor

was greater than the optimal adjustment factor of zero. Models 3 and 4 are the conditions with high change, where underreaction was observed, i.e., the observed adjustment factor was smaller than the optimal adjustment factor. The base treatment replicates the original Kremer et al. (2011) study. Our research objective is to mitigate the effects of underreaction and overreaction in order to help improve forecast accuracy. We aim to achieve this by using various treatments that prior research has proven effective in improving forecast accuracy in other types of time series demand forecasting models (e.g., with seasonality, trend, etc.).

5.4 Hypothesis Testing: In the base treatment, the participants are asked to forecast one time period at a time and are presented with simple line graphs of actuals and forecasts. We are able to reproduce the findings from Kremer's original study and find evidence of overreaction in stable conditions and underreaction in unstable situations (compared with the optimal response).

| | Base | |
|------------|----------------|----------------|
| | N10 | N40 |
| C0 | 0.40*** (0) | 0.46*** (0) |
| C40 | 0.55*** (0.94) | 0.54*** (0.62) |

Table 10 Wald test of Optimal Alpha vs. Subject Adj Score

Note: The number in parenthesis is the Optimal Alpha for each condition

We now discuss the results from testing each of the hypotheses associated with the treatments.

5.4.1 Fan Charts:

Overreaction:

H1a: *Use of visual fan charts in the judgmental forecasting process reduces the underreaction*

In earlier sections, we argued that underscoring the uncertainty (noise) associated with the forecasting process will aid in reducing the salience of the forecast errors, which is desirable in stable scenarios. Fan charts are well suited to highlight that uncertainty; hence, we predicted fan charts would help reduce overreaction.

| | Fan | |
|------------|-------------|---------------|
| | N10 | N40 |
| C0 | 0.44 (0.40) | 0.62** (0.46) |
| C40 | 0.55 (0.56) | 0.55 (0.54) |

Table 11 Wald test of Treatment vs. Base Adjustment Score

Note: The number in parenthesis is the Adjustment Score (α) for the base condition

Compared to the base treatment, our results show that providing the subjects with fan charts causes the overreaction to increase in both stable conditions. Furthermore, this increase is significant in stable conditions with high noise. The fan charts did not have the desired effect of reducing the salience of forecast errors.

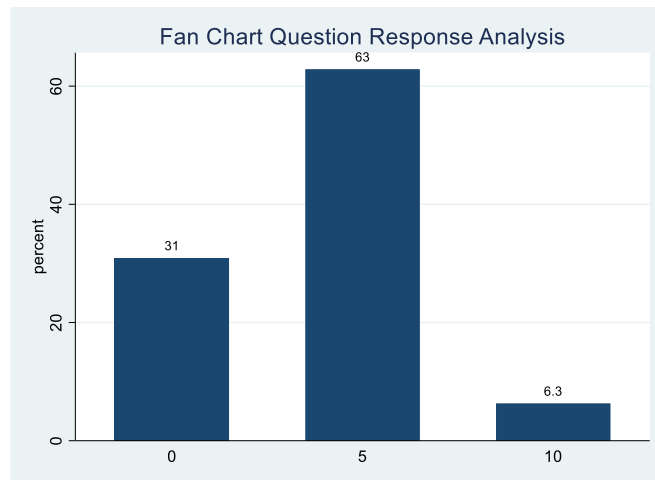


Figure 12 Fan Chart Questionnaire Response Analysis

To better understand the lack of positive impact of fan charts on overreaction, we reviewed the participant's responses to questions probing their understanding of the confidence interval's lower and upper bound. Their responses show that only about 6.3% of the participants could answer both questions correctly, 63% were able to answer one question, and 31% of the respondents could not answer either of the questions (see Figure 12). Consequently, we suspect that the subjects may not have fully appreciated the concept of fan charts and could not apply them effectively in this exercise and realize that, at least in stable conditions, the variations in demand are noise and should be ignored. We categorize participants based on their answers to the understanding questions and add this as a control to the regression. However, the results in Table 12 highlight that even participants with a full understanding of fan charts do not significantly differ in their adjustment scores (α) from the participants with partial knowledge.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|-----------------|------------------------|------------------------|-------------------------|-------------------------|
| | Condition 1 (C0N10) | Condition 2 (C0N40) | Condition 3 (C40N10) | Condition 4 (C40N40) |
| | Coefficient | Coefficient | Coefficient | Coefficient |
| FC Know. Leve 1 | 0.09 (0.08) | 0.01 (0.09) | 0.10 (0.08) | -0.04 (0.08) |
| FC Know. Leve 2 | -0.08 (0.14) | -0.16 (0.17) | -0.11 (0.17) | -0.21 (0.20) |
| FC Treatment | 0.02 (0.08) | 0.19** (0.08) | -0.04 (0.08) | 0.02 (0.08) |
| Constant | 0.45*** (0.15) | 0.03 (0.20) | 0.33** (0.16) | 0.76*** (0.20) |

Table 12 Nested Models Fan Charts Estimation

Note: Controls included in the above estimation

Looking at the output of Models 1 and 2 in Table 12, fan charts contributed to an increase of 0.02 and 0.19 to the adjustment scores, respectively, when compared to the base treatment. The impact was significant only in condition 2.

Underreaction:

H1b-i): *Use of visual fan charts in the judgmental forecasting process reduces the underreaction*

H1b-ii) *Use of visual fan charts in the judgmental forecasting process increases the underreaction*

We argued in our earlier sections that fan charts could reduce the salience of errors. In contrast, in conditions with a high change-to-noise ratio (high values of W), forecast errors must influence future forecasts more (reduce system neglect). We were unsure of the directional impact of fan charts on underreaction. Hence, we stated our hypothesis related to underreaction as a dual one. The results from our analysis show that

there is no significant difference in adjustment scores between the base and the treatment conditions. The fan chart coefficients from the nested models also present the same picture (Table 12).

| | Model 1 | Model 2 | Model 3 | Model 4 |
|-------------------------|--------------------------------|--------------------------------|---------------------------------|---------------------------------|
| | Condition 1 (C0N10) | Condition 2 (C0N40) | Condition 3 (C40N10) | Condition 4 (C40N40) |
| | Coefficient | Coefficient | Coefficient | Coefficient |
| Abs. Forecast Error | -0.45*** (0.04) | -0.27*** (0.02) | -0.34*** (0.02) | -0.28*** (0.01) |
| Fan | 0.02 (0.08) | 0.19** (0.08) | -0.04 (0.08) | 0.02 (0.08) |
| Decomposition | -0.09* (0.05) | -0.11** (0.05) | -0.14*** (0.06) | -0.26*** (0.06) |
| Bias | 0.06 (0.05) | -0.01 (0.05) | 0.04 (0.06) | -0.05 (0.06) |
| Bias w Rolling Training | 0.01 (0.05) | -0.08 (0.05) | -0.16** (0.01) | -0.08 (0.06) |
| Constant | 0.45*** (0.15) | 0.03 (0.20) | 0.33** (0.16) | 0.76*** (0.20) |

Table 13 Consolidated Nested Models Estimation Results

Note: The Base Treatment is used as the reference category. Numbers in bold indicate the value of the coefficient. Numbers in parentheses indicate the standard errors. Controls included in the above estimation

5.4.2 Decomposition

H2: *Decomposition of the time series into individual components reduces overreaction and underreaction*

Participants were provided information on the nature of the time series for all conditions. Subjects were informed that the time series consisted of two components: i) noise and ii) change. They were not explicitly informed about the actual noise and change levels. In the decomposition treatment, participants were asked to separately estimate the past average demand and the change in demand. Their estimates were then aggregated mechanically to derive the total forecast. We expected that the mechanism of estimating the components would emphasize the demand generation process. Understanding the process should help in both overreaction and underreaction situations by reducing errors' salience and system neglect.

| | Decomposition | |
|------------|----------------------|----------------|
| | N10 | N40 |
| C0 | 0.14* (0.40) | 0.25** (0.46) |
| C40 | 0.28*** (0.56) | 0.20*** (0.54) |

Table 14 Wald test of Treatment vs. Base Adjustment Score

Note: The number in parenthesis is the Alpha for the base condition

Compared to the base treatment, our results (Table 14) show that getting subjects to think and estimate the demand components explicitly reduced their adjustment factor across all the conditions. While we

achieved the desired response in condition 1 and condition 2, in conditions 3 and 4, the participants' behavior was the opposite of our prediction. The nested model estimation results in Table 15 confirm this.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---------------|--------------------------------|--------------------------------|---------------------------------|---------------------------------|
| | Condition 1 (C0N10) | Condition 2 (C0N40) | Condition 3 (C40N10) | Condition 4 (C40N40) |
| | Coefficient | Coefficient | Coefficient | Coefficient |
| Decomposition | -0.09* (0.05) | -0.11** (0.05) | -0.14*** (0.06) | -0.26*** (0.06) |
| Constant | 0.45*** (0.15) | 0.03 (0.20) | 0.33** (0.16) | 0.76*** (0.20) |

Table 15 Nested Models Decomposition Estimation

Note: Controls included in the above estimation

Thus, decomposition reduced overreaction in stable conditions but also further worsened underreaction in unstable conditions. Our hypothesis is only partially supported.

To understand the strong dampening effect of decomposition on the participants' adjustment scores across all conditions, we investigated how the total forecasts (algebraic sum of average demand and change) of the subjects in the decomposition treatment differ from the forecasts of the base treatment. Compared to the base treatment, the total forecast of the subjects was consistently higher across all the conditions in the decomposition treatment. Table 16 below shows a t-test comparison of the subject's final forecast in the decomposition treatment against the base treatment.

| | Forecast (Deco vs. Base) | |
|------------|---------------------------------|------------------|
| | N10 | N40 |
| C0 | 558.7*** (517.3) | 567.3*** (518.4) |
| C40 | 551.3*** (533.4) | 562.6*** (505.3) |

Table 16 T-test Comparison of Participants Forecast with Decomposition Treatment and Base.

Note: The number in parenthesis indicates the forecast from the base condition.

To better understand the cause of the higher average demand estimation in the decomposition treatment, we plotted the average change, demand, and forecast values for each period in the decomposition treatment and the forecast from the base treatment in **Figure 13**. Triangular markers represent the demand forecast in the base treatment, whereas circular markers represent the decomposition treatment. In the decomposition treatment, the green line represents the average demand estimation, the orange line represents the change estimation, and the maroon line represents the forecast.

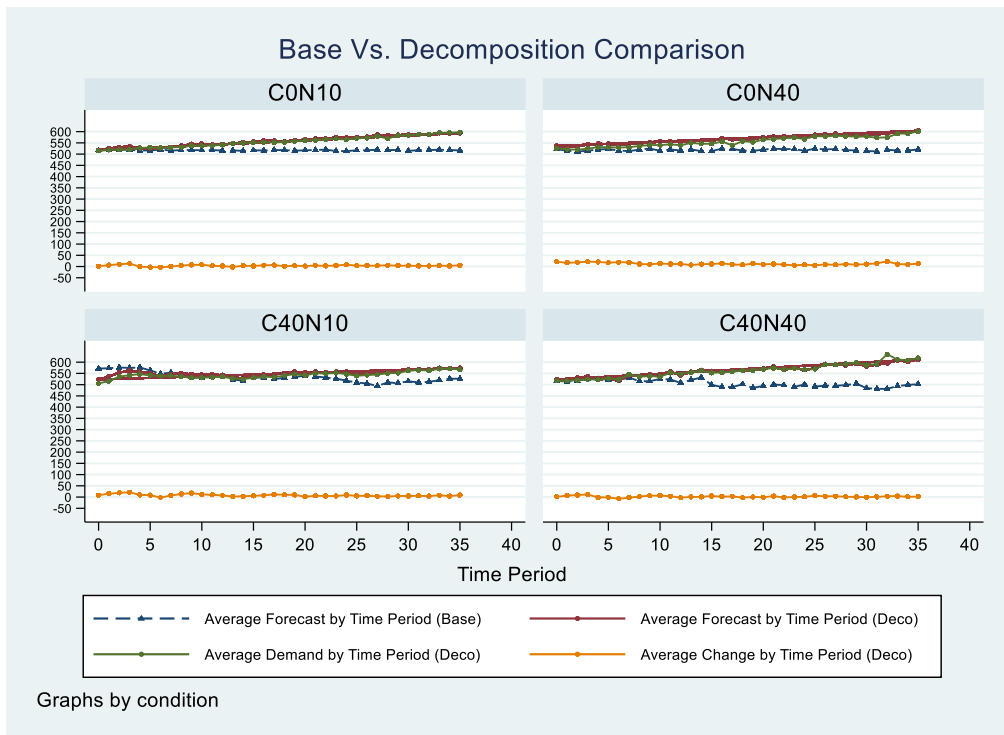


Figure 13 Comparison of Base vs. Decomposition

A visual inspection reveals a key factor that ultimately affects the adjustment score. The visible upwards trend in demand values entered by the participants combined with the average change implies that participants implicitly impute a trend in the demand series. Given that none of the underlying demand conditions exhibited a trend, the lack of adjustment to the resulting consistent over-forecasting resulted in a low adjustment score across all the conditions.

5.4.3 Bias Treatment

Bias is a directional measure of forecast accuracy. A positive bias indicates a tendency to under-forecast, while a negative bias denotes an over-forecast. Bias is classified as performance feedback since it conveys more information than outcome feedback, where the subjects are only informed about the actual demand. We predicted that bias

feedback would help alleviate overreaction. In case of unstable conditions, given that the average demand itself changes from period to period, we did not offer a prediction on the direction of the impact of the bias feedback. Hence we put forward a dual-part hypothesis about underreaction.

H3a: *Providing bias feedback reduces overreaction*

H3b-i): *Providing bias feedback in the judgmental forecasting process reduces the underreaction*

H3b-ii): *Providing bias feedback in the judgmental forecasting process increases the underreaction*

| | Bias | |
|------------|-------------|-------------|
| | N10 | N40 |
| C0 | 0.44 (0.40) | 0.45 (0.46) |
| C40 | 0.56 (0.56) | 0.42 (0.54) |

Table 17 Wald test of Treatment vs. Base Adjustment Score

Note: The number in parenthesis is the Alpha for the base condition

Table 17 compares the bias treatment condition's adjustment scores (α) values with the base condition. We see that bias feedback did not significantly impact the adjustment scores (α) across the conditions. The subject adjustment scores in the bias and base treatment are similar in all conditions. Table 18 shows the results of the nested model estimation that show no significant impact from the bias feedback on the adjustment scores.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|----------|------------------------|------------------------|-------------------------|-------------------------|
| | Condition 1 (C0N10) | Condition 2 (C0N40) | Condition 3 (C40N10) | Condition 4 (C40N40) |
| | Coefficient | Coefficient | Coefficient | Coefficient |
| Bias | 0.06 (0.05) | -0.01 (0.05) | 0.04 (0.06) | -0.05 (0.06) |
| Constant | 0.45*** (0.15) | 0.03 (0.20) | 0.33** (0.16) | 0.76*** (0.20) |

Table 18 Nested Models Bias Estimation

Note: Controls included in the above estimation

Bias as a performance measure should help subjects identify consistent over-forecasting and under-forecasting compared to the base treatment. Given that there was no significant difference between the base and the bias treatment, we proceeded to look at instances of over-forecasting and under-forecasting in both the bias and the base treatments.

We defined over-forecasting and under-forecasting as any four consecutive observations with a positive bias or negative bias. Table 19 below shows the classification based on the above definition. We can make a few critical observations from the table below. Firstly, the number of people who do not suffer from persistent forecasting bias –positive or negative – is relatively high at 83% for the base and 81% for the bias treatment, respectively. Bias feedback will offer limited additional information for this group of people as they do not suffer from bias.

| | Total Obs. | %No Bias | %Over | %Under |
|------|------------|----------|-------|--------|
| Base | 7,308 | 83% | 16% | 1% |
| Bias | 8,136 | 81% | 18% | 1% |

Table 19 Percentage of Observations classified as Over-forecasting & Under-forecasting in Base and Bias Treatments

Second, when participants display a bias, it is towards over-forecasting than under-forecasting. We compared the average bias performance of instances of over-forecasting and under-forecasting in the bias treatment against the base treatment. Tables 20a and 20b show the results of our comparison.

| | <u>Bias Metric (Over-forecasting)</u> | |
|------------|--|-------------------|
| | <u>N10</u> | <u>N40</u> |
| C0 | -7.2 (-6.6) | -9.29*** (-4.9) |
| C40 | -24.5 (-22.5) | -32.24 (-28.8) |

Table 20a Bias Performance of Observations Classified as Over-forecasting from Bias and Base Treatment

Note: Numbers in parenthesis indicate bias performance of subjects in the base treatment

| | <u>Bias Metric (Under-forecasting)</u> | |
|------------|---|-------------------|
| | <u>N10</u> | <u>N40</u> |
| C0 | 1.78*** (6.8) | 3.8*** (4.9) |
| C40 | 10.34*** (5.07) | 13.9 (12.83) |

Table 20b Bias Performance of Observations Classified as Over-forecasting from Bias and Base Treatment

Based on Table 20a, bias feedback was ineffective at reducing over-forecasting or, even worse, increased it. Bias feedback seemed to be more effective at reducing under-forecasting. To conclude, given that under-forecasting was just observed in 1% of the cases, we observe that either bias feedback is unnecessary – because the forecasters do not display bias – or was ineffective, as in the case of over-forecasting. This may explain why there is essentially no difference in the adjustment scores between the bias and the base treatments.

5.4.4 Bias with Rolling Training

The rolling training approach is characterized by providing performance feedback at regular intervals instead of every time period. We argued earlier that limiting the number of reviews emphasizes the review's "decision-making value" (Long et al., 2020). With the limited number of reviews, participants are expected to be cognitively more attentive. Petropoulos et al. (2017) claim that the rolling training approach enables the *"balance between the sensitivity and stability of the feedback."*

We argue that compared with the base treatment, where participants are exposed to forecast errors every time period, rolling training, with its emphasis on balanced feedback and limited reviews, could make them more aware of the demand generation process. For example, they could notice the minor variation in the demand around the mean in a stable series and realize it is a noise. Similarly, in an unstable process, while there may be consecutive series of increasing and decreasing demand, attentive participants are likely to realize that the variation from one bucket to another is indicative of a change than a persistent trend. Hence, our hypothesis proposed that the rolling training approach would help in both the underreaction and overreaction scenarios.

H4: *A rolling approach to providing bias feedback reduces overreaction and underreaction*

| | BRT | |
|------------|---------------|-------------|
| | N10 | N40 |
| C0 | 0.34 (0.40) | 0.29 (0.46) |
| C40 | 0.30** (0.56) | 0.32 (0.54) |

Table 21 Wald test of Treatment vs. Base Adjustment Score

As defined previously, the adjustment scores and forecast errors in the rolling training conditions are calculated per forecasting bucket consisting of 4 time periods rather than for each time period as was done in the other treatments.

Compared with the base treatment, the rolling training approach reduces overreaction (although not significantly) in both stable conditions (Table 21). In the case of unstable conditions, the rolling treatment could not trigger an increase in the subjects' adjustment scores (α) compared to the base treatment (increasing reaction). In unstable conditions (conditions 3 and 4), the desired reaction is that more significant errors must greatly influence the forecast. The individual adjustment scores (α) were reduced across the board.

The directional bias feedback did not help participants distinguish the variation in demand around the mean in stable conditions as noise.

Counter to our argument, the participants' limited number of reviews also did not help reduce system neglect. If participants gained an insight into the demand generation process and, subsequently, the system

parameters, we would have observed increases in their adjustment scores (α) in unstable conditions.

| | Model 1 | Model 2 | Model 3 | Model 4 |
|-----------------|--------------------------------|--------------------------------|---------------------------------|---------------------------------|
| | Condition 1 (C0N10) | Condition 2 (C0N40) | Condition 3 (C40N10) | Condition 4 (C40N40) |
| | Coefficient | Coefficient | Coefficient | Coefficient |
| Bias w Rol.Trng | 0.01 (0.05) | -0.08 (0.05) | -0.16** (0.01) | -0.08 (0.06) |
| Constant | 0.45*** (0.15) | 0.03 (0.20) | 0.33** (0.16) | 0.76*** (0.20) |

Table 22 Nested Models Rolling Training Estimation

Note: Controls included in the estimation.

The nested model estimation outputs are shown in Table 22 and include the control variables. Rolling training does not significantly impact underreaction or overreaction. Thus, our hypothesis is not supported.

5.4.5 Full Parameter Knowledge Treatment:

Our intent with the various treatments was to provide subjects with decision tools to help remedy the system neglect and improve forecasting performance. Thus, the base condition theoretically provides a lower bound for forecasting performance. Kremer et al. further claim, *"In most instances, system parameters c and n are unknown or even unknowable."* Therefore, knowledge of system parameters should help decide the right adjustment scores. Of course, this assumes that the subjects can incorporate the information about the system parameters in their demand estimation.

We conducted a separate experiment where participants were informed about the nature of the demand function, and the system parameters c and n were disclosed to the participants. This scenario was conducted

for the high noise and high change conditions (since this condition represents the most complex of all conditions). This is different from the original decomposition approach, where participants were provided information about the nature of the time series but were not provided any information on the system parameters.

| Wald test of Optimal Alpha vs. Subject Adj Score | | | |
|---|----------------|----------------------|--------------------------|
| Treatment | | | |
| Condition | Base | Decomposition | Decomposition(PK) |
| C40N40 | 0.54*** (0.62) | 0.20*** (0.62) | 0.30*** (0.62) |

Table 23a Wald test of Optimal Alpha vs. Subject Adj Score

| Wald test of Base vs. Subject Adj Score | | |
|--|----------------------|--------------------------|
| Condition | Decomposition | Decomposition(PK) |
| C40N40 | 0.20*** (0.54) | 0.30*** (0.54) |

Table 23b Wald test of Base vs. Subject Adj Score

Table 23a and Table 23b compare the adjustment scores from the parameter knowledge treatment against the optimal values and the base treatment.

The average adjustment scores continue to show underreaction compared to the base treatment (Table 23b). Compared to the original decomposition approach, the adjustment scores have increased from 0.20 to 0.30. It seems that the participants cannot still effectively

incorporate the knowledge about the c and n into their forecast estimates.

Improving accuracy in unstable conditions requires subjects to weigh recent forecast errors higher in their adjustment scores. The treatment has not triggered such a response in the subjects. Table 24 shows the estimation results from the multi-level model. It shows that parameter knowledge reduces the subject adjustment score compared to the base treatment. However, the adjustment scores are higher when compared with the decomposition treatment (without parameter knowledge).

| Model 4 | |
|-----------------------------|------------------------|
| Condition 4 (C40N40) | |
| | Coefficient |
| PK Decomposition | -0.17*** (0.07) |
| Constant | 0.67*** (0.04) |

Table 24 Estimation Results for Condition 4, including Parameter Knowledge Treatment

Table 25 T-test of Observed MAPE vs. Opt MAPE

| | Base | | Bias | | BRT | |
|------------|----------------|----------------|---------------------------|----------------|-------------------------|-----------------|
| | N10 | N40 | N10 | N40 | N10 | N40 |
| C0 | 2.20(1.18)*** | 5.0(5.56)*** | 2.86(1.37)*** | 6.89(5.54)*** | 4.97(1.38)*** | 10.20(5.93)*** |
| C40 | 10.82(3.96)*** | 13.31(6.26)*** | 12.59(4.01)*** | 17.27(6.38)*** | 15.52(3.95)*** | 21.67(6.32)*** |
| | Fan | | Decomposition | | PK Decomposition | |
| | N10 | N40 | N10 | N40 | N10 | N40 |
| C0 | 4.52(1.38) | 11.08(6.26)*** | C0 9.10(1.37)*** | 12.29(5.82)*** | - | - |
| C40 | 10.02(4.01)*** | 9.23(5.72)*** | C40 15.04(4.04)*** | 11.(5.48)*** | - | 13.81 (5.63)*** |

Table 26 T-test of MAPE Treatment vs. Base

| | Bias | | BRT | | Fan | |
|------------|----------------------|-----------------|-------------------------|-----------------|---------------|----------------|
| | N10 | N40 | N10 | N40 | N10 | N40 |
| C0 | 2.86(2.20)*** | 6.89(5.00)*** | 4.97(2.20)*** | 10.20(5.00)*** | 4.52(2.20) | 11.08(5.00)*** |
| C40 | 12.59(10.82)*** | 17.27(13.31)*** | 15.52(10.82)*** | 21.67(13.31)*** | 10.02(10.82)* | 9.23(13.31)*** |
| | Decomposition | | PK Decomposition | | | |
| | N10 | N40 | N10 | N40 | | |
| C0 | 9.10(2.20)*** | 12.29(5.00)*** | - | - | | |
| C40 | 15.04(10.82)*** | 11.06(13.31)*** | - | 13.81 (13.31) | | |

6. Forecasting Performance Implications

The ultimate objective of any forecasting exercise is to reduce errors and improve accuracy. Accordingly, subjects in our forecast exercise were instructed to minimize their forecasts' MAPE, and respondents with the lowest MAPE (top three) were incentivized with a bonus payment of USD 15. Except for the decomposition treatment, the participants did not have any indication of the demand function and the research hypothesis. There was no apriori forecasting knowledge required from the participants. We designed the treatments to address the causes of underreaction and overreaction to help improve the MAPE of the forecasting process. We deployed Fan Charts and Bias treatments to reduce the salience of forecast errors. We used rolling training and decomposition approaches to emphasize the system parameters and reduce the salience of forecast errors.

The MAPE values comparisons for each treatment are presented in Tables 25 (Observed vs. Optimal) and 26 (Treatment vs. Base). Table 25 shows that the observed MAPEs are higher than the optimal MAPEs in all treatment and condition combinations. In Table 26, we see that only conditions 3 and 4 in the Fan Chart treatment and condition 4 in the decomposition treatment resulted in MAPE values lower than the base condition. We see an increase in MAPE in all other conditions compared to the Base treatment.

A possible explanation for not achieving any reduction in MAPE was that the participants could not fully leverage the treatment mechanisms and

apply them to the forecasting process. Although instructions and short training videos were available for participants' reference, given the limited time, participants might not have been able to comprehend fully. Recent research by Aguinis et al. (2021) has identified some challenges with using the MTurk subject pool. MTurk subject pools generally work to maximize monetary gains by completing as many tasks as possible. To quote Aguinis et al. 2021 "*Compared with student samples, online participants are significantly more likely to be distracted due to cell phone use (MTurker = 21% vs. student = 9%), internet surfing (MTurker = 11% vs. student = 1%), or conversing with another person (MTurker = 21% vs. student = 2%)*". Nevertheless, we do not find significant support for the attention check questions or the knowledge of fan charts as a predictor of performance. Therefore, we cannot conclude whether this is the main contributor to the poor performance of the treatments tested.

The base treatment is cognitively the least demanding of all the treatments, which may be why it resulted in lower MAPE for the online pool. Some exploratory analysis points to the possibility of limited cognitive attention. The base treatment, which is the least informative but cognitively the least demanding, nevertheless resulted in lower MAPE. To understand whether the participants did suffer from cognitive load in the treatments provided, we plotted the MAPE across treatments against the self-declared usage of the graphs (Figure 14). Except for the base treatment in which usage of the graph is beneficial, the usage versus MAPE relationship resembles an inverted "U," such that very low or very high usage levels are associated with lower MAPE than

intermediate use. This inferior performance of intermediate graph usage in the treatments where the graphs provided significantly different information from the data tables supports the idea of a switching cost where the participants need to adjust between the various elements of the complex graphic and the numeric feedback provided. This increases the cognitive burden and reduces forecasting performance.

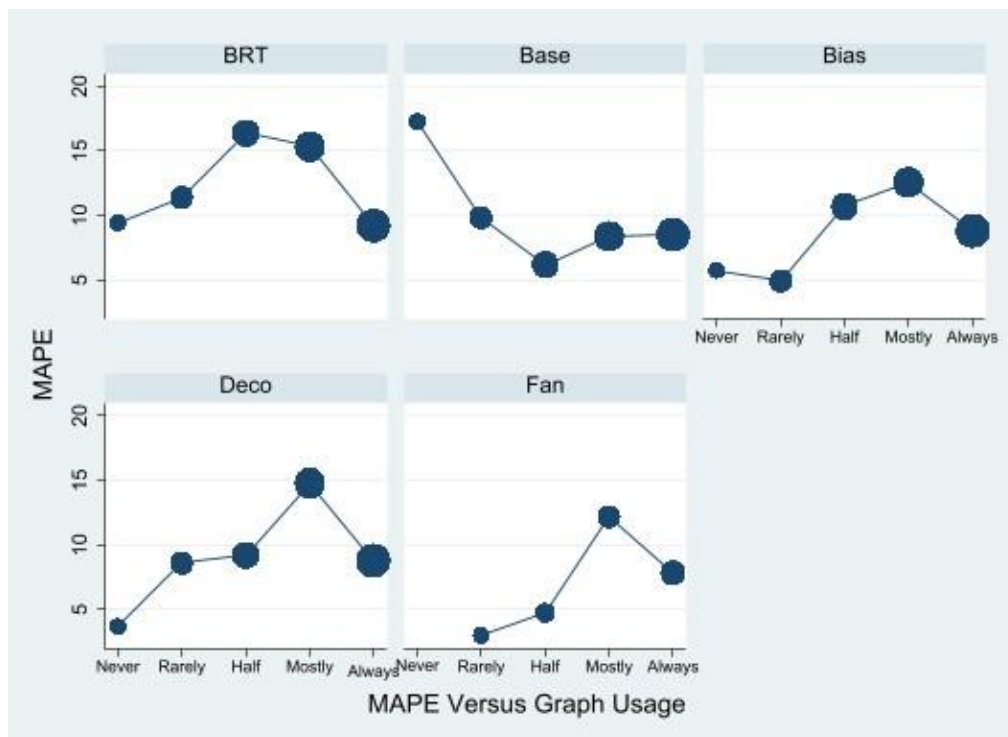


Figure 14 Subject MAPE versus the usage of Graphs

Note: The size of the bubble represents the frequency of the observation.

| In | Treatment | Prediction | Result | MAPE |
|----------|-------------------------|--|---------------------------------|--|
| H1a | Fan Chart | Fan Chart Reduces Overreaction | Not Supported | Lower for condition 3. Increased for conditions 1,2 and 4 |
| H1b-(i) | Fan Chart | Fan Chart Reduces Underreaction | Not Supported | |
| H1b-(ii) | Fan Chart | Fan Chart Increases Underreaction | Not Supported | |
| H2 | Decomposition | Decomposition Reduces Overreaction and Underreaction | Supported for Overreaction only | Lower for condition 4. Increased for conditions 1,2, and 3. |
| H3a | Bias | Bias Feedback Reduces Overreaction | Not Supported | Increased for all conditions |
| H3b-(i) | Bias | Bias Feedback Reduces Underreaction | Not Supported | |
| H3b-(ii) | Bias | Bias Feedback Increases Underreaction | Not Supported | |
| H4 | Bias + Rolling Training | BRT Reduces Overreaction and Underreaction | Not Supported | Increased for all conditions |

Table 27. Summary of Results

7. Managerial Implications

Given the prevalence of judgmental forecasting in industry, our research objective was to identify approaches to mitigate the overreaction and underreaction observed in time series with noise and change. Kremer et al. (2011) identified overreaction and underreaction as the primary factors impacting forecasting performance. We deployed and tested decision support tools to help improve the forecast quality in our context characterized by noise and change. We confirmed that rolling training and decomposition approaches were able to reduce overreaction in the stable time series. However, this did not translate into improvements in forecast quality, and we failed to establish a consistent relationship between adjustment scores (α) and MAPE.

In fact, the cases in which a significant improvement in MAPE was observed were all cases in which the adjustment factor changed in a direction opposite to the one theorized by Kremer et al. (2011). In instances where we significantly reduced underreaction in stable conditions, the MAPE nevertheless increased. Achieving an improvement in MAPE proved to be a challenging task.

This leads us to question whether the demand forecasting model tested by Kremer et al. (2011) is a good model of the individual participants' forecasting decisions, as the forecast models that individuals seem to deploy do not conform to the single exponential model. If the exponential smoothing model were not to apply, then focusing on the adjustment

factor would not achieve the desired outcome, which is improved forecasting accuracy.

Another important finding from our research is that individuals cannot effectively distinguish between noise and change. Only demand series with low change and low noise is reliably recognized as stable demand series. Demand patterns with a high noise but low change are frequently misclassified as unstable. This is unfortunate as the desired forecasting behavior depends on correctly attributing demand fluctuations to change vs. noise. Demand patterns without change require a stable forecast and a stronger weight to past observations and downplay errors. In contrast, demand patterns displaying significant change require a stronger weight towards recent observations that reflect the change.

Even though we could not improve forecast performance by altering adjustment scores (α), there are still a few insights that managers can leverage from our study.

First, in stable demand conditions, more straightforward approaches are preferred. They are intuitive and cognitively less demanding. This is likely why the base, bias, and fan chart-based methods had lower MAPE than rolling training and decomposition-based approaches. The base approach was the better-performing one amongst the others, indicating limited incremental benefits with the additional tools. To fully leverage fan chart-based approaches requires an appreciation of uncertainty and confidence intervals. The results from the survey questions suggest a lack of such knowledge in the general population. Developing such an

appreciation requires some level of training and education that managers need to cater to. The enhanced knowledge could help further improve performance in stable conditions.

Stable time series are, by nature, more suited for automation. However, the well-known algorithm aversion phenomenon means that humans are hesitant to use algorithms since they are imperfect, even if they perform better than humans. Dietvorst et al. (2018) find that people may be more willing to accept the outcomes from an algorithm if they are allowed to change their forecasts even by a small amount. This approach may help balance automation and the manual judgmental approach to improve judgmental forecasts.

Second, distraction or additional cognitive effort may also lead to poorer performance. Bias, Rolling Training, Fan Chart, and Decompositions were all treatments with additional intervention and action required from the participants. The most straightforward base treatment outperformed the treatments in the M-Turk participant pool, which is drawn from the general population.

Third, and interestingly, none of the treatments worked in all the conditions. This suggests the need for a multi-pronged approach based on the nature of time series. Automating decisions in a stable situation and simplifying the decomposition-based approach in an unstable situation hold the most promise. Such an approach would require some statistical analysis of the time-series a priori. Budescu and Chen (2015)

suggest an approach that extends the crowds' wisdom using an expert pool of forecasters who consistently outperform the others.

8. Conclusion

We started our research by replicating the findings from the original Kremer et al. (2011) research about system neglect. Our goal was to identify treatments that could help mitigate system neglect's effects on forecasting performance by influencing the individual adjustment scores (α). Table 27 summarizes our findings. We have only been able to trigger a reduction in over-reaction with the decomposition approach. None of the treatments were able to trigger an increase in adjustment scores (α) for the unstable conditions (reduce underreaction). In terms of forecasting performance, we reduced MAPE in only three conditions. The Fan Chart treatment delivered a reduction in MAPE of 0.8% for condition 3 and 3.08% for condition 4, whereas the decomposition treatment delivered a reduction in MAPE of 2.25% compared to the base treatment.

The treatments were ineffective in delivering consistent improvement across all the conditions. The results find some support in prior research highlighted in the literature review. For example, Blogger and Harvey (1993) found that individual forecasting performance is highly dependent on the characteristics of time series, and small changes in the time series and its presentation can significantly affect forecasters' performance.

The current research points to several areas for further investigation. Firstly, the normative model used in our and the Kremer et al. study was

the single exponential smoothing model. We have evaluated individual performance against the normative benchmark from the exponential smoothing model. Our results show that better alignment with the benchmark need not lead to improved forecasting accuracy. Future research could investigate other normative models for the demand function and evaluate individual responses against it. There is also a potential to explore other behavioral models of forecasting which could better explain the respondent's behavior. We noted earlier that Tong and Feiler (2017) advanced a behavioral forecasting model, arguing that individuals generate a forecast based on a small (less than 7) and randomly generated sample of the series and naively assume that the sample represents the true population. Researchers could evaluate if Tong and Feiler's model could better describe the individual responses.

Second, the experimental setup could be varied to understand how individual responses vary based on the time series characteristics. Within a given treatment, a combination of the conditions can be deployed such that the same individual is exposed to different conditions within a given treatment. Such treatments can help isolate individual responses to differences in time series. E.g., when the demand conditions switch from a stable to an unstable condition, we could confirm if the individual response switches from overreaction to underreaction or vice-versa. In real-world scenarios, planners often have to deal with time series with varying characteristics in no specific order; insights from treatments with varying demand functions can help design effective interventions from a practitioner's standpoint.

Third, as prior research has highlighted the prevalence of judgmental forecasting among prior practitioners, it is critical to ensure that forecasters are adequately trained and qualified for the forecasting task. Budescu and Chen (2015) showed that using an expert pool of forecasters who consistently outperform the others delivers significant improvements in the quality of aggregation when compared with the standard pool. A recent study by Kim et al. (2019) studied how individuals and groups respond to the advice they receive as part of a judgmental forecasting activity. They find that groups are better at discerning the quality of forecasting advice than individuals and generally perform better than individuals. Both studies suggest that the potential use of a qualified pool of experts may help improve the forecasting performance compared to a normal pool.

From a practitioner standpoint, this is quite relevant where organizations are building Centres of Excellence (COE) to develop and leverage expertise. While the forecasting decisions involving stable time series could be automated to deliver performance improvements, the pool of experts could help drive improvement in performance for unstable series could be derived. Future studies could look at evaluating these scenarios impact on the system neglect phenomenon.

Forecasting is a critical business process that significantly impacts the top and bottom lines. Our research has attempted to find interventions that could improve forecasting by mitigating the effects of underreaction and underreaction to forecast errors. We found no treatment worked across the conditions, and the individual behavior did not conform to the

single exponential smoothing forecasting model. Based on our results, we have highlighted some fruitful areas for further research for practitioners and academicians.

References

- Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgment into quantitative forecasting methods: A review. *Omega*, 86, 237-252.
- Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, 47(4), 823-837.
- Bazerman, M. H., & Moore, D. A. (2013). Judgment in managerial decision-making.
- Bendoly, E., Croson, R., Goncalves, P., & Schultz, K. (2010). Bodies of knowledge for research in behavioral operations. *Production and Operations Management*, 19(4), 434-452.
- Bolger, F., & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *The Quarterly Journal of Experimental Psychology Section A*, 46(4), 779-811.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155-1170.
- Donohue, K., Katok, E., & Leider, S. (Eds.). (2019). *The handbook of behavioral operations*. John Wiley & Sons.
- Edmundson, R. H. (1990). Decomposition; a strategy for judgemental forecasting. *Journal of Forecasting*, 9(4), 305-314.
- Eroglu, C., & Croxton, K. L. (2010). Biases in judgmental adjustments of statistical forecasts: The role of individual differences. *International Journal of Forecasting*, 26(1), 116-133.
- Fahimnia, B., Pournader, M., Siemsen, E., Bendoly, E., & Wang, C. (2019). Behavioral operations and supply chain management—a

- review and literature mapping. *Decision Sciences*, 50(6), 1127-1183.
- Feiler, D. C., Tong, J. D., & Larrick, R. P. (2013). Biased judgment in censored environments. *Management Science*, 59(3), 573-591.
- Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6), 570-576.
- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International journal of forecasting*, 25(1), 3-23.
- Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting*, 36, 5-12.
- Gino, F., & Pisano, G. (2008). Toward a theory of behavioral operations. *Manufacturing & Service Operations Management*, 10(4), 676-691.
- Goodwin, P. (2000) Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting*, 16, 261– 275.
- Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, 9(2), 147-161.
- Harvey, N. (2001). Improving judgment in forecasting. In *Principles of forecasting* (pp. 59-80). Springer, Boston, MA.
- Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *Thinking & reasoning*, 13(1), 5-24.
- Kim, H. Y., Lee, Y. S., & Jun, D. B. (2020). Individual and group advice taking in judgmental forecasting: Is group forecasting superior to

- individual forecasting?. *Journal of Behavioral Decision Making*, 33(3), 287-303.
- Kremer, M., Moritz, B., & Siemsen, E. (2011). Demand forecasting behavior: System neglect and change detection. *Management Science*, 57(10), 1827-1843.
- Kremer, M., Siemsen, E., & Thomas, D. J. (2016). The sum and its parts: Judgmental hierarchical forecasting. *Management Science*, 62(9), 2745-2764.
- Kreye, M. E., Goh, Y. M., Newnes, L. B., & Goodwin, P. (2012). Approaches to displaying information to assist decisions under uncertainty. *Omega*, 40(6), 682-692.
- Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, 1(1), 25-35.
- Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3), 493-518.
- Legerstee, R., & Franses, P. H. (2014). Do experts' SKU forecasts improve after feedback? *Journal of Forecasting*, 33(1), 69-79.
- Lee, Y. S., & Siemsen, E. (2017). Task decomposition and newsvendor decision making. *Management Science*, 63(10), 3226-3245.
- Lee, Y. S., Seo, Y. W., & Siemsen, E. (2018). Running behavioral operations experiments using Amazon's Mechanical Turk. *Production and Operations Management*, 27(5), 973-989.
- Lin, V. S., Goodwin, P., & Song, H. (2014). Accuracy and bias of experts' adjusted forecasts. *Annals of Tourism Research*, 48, 156-174.
- Long, X., Nasiry, J., & Wu, Y. (2020). A behavioral study on abandonment decisions in multistage projects. *Management Science*, 66(5), 1999-2016.

- Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd ed.). John Wiley & Sons.
- MacGregor, D. G. (2001). Decomposition for judgmental forecasting and estimation. In *Principles of forecasting* (pp. 107-123). Springer, Boston, MA.
- Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 51(6), 932-947.
- Moritz, B., Siemsen, E., & Kremer, M. (2014). Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management*, 23(7), 1146-1160.
- Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, 115(2), 502.
- Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*, 55(290), 299-306.
- O'Dwyer, L. M., & Parker, C. E. (2014). *A Primer for Analyzing Nested Data: Multilevel Modeling in SPSS Using an Example from a REL Study*. REL 2015-046. Regional Educational Laboratory Northeast & Islands.
- Oliva, R., & Watson, N. (2009). Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production and Operations Management*, 18(2), 138-151.
- Perera, H. N., Hurley, J., Fahimnia, B., & Reisi, M. (2019). The human factor in supply chain forecasting: A systematic review. *European Journal of Operational Research*, 274(2), 574-600.
- Petropoulos, F., Goodwin, P., & Fildes, R. (2017). Using a rolling training approach to improve judgmental extrapolations elicited from

- forecasters with technical knowledge. *International Journal of Forecasting*, 33(1), 314-324.
- Sanders, N. R., & Ritzman, L. P. (1995). Bringing judgment into combination forecasts. *Journal of Operations Management*, 13(4), 311-321.
- Scheele, L. M., Thonemann, U. W., & Slikker, M. (2018). Designing incentive systems for truthful forecast information sharing within a firm. *Management Science*, 64(8), 3690-3713.
- Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. (2015). Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36, 33-45.
- Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.
- Tong, J., & Feiler, D. (2017). A behavioral model of forecasting: Naive statistics on mental samples. *Management Science*, 63(11), 3609-3627.
- Tong, J., Feiler, D., & Larrick, R. (2018). A behavioral remedy for the censorship bias. *Production and Operations Management*, 27(4), 624-643.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, 39(7), 1275-1289.

Appendix 1

Recruitment material for the Base Treatment:

Instructions

Study Overview and Context

This is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process. The experiment is to support Doctoral Research.

You have been given the actual demand for the past 36 time periods. You will need to predict the subsequent 36 time periods.

Software Required

You will need to have Microsoft Excel 2012 or a later version, and you will need to enable Macros.

Estimated Completion Time

This task is estimated to take **25 minutes**.

To Receive Credit and Avoid Your Submission Being Rejected

You must complete and upload the excel simulation with your forecast into the survey link and complete the post-simulation survey questionnaire.

Not following the prescribed steps will result in the submission being rejected.

Bonus Payment

The top 3 entries with the highest forecast accuracy will be awarded a bonus payment of USD 15, which will be provided approximately 2 weeks from the study completion.

Study Withdrawal

You may withdraw from the study within 72 hours of completion by informing the principal investigator Srikant Vinakota via email at srikant.v@2017@dba.smu.edu.sg. You may also email the principal investigator's supervisor Prof. Pascale Crama at pcrama@smu.edu.sg.

Withdrawn entries will not be reimbursed for the participation fee.

Institutional Research Board:

For questions about IRB, you may contact irb@smu.edu.sg. Please include IRB approval number:IRB-21-105-E026-M1(522)

Select the link below to complete the survey. You will receive a code to paste into the box below to receive credit for taking our survey at the end of the survey.

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box below.

Survey link:

The link will appear here only if you accept this HIT.

Provide the spreadsheet completion code here:

Appendix 2a
Instructions For Base Treatment
Judgmental Demand Forecasting

The following is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process.

Context: You are the demand planner for X-mart, a supermarket. Your role is to forecast the future demand for products based on actual past demand.

The demand forecast forms the input for subsequent activities such as procurement from the supplier and replenishment of the store. The demand forecast is the first step in the planning process, enabling product availability, and an accurate forecast leads to improved business performance.




You have been given the actual demand for the past 36 time periods. You will need to predict the subsequent 36 time periods one step ahead for the next selling period, i.e., forecast period 37, using data available through period 36. Actual demand for period 37 is then realized, and you will be asked to forecast period 38.

Forecast Performance Measures:

Note below the definition of the forecast performance measures. The table below shows an example calculation for the measures. The number in the red circle corresponds to the measure described below.

1. Forecast error is the absolute difference between the actual demand and the forecast. **For time period 38, this is the absolute difference between 90 (actual) and 110 (forecast), and it is 20.**
2. Absolute Percentage Error (APE) is the forecast error expressed as a percentage of the actual demand. **For time period 38, this is calculated as $(20/90)$, which is 22.1%**
3. MAPE refers to the Mean of the absolute percentage error (APE) over the given time periods. **For period 38, this is the average of period 37 (10%) and period 38 (22.2%), which is 16.1%.**

Your Objective - *Minimize* your forecast error which leads to the minimization of MAPE

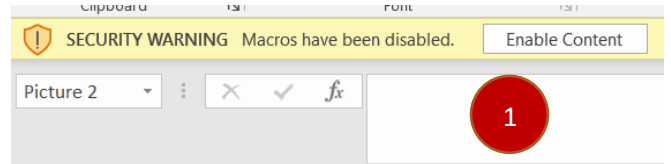
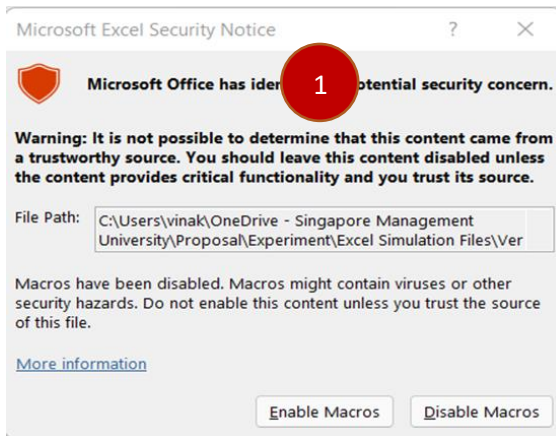
| | | |  |  |  |
|-------------|---------------|----------|--|---|---|
| Time Period | Actual Demand | Forecast | Forecast Error | APE(%) | MAPE(%) |
| T37 | 100 | 90 | 10 | 10% | 10.0% |
| T38 | 90 | 110 | 20 | 22% | 16.1% |
| T39 | 110 | 100 | 10 | 9% | 13.8% |

Reward

Participants will be paid **USD 2** for fully completed responses. In addition, there is a bonus reward of **USD 15** for the top **3** respondents with the lowest MAPE for the 36 periods.

Spreadsheet set up and overview

1. The spreadsheet uses macros. You may receive security warning messages similar to the one shown below. Please Enable Macros to use the spreadsheet
2. Please ensure you key in your Participant ID in **Cell B1** in the Tab -Spreadsheet Simulation
3. You are requested to complete the task in one sitting



Please Enter
Your
Participant ID



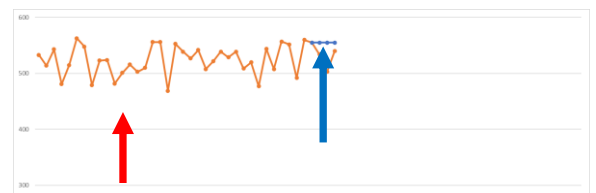
Steps

The Actual Demand for the Product are shown in the column labeled Actual Demand (Column C in YELLOW).

1. Provide your forecast in the column labeled User Forecast (Column D in BLUE) from Row 37 onwards.
Note: i) enter values greater than 0, ii) text entries are not permitted, iii) forecasts once entered cannot be changed.
2. The performance metrics will be updated after every submission
3. The actuals and the forecast are also plotted in a graph that you are encouraged to review before your next submission.
4. Once you have completed the submission for time period 72, the spreadsheet simulation task is complete.
5. YOU MUST THEN SAVE THE SPREADSHEET AND UPLOAD IT BACK TO THE SURVEY.
6. Proceed to complete the questionnaire in the survey. You will receive a completion code at the end of the questionnaire, which must be typed in **MTurk** to get credit for your effort.

| Time Period | Actual Sale | Your Forecast | Forecast Error | APE (%) | MAPE (%) | Mean Forecast Accuracy (100-MAPE) |
|-------------|-------------|---------------|----------------|---------|----------|-----------------------------------|
| T37 | 555 | 555 | 0 | 0.0% | 0.0% | 100.0% |
| T38 | 534 | 555 | 21 | 3.9% | 2.0% | 98.0% |
| T39 | 503 | 555 | 52 | 10.3% | 4.8% | 95.2% |
| T40 | 540 | 555 | 15 | 2.8% | 4.3% | 95.7% |

SAMPLE DATA



Key in your inputs here

Actual Demand

Your Forecast

Appendix 2b

Instructions for Fan Chart Treatment

Judgmental Demand Forecasting

The following is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process.

Context: You are the demand planner for X-mart, a supermarket. Your role is to forecast the future demand for products based on actual past demand.

The demand forecast forms the input for subsequent activities such as procurement from the supplier and replenishment of the store. The demand forecast is the first step in the planning process, enabling product availability, and an accurate forecast leads to improved business performance.

You have been given the actual demand for the past 36 time periods. You will need to predict the subsequent 36 time periods one step ahead for the next selling period, i.e., forecast period 37, using data available through period 36. Actual demand for period 37 is then realized, and you will be asked to forecast period 38.

Forecast Performance Measures:

Note below the definition of the forecast performance measures. The table below shows an example calculation for the measures. The number in the red circle corresponds to the measure described below.

1. Forecast error is the absolute difference between the actual demand and the forecast. **For time period 38, this is the absolute difference between 90 (actual) and 110 (forecast) and is 20.**
2. Absolute Percentage Error (APE) is the forecast error expressed as a percentage of the actual demand. **For time period 38, this is calculated as (20/90), which is 22.1%**
3. MAPE refers to the Mean of the absolute percentage error (APE) over the given time periods. **For period 38, this is the average of period 37 (10%) and period 38 (22.2%), which is 16.1%.**

Your Objective - *Minimize* your forecast error which leads to minimization of MAPE

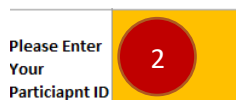
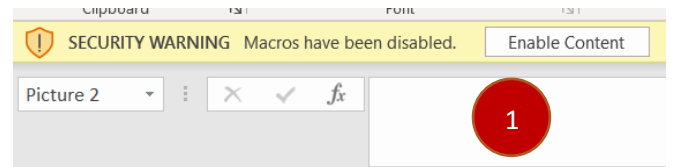
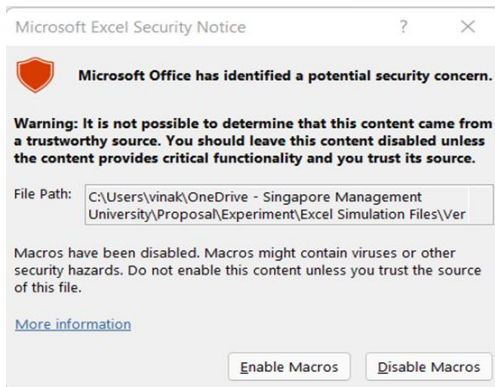
| | | | Forecast Measures | | |
|-------------|---------------|----------|-------------------|--------|---------|
| | | | 1 | 2 | 3 |
| Time Period | Actual Demand | Forecast | Forecast Error | APE(%) | MAPE(%) |
| T37 | 100 | 90 | 10 | 10.0% | 10.0% |
| T38 | 90 | 110 | 20 | 22.2% | 16.1% |
| T39 | 110 | 100 | 10 | 9.1% | 13.8% |

Reward

Participants will be paid **USD 2** for fully completed responses. In addition, there is a bonus reward of **USD 15** for the top **3** respondents with the lowest MAPE for the 36 periods.

Spreadsheet set up and overview

1. The spreadsheet uses macros. You may receive security warning messages similar to the one shown below. Please **Enable Macros** to use the spreadsheet
2. Please ensure you key in your Participant ID in **Cell B1** in the Tab -Spreadsheet Simulation
3. You are requested to complete the task in one sitting



Steps

The Actual Demand for the Product is shown in the column labeled Actual Demand (Column C in **YELLOW**).

1. Provide your forecast in the column labeled "User Forecast" (Column D in **BLUE** from Row 37 onwards.
Note: i) enter values greater than 0, ii) text entries are not permitted, iii) forecasts once entered cannot be changed.

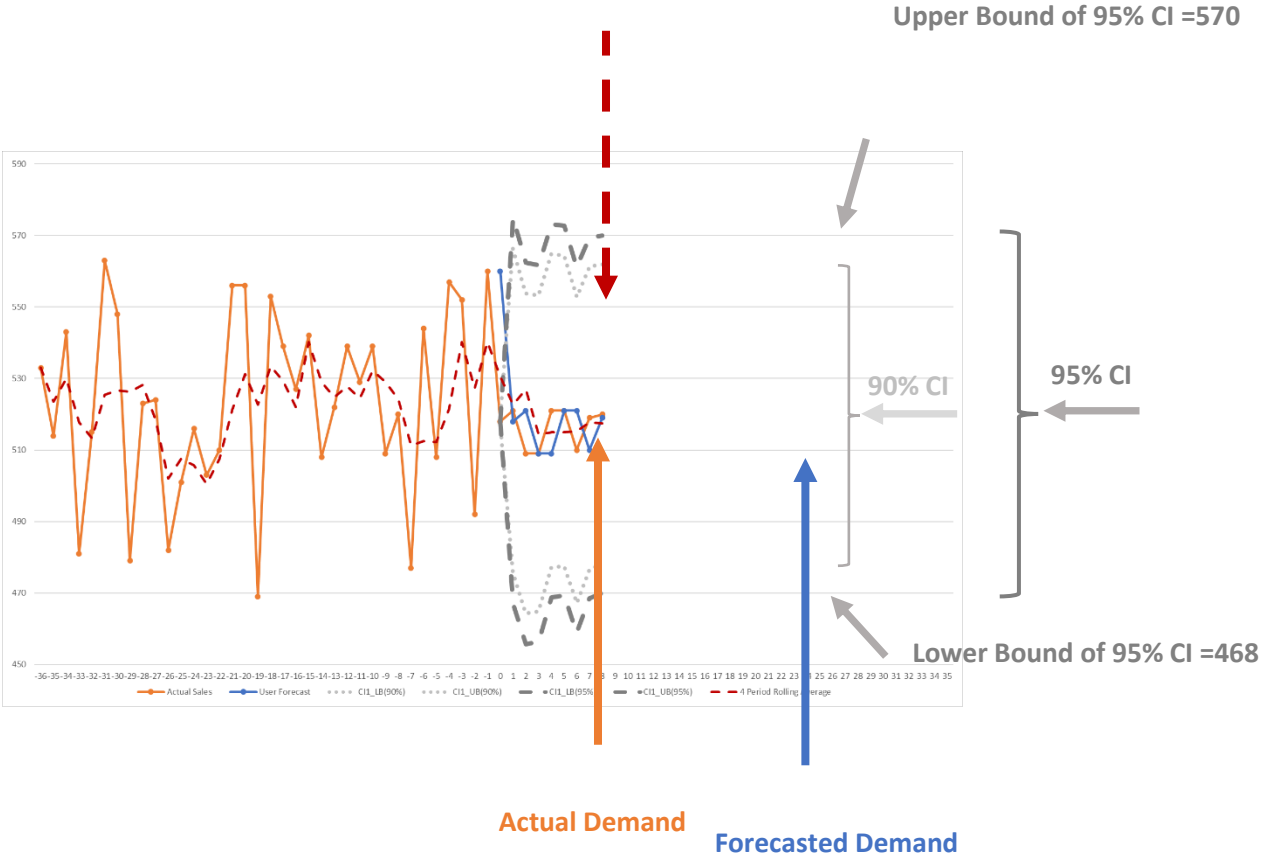
2. The actuals and the forecast are also plotted in a graph that includes a line chart, a moving average plot, and 90% and 95% confidence interval plots.
3. The 90% confidence interval means there is a 90% likelihood that the next demand will be in this interval. The midpoint of the interval is the forecast. In other words, there is a 90% chance that the correct demand is included by the lower and upper bounds of the interval, and there is only a 5% chance that demand will be higher than the upper bound and a 5% chance that it will be lower than the lower bound (see Figure 1)
4. The past four periods moving average of actuals is included to give you an indication of the evolution of the demand
5. Once you have completed the submission for time period 72, the spreadsheet simulation task is complete.
6. YOU MUST THEN SAVE THE SPREADSHEET AND UPLOAD IT BACK TO THE SURVEY.
7. Proceed to complete the questionnaire in the survey. You will receive a completion code at the end of the questionnaire, which must be typed in **MTurk** to get credit for your effort.

| Time Period | Actual Sale | Your Forecast | Forecast Error | APE (%) | MAPE (%) | Mean Forecast Accuracy | MAPE |
|-------------|-------------|---------------|----------------|---------|----------|------------------------|-------|
| T37 | 555 | 555 | 0 | 0.0% | 0.0% | | |
| T38 | 534 | 555 | 21 | 3.0% | | | |
| T39 | 503 | 555 | 52 | 10.3% | | | 95.2% |
| T40 | 540 | 555 | 15 | 2.8% | | | 95.7% |

SAMPLE DATA

Key in your inputs here

Past 4 period moving average of actuals



Appendix 2c

Instructions for Bias Treatment

Judgmental Demand Forecasting

The following is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process.

Context: You are the demand planner for X-mart, a supermarket. Your role is to forecast the future demand for products based on actual past demand.

The demand forecast forms the input for subsequent activities such as procurement from the supplier and replenishment of the store. The demand forecast is the first step in the planning process, enabling product availability, and an accurate forecast leads to improved business performance.

You have been given the demand for the past 36 time periods. You will need to predict the subsequent 36 time periods one step ahead for the next selling period, i.e., forecast period 37, using data available through period 36. Actual demand for period 37 is then realized, and you will be asked to forecast period 38.

Forecast Performance Measures:

Note below the definition of the forecast performance measures. The table below shows an example calculation for the measures. The number in the red circle corresponds to the measure described below.

- Forecast error is the absolute difference between the actual demand and the forecast. **For time period 38, this is the absolute difference between 90 (actual) and 110 (forecast) and is 20.**
- Absolute Percentage Error (APE) is the forecast error expressed as a percentage of the actual demand. **For time period 38, this is calculated as $(20/90)$, which is 22.2%**
- MAPE refers to the Mean of the absolute percentage error (APE) over the given time periods. **For period 38, this is the average of period 37 (10%) and period 38 (22.2%), which is 16.1%.**
- Bias is a directional measure of forecast error percentage error. A positive bias indicates the tendency to under forecast, whereas a negative bias indicates the tendency to over forecast. Bias is calculated as the $(\text{Actual}-\text{Forecast})/\text{Actual}$. **For period 38, it is $(90-110)/90$, which is -22.2%.**

Your Objective - *Minimize* your forecast error which leads to minimization of MAPE

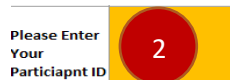
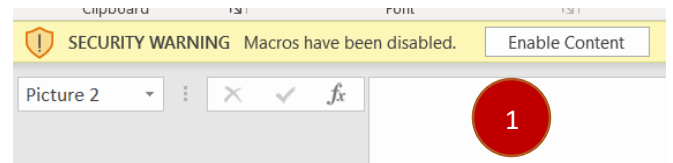
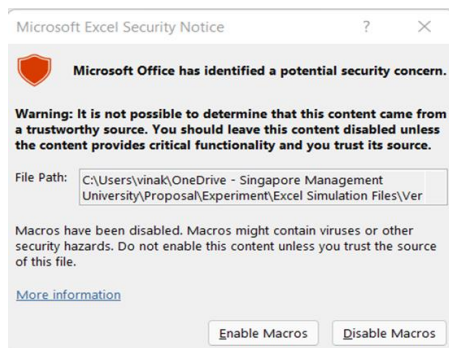
| | | | Forecast Measures | | | |
|-------------|---------------|----------|-------------------|--------|---------|--------|
| | | | 1 | 2 | 3 | 4 |
| Time Period | Actual Demand | Forecast | Forecast Error | APE(%) | MAPE(%) | Bias |
| T37 | 100 | 90 | 10 | 10.0% | 10.0% | 10.0% |
| T38 | 90 | 110 | 20 | 22.2% | 16.1% | -22.2% |
| T39 | 110 | 100 | 10 | 9.1% | 13.8% | 9.1% |

Reward

Participants will be paid **USD 2** for fully completed responses. In addition, there is a bonus reward of **USD 15** for the top **3** respondents with the lowest MAPE for the 36 periods.

Spreadsheet set up and overview

1. The spreadsheet uses macros. You may receive security warning messages similar to the one shown below. Please Enable Macros to use the spreadsheet
2. Please ensure you key in your Participant ID in **Cell B1** in the Tab -Spreadsheet Simulation
3. You are requested to complete the task in one sitting



Steps

The Actual Demand for the Product are shown in the column labeled Actual Demand (Column C in YELLOW).

1. Provide your forecast in the column labeled **"User Forecast"** (Column D in BLUE) from Row 37 onwards.

Note: i) enter values greater than 0, ii) text entries are not permitted, iii) forecasts once entered cannot be changed.

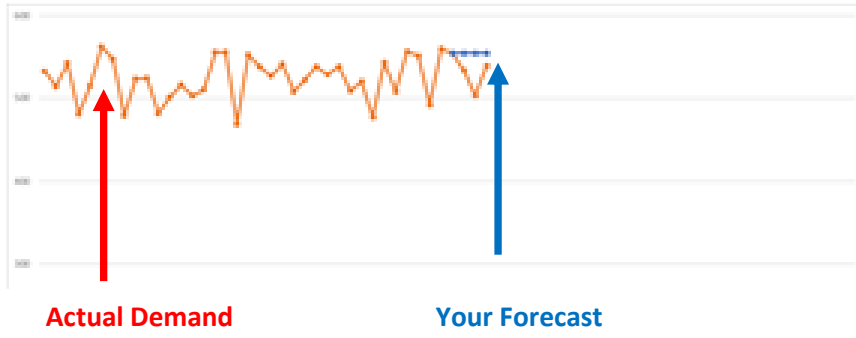
2. The actuals and the forecast are plotted in a line chart that depicts the actual demand and your forecast.
3. A plot of the forecast bias in the form of a bar chart is also presented. The bar chart indicates the magnitude and direction of forecast error (e.g., under forecasting or over-forecasting).
4. Once you have completed the submission for time period 72, the spreadsheet simulation task is complete.
5. YOU MUST THEN SAVE THE SPREADSHEET AND UPLOAD IT BACK TO THE SURVEY.
6. Proceed to complete the questionnaire in the survey. You will receive a completion code at the end of the questionnaire, which must be typed in **MTurk** to get credit for your effort.

| Time Period | Actual Sale | Your Forecast | Forecast Error | APE (%) | MAPE (%) | Mean Forecast Accuracy (100-MAPE) | Bias |
|-------------|-------------|---------------|----------------|---------|----------|-----------------------------------|-------|
| T37 | 555 | 500 | 55 | 9.9% | 9.9% | 90.1% | 11.0% |
| T38 | 534 | 555 | 21 | 3.9% | 3.9% | 93.1% | -3.8% |
| T39 | 503 | 555 | 52 | 10.3% | 10.3% | 91.9% | -9.4% |
| T40 | 540 | 555 | 15 | 2.8% | 6.7% | 93.3% | -2.7% |
| T41 | | | | | | | |

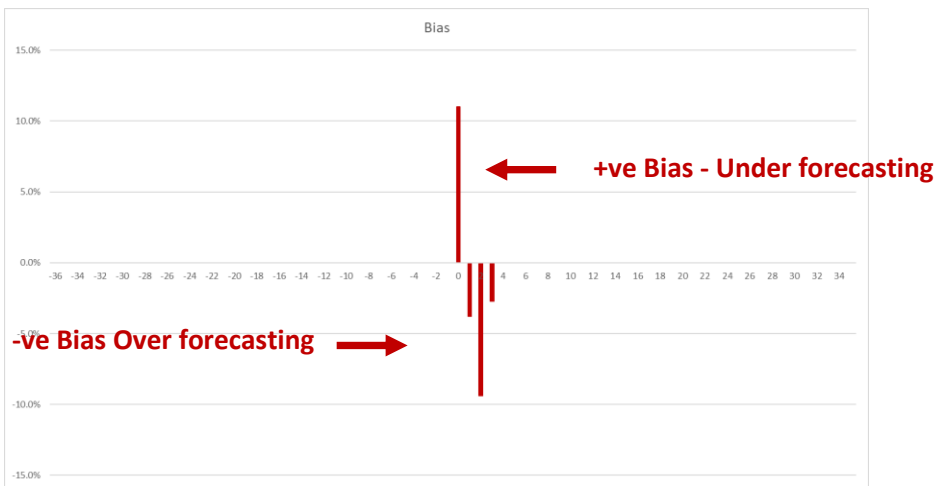
SAMPLE DATA



Key in your inputs here



Key in your inputs here



Appendix 2d

Instructions for Rolling Training

Judgmental Demand Forecasting

The following is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process.

Context: You are the demand planner for X-mart, a supermarket. Your role is to forecast the future demand for products based on actual past demand.

The demand forecast forms the input for subsequent activities such as procurement from the supplier and replenishment of the store. The demand forecast is the first step in the planning process, enabling product availability, and an accurate forecast leads to improved business performance.

You have been given the actual demand for the past 36 time periods. You will need to predict the subsequent 36 time periods one step ahead for the next selling period, i.e., forecast period 37, using data available through period 36. Actual demand for period 37 is then realized, and you will be asked to forecast period 38.

Forecast Performance Measures:

Note below the definition of the forecast performance measures. The table below shows an example calculation for the measures. The number in the red circle corresponds to the measure described below.

- Forecast error is the absolute difference between the actual demand and the forecast. **For time period 38, this is the absolute difference between 90 (actual) and 110 (forecast) and is 20.**
- Absolute Percentage Error (APE) is the forecast error expressed as a percentage of the actual demand. **For time period 38, this is calculated as (20/90), which is 22.2%**
- MAPE refers to the Mean of the absolute percentage error (APE) over the given time periods. **For period 38, this is the average of period 37 (10%) and period 38 (22.2%), which is 16.1%.**
- Bias is a directional measure of forecast error percentage error. A positive bias indicates the tendency to under forecast, whereas a negative bias indicates the tendency to over forecast. It is calculated as the (Forecast – Actual)/Actual. **For period 38, it is (110-90)/90, which is -22%.**

Your Objective - *Minimize* your forecast error which leads to minimization of MAPE

| | | | Forecast Measures | | | |
|-------------|---------------|----------|-------------------|--------|---------|--------|
| | | | 1 | 2 | 3 | 4 |
| Time Period | Actual Demand | Forecast | Forecast Error | APE(%) | MAPE(%) | Bias |
| T37 | 100 | 90 | 10 | 10.0% | 10.0% | 10.0% |
| T38 | 90 | 110 | 20 | 22.2% | 16.1% | -22.2% |
| T39 | 110 | 100 | 10 | 9.1% | 13.8% | 9.1% |

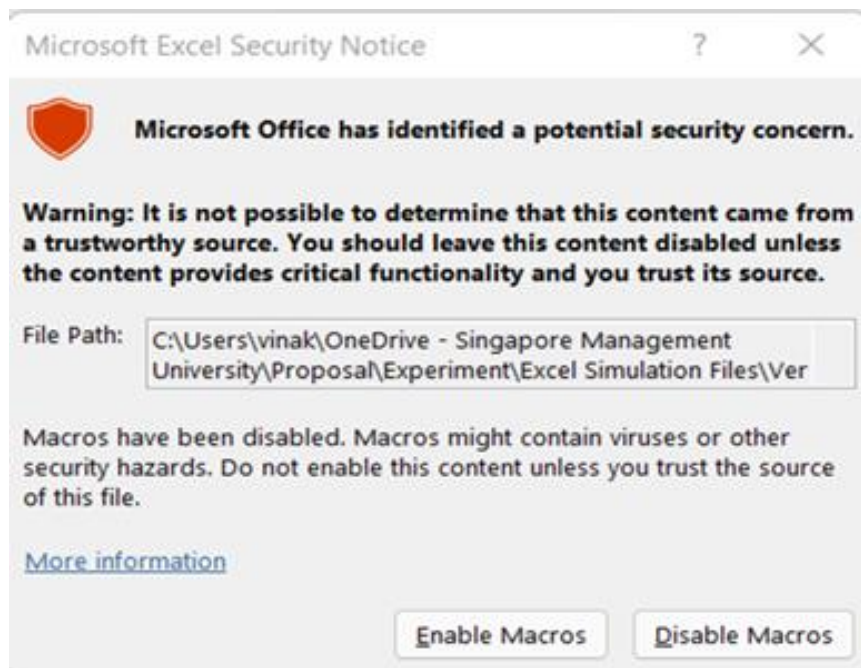
Reward

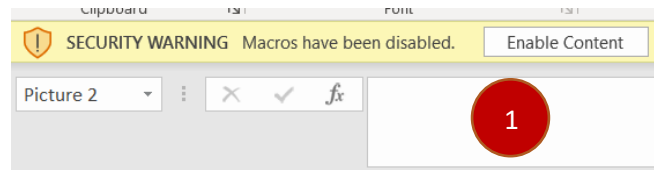
Participants will be paid **USD 2** for fully completed responses. In addition, there is a bonus reward of **USD 15** for the top **3** respondents with the lowest MAPE for the 36 periods.

Spreadsheet set up and overview

1. The spreadsheet uses macros. You may receive security warning messages similar to the one shown below. Please **Enable Macros** to use the spreadsheet
2. Please ensure you key in your Participant ID in **Cell B1** in the Tab -Spreadsheet Simulation

You are requested to complete the task in one sitting





Steps

The Actual Demand for the Product are Shown in the column labeled Actual Demand (Column C in **YELLOW**).

1. Provide your forecast in the column labeled "**User Forecast**" (Column D in **BLUE** from Row 37 onwards.
Note: i) enter values greater than 0, ii) text entries are not permitted, iii) forecasts once entered cannot be changed.
2. The actuals and the forecast are plotted in a line chart that depicts the actual demand and your forecast.
3. A plot of the forecast bias in the form of a bar chart is also presented. The bar chart indicates the magnitude and direction of forecast error (e.g., under forecasting or over-forecasting).
4. Once you have completed the submission for time period 72, the spreadsheet simulation task is complete.
5. **YOU MUST THEN SAVE THE SPREADSHEET AND UPLOAD IT BACK TO THE SURVEY.**

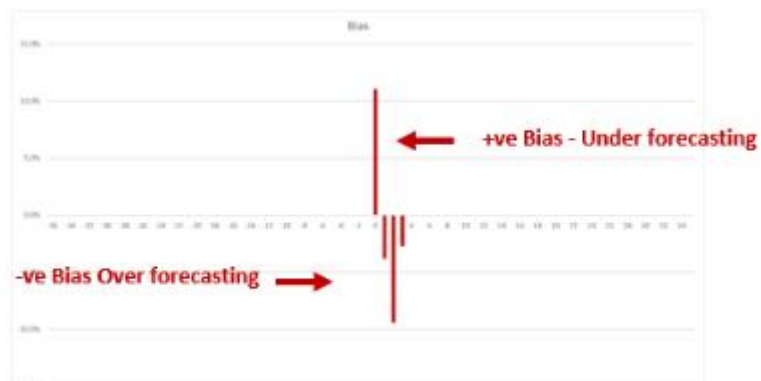
- Proceed to complete the questionnaire in the survey. You will receive a completion code at the end of the questionnaire, which must be typed in **MTurk** to get credit for your effort.

| Time Period | Actual Sale | Your Forecast | Forecast Error | APE (%) | MAPE (%) | Mean Forecast Accuracy (100-MAPE) | Bias |
|-------------|-------------|---------------|----------------|---------|----------|-----------------------------------|-------|
| T37 | 555 | 500 | 55 | 9.9% | 9.9% | 90.1% | 11.0% |
| T38 | 534 | 555 | 21 | 3.9% | 3.9% | 93.1% | -3.8% |
| T39 | 503 | 555 | 52 | 10.3% | 10.3% | 91.9% | -9.4% |
| T40 | 540 | 555 | 15 | 2.8% | 6.7% | 93.3% | -2.7% |
| T41 | | | | | | | |

Key in your inputs here



Key in your inputs here



Appendix 2e
Instructions for Decomposition Training

Judgmental Demand Forecasting

The following is an experiment on individual judgments. The experiment is not intended to test your knowledge but is a means to understand the individual decision-making process.

Context: You are the demand planner for X-mart, a supermarket. Your role is to forecast the future demand for products based on actual past demand.

The demand forecast forms the input for subsequent activities such as procurement from the supplier and replenishment of the store. The demand forecast is the first step in the planning process, enabling product availability, and an accurate forecast leads to improved business performance.

You have been given the actual demand for the past 36 time periods. You will need to predict the subsequent 36 time periods one step ahead for the next selling period, i.e., forecast period 37, using data available through period 36. Actual demand for period 37 is then realized, and you will be asked to forecast period 38.

Forecast Performance Measures:

Note below the definition of the forecast performance measures. The table below shows an example calculation for the measures. The number in the red circle corresponds to the measure described below.

1. Forecast error is the absolute difference between the actual demand and the forecast. ***For time period 38 this is the absolute difference between 90 (actual) and 110 (forecast) and is 20.***
2. Absolute Percentage Error (APE) is the forecast error expressed as a percentage of the actual demand. ***For time period 38 this is calculated as $(20/90)$ which is 22.1%***
3. MAPE refers to the Mean of the absolute percentage error (APE) over the given time periods. ***For period 38 this is the average of period 37 (10%) and period 38 (22.2%) which is 16.1%.***

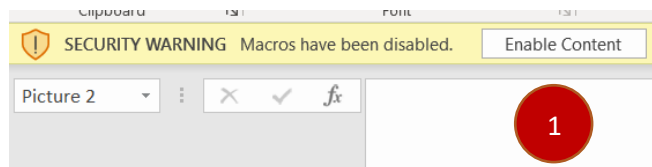
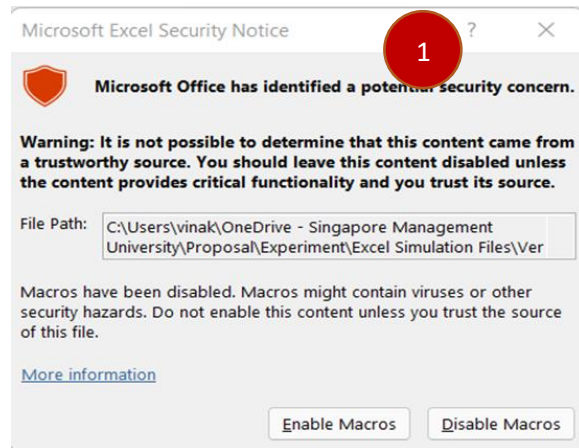
Your Objective - *Minimize your forecast error which leads to minimization of MAPE*

Reward

Participants will be paid **USD 2** for fully completed responses. In addition, there is a bonus reward of **USD 15** for the top **3** respondents with the lowest MAPE for the 36 periods.

Spreadsheet set up and overview

3. The spreadsheet uses macros. You may receive security warning messages similar to the one shown below. Please **Enable Macros** to use the spreadsheet
4. Please Ensure you key in your Participant ID in **Cell B1** in the Tab -Spreadsheet Simulation
5. You are requested to complete the task in one sitting



Steps

The Actual Demand for the Product is Shown in the column labelled Actual Demand (Column C in **YELLOW**).

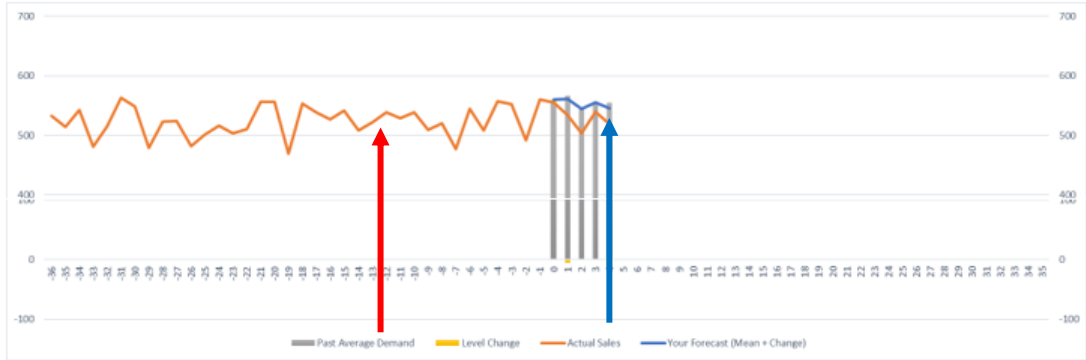
7. The forecast can be expressed as the combination of 3 components
 - i. The past average demand
 - ii. A level change to the past average demand (can be positive or negative)
 - iii. A noise component (can be positive or negative) – Noise being random in nature should be effectively ignored

8. Enter your estimates of past average demand and the change, the spreadsheet will then automatically calculate the total demand as the sum of average demand and the change and is populated in Column F in **Yellow**
9. Note: i) enter values greater than 0 for past average demand, ii) level changes can be positive or negative iii) text entries are not permitted, iii) forecasts once entered cannot be changed"
10. The performance metrics will be updated after every submission. The actuals, forecast components and the total forecast are also plotted in a graph that you are encouraged to review before your next submission.
11. Once you have completed submission for time period 72, the spreadsheet simulation task is complete.
12. YOU MUST THEN SAVE THE SPREADSHEET AND UPLOAD IT BACK IN THE SURVEY.
13. Proceed to complete the questionnaire in the survey. You will receive a completion code at the end of the questionnaire, which must be typed in **MTurk** to get credit for your effort.

| Serial Number | Week | Actual Sales | Past Average Demand | Level Change | Your Forecast (Mean +) |
|---------------|------|--------------|---------------------|--------------|------------------------|
| 37 | T37 | 555 | 500 | -15 | 485 |
| 38 | T38 | 534 | 490 | 25 | 515 |
| 39 | T39 | 500 | 505 | 10 | 515 |
| 40 | T40 | 519 | 500 | 15 | 515 |
| 41 | T41 | 519 | 480 | 35 | 515 |
| 42 | T42 | 548 | 470 | 45 | 515 |
| 43 | T43 | 491 | 495 | 20 | 515 |
| 44 | T44 | 487 | 485 | 30 | 515 |

SAMPLE DATA

Key in your inputs here



Actual Demand

Your Forecast

Appendix 3

List Of Figures

| Figure No | Description | Page No |
|------------------|--|----------------|
| Figure 1 | Typical Forecasting Process in a Firm | 2 |
| Figure 2 | Base Treatment Respondent Forecast Input Screen | 30 |
| Figure 3 | Base Treatment Graphs | 31 |
| Figure 4 | Fan Chart Graph | 32 |
| Figure 5 | Decomposition Treatment Respondent Forecast Input Screen | 33 |
| Figure 6 | Decomposition Treatment Graphs | 34 |
| Figure 7 | Bias Graph | 35 |
| Figure 8 | Rolling Treatment Respondent Forecast Input Screen | 36 |
| Figure 9 | Forecasting process | 43 |
| Figure 10 | Participant Response to the Nature of Demand | 47 |
| Figure 11 | Nesting and Multilevel Structure of the Study | 50 |
| Figure 12 | Fan Chart Questionnaire Response Analysis | 55 |
| Figure 13 | Comparison of Base vs. Decomposition | 62 |
| Figure 14 | Figure 14 Subject MAPE versus the usage of Graphs | 75 |

Appendix 4

List Of Tables

| Table No | Description | |
|-----------------|--|----|
| 1 | Table 1 Calculation of Forecast Measures | 19 |
| 2 | Listing of Experiment Conditions | 28 |
| 3 | Optimal Alpha by Condition | 28 |
| 4 | Number of Participants Per Treatment Condition and Demand Set | 38 |
| 5 | Two Tail T-Test Comparison of Optimal vs. Observed MAE | 45 |
| 6 | Subject Adjustment Scores Analysis by Observation | 46 |
| 7 | Split of Participants Perception of Nature of Demand as Stable or Stable with Noise based on demand condition. | 47 |
| 8a | Adjustment scores based on Nature of demand | 49 |
| 8b | Estimation Results based on Nature of demand | 49 |
| 9 | Interclass Estimates for Null Model | 52 |
| 10 | Wald test of Optimal Alpha vs. Subject Adj Score | 53 |
| 11 | Wald test of Treatment vs. Base Adjustment Score | 54 |
| 12 | Nested Models Fan chart estimation | 56 |
| 13 | Consolidated Nested Models Estimation | 58 |
| 14 | Wald test of Treatment vs. Base Adjustment Score | 59 |
| 15 | t-test Comparison of Participants Forecast with Decomposition Treatment and Base. | 60 |
| 16 | Nested Models Decomposition Estimation | 61 |
| 17 | Wald test of Treatment vs. Base Adjustment Score | 63 |
| 18 | Nested Models Bias Estimation | 64 |
| 19 | Percentage of Observations classified as Over-forecasting & Under-forecasting in Base and Bias Treatments | 64 |
| 20a | Bias Performance of Observations Classified as Over-forecasting from Bias and Base Treatment | 65 |
| 20b | Bias Performance of Observations Classified as Over-forecasting from Bias and Base Treatment | 65 |
| 21 | Wald test of Treatment vs. Base Adjustment Score | 67 |
| 22 | Nested Models Rolling Training Estimation | 68 |
| 23a | Wald test of Optimal Alpha vs. Subject Adj Score | 69 |
| 23b | Wald test of Base vs. Subject Adj Score | 69 |
| 24 | Estimation Results for Condition 4, including Parameter Knowledge Treatment | 70 |
| 25 | T-test of Observed MAPE vs Opt MAPE | 71 |
| 26 | T-test of MAPE Treatment vs. Base | 72 |
| 27 | Summary of Results | 76 |