IE UNIVERSIDAD

TESIS DOCTORAL/ DOCTORAL DISSERTATION

ENSAYOS SOBRE LA SABIDURÍA DE LOS GRUPOS

/ ESSAYS ON THE WISDOM OF CROWD

SHIJITH KUMAR PAYYADAKKATH MEETHALE

SEGOVIA, 2021

IE UNIVERSIDAD

TESIS DOCTORAL/ DOCTORAL DISSERTATION

ENSAYOS SOBRE LA SABIDURÍA DE LOS GRUPOS

/ ESSAYS ON THE WISDOM OF CROWD

SHIJITH KUMAR PAYYADAKKATH MEETHALE

Doctoral Thesis Advisor: Professor. MATTHIAS SEIFERT

Abstract

This dissertation consists of three essays. These essays together attempt to investigate on how to use different decision rules and aggregation mechanisms to improve collective judgment accuracy in wisdom of crowd applications.

The first chapter focusses on the application area of open innovation and compares the performance of commonly used decision rules (scoring and ranking rules) for idea evaluation & selection. The results suggest that the crowd wisdom extracted using the scoring rule outperforms the crowd wisdom derived using ranking rule in terms of the likelihood of selecting the highest-quality ideas. From a managerial perspective, this study provides guidance for the choice of idea evaluation process and thereby the design of open innovation initiatives.

The second and the third chapters focus on the application area of judgmental forecasting and investigate an important context, that of forecasting environments characterized by structural breaks. The second chapter demonstrates the presence of systematic biases (under and over forecasting) in forecast judgments under the presence of structural breaks. Further, this chapter proposes a novel and effective asymmetric trimming rule of forecast aggregation and prescribes two forecast ensemble methods to leverage such aggregated forecasts and improve forecast performance. Chapter 3 further tests the performance of the proposed forecast ensemble methods under extended conditions and suggests improvements to one of the proposed ensemble methods.

Overall, the dissertation aims to unfold decision rules and aggregation heuristics by demonstrating potential cognitive limitations and biases of the decision maker and further, prescribes aggregation mechanisms to improve decision quality.

Resumen

La presente tesis consta de tres ensayos. En conjunto, intentan investigar cómo emplear diferentes reglas de decisión y mecanismos de agregación para mejorar la precisión del juicio colectivo en las aplicaciones de la sabiduría de los grupos.

El primer capítulo se centra en el área de aplicación de la innovación abierta y compara el rendimiento de las reglas de decisión más utilizadas (reglas de puntuación y clasificación) para la evaluación y selección de ideas. Los resultados indican que la sabiduría de los grupos extraída mediante la regla de puntuación supera a la sabiduría de los grupos derivada mediante la regla de clasificación en cuanto a la probabilidad de seleccionar las ideas de mayor calidad. Desde el punto de vista de la gestión, este estudio proporciona orientación para la elección del proceso de evaluación de ideas y, por tanto, para el diseño de iniciativas de innovación abierta.

El segundo y el tercer capítulo se centran en el área de aplicación del pronóstico por juicio e investigan un contexto importante de entornos de pronóstico caracterizados por rupturas estructurales. El segundo capítulo demuestra la presencia de sesgos sistemáticos (infra y sobrepronóstico) en los juicios de pronóstico en presencia de rupturas estructurales. Además, este capítulo propone una regla de recorte asimétrica novedosa y eficaz de agregación de pronósticos y prescribe dos métodos de conjunto de pronósticos para aprovechar dichos pronósticos agregados para mejorar el rendimiento de los pronósticos. En el capítulo 3 se comprueba el rendimiento de los métodos de conjuntos de pronóstico

propuestos en condiciones ampliadas y se proponen mejoras en uno de los métodos de conjuntos propuestos.

El objetivo de la tesis es desplegar las reglas de decisión y la heurística de agregación demostrando las posibles limitaciones cognitivas y los sesgos del decisor y, además, prescribe mecanismos de agregación para mejorar la calidad de las decisiones.

Acknowledgement

Words cannot express my feelings, nor my thanks for all the support and help I have received during my PhD journey. Nevertheless, I would like to express gratitude from the bottom of my heart to all who have continuously inspired and supported me through this journey both professionally and personally.

I must start by expressing by deepest gratitude to my advisor, Professor. Matthias Seifert for handholding me through the most challenging times, instilling confidence in me through this journey and providing me with valuable guidance whenever it mattered. Thanks a ton for making me live through the value of two most important elements in life - *patience* and *perseverance*.

I am deeply indebted to Professor. Dilney Goncalves and Professor. Yun Shin Lee for their vital guidance and support in my studies for this dissertation. I would also like to extend by deepest gratitude to Professor. Jeeva Somasundaram and Professor. Shameek Sinha for their invaluable guidance towards my dissertation.

My journey in the PhD and this dissertation would not have been complete without the continuing support and nurturing from Dr. Laura Maguire. My heart goes all out to express my heartfelt gratitude for her sustained encouragement. I would also like to thank Ms. Maria Muriel De Alba for her kind patience in handling all my requests and providing all the support needed.

I would like to extend special thanks to my friends and colleagues. Thank you, Akhil, Claudio, David, Mohamad, Polina, Prana, and Sumit! Special thanks to my friend

and buddy, Sreyaa for her support and help through this journey. I would also like to extend my sincere thanks to my colleagues at Solbridge International School of Business.

I very much appreciate the contribution of all my teachers who have played an instrumental role at different stages in my life. I shall ever remain thankful for all those learnings which have helped me to pursue my dreams.

I would like to thank my beloved wife, Dr. Nidhina, who became an integral part of my life during my PhD journey, and from then on helped me sail through with unconditional personal support.

I wouldn’t have been able to pursue my dreams without the prayers of my mother and the wishes of my father. Nothing in this world is possible for me without their love and blessings. My brother, his wonderful wife, and their little princess (Ammu) have always been there for me, and this support is invaluable. Thanks a ton, to my family!

Finally, I would like to thank Almighty for blessing me with valuable support and timely help through all these wonderful people.

Table of Contents

List of Figures

List of Tables

Introduction

My scholarly goal is centered on understanding the effectiveness of different decision rules and aggregation mechanisms used to distil collective wisdom in different application contexts. Leveraging collective knowledge by relying on the wisdom of crowds (Surowiecki, 2004), a phenomenon first demonstrated by Galton (1907), has become an increasingly popular approach to find solutions to business problems. With rapid advancements in information technology, internet, and social media, we can observe that in the recent past, the 'crowd' (for example consisting of customers, users, domain experts, etc.) is playing a major role in making complex decisions for firms and other institutions - some prominent emerging application fields being - prediction polls and forecasting (Jose, et al, 2014; Tetlock & Gardner, 2016; Grushka-Cockayne, et al, 2017), prediction markets (Atanasov, et al, 2017), and open innovation and innovation tournaments (von Hippel, 2005; Terwiesch & Ulrich, 2009). Real world examples are in plenty - selection of T-shirt designs at Threadless (Malone, et al, 2010), Google's 10K project, IBM-innovation jam (Klien & Garcia, 2015), aggregation of macroeconomic forecasts at Central Banks (Mannes, Soll & Larrick, 2014) and Dell's Ideastorm (Bayus, 2013) being among many others.

Behavioral decision research identifies the collective judgments stream as an important and relevant area of research (Durbach & Montibeller, 2019). The benefits of collective judgments (i.e., crowd wisdom) are attributed to the independence and the diverse knowledgebase of the decision-making participants (i.e., predictors, forecasters, etc.) (Surowiecki, 2004; Davis-Stober, et al, 2014). Under increasing

use of crowd wisdom in various decision-making applications, recent research has emphasized the need for developing insights on decision rules (King & Lakhani, 2013; Kornish, et al, 2017). Research has also focused on extracting crowd wisdom by improving aggregation mechanisms (Larrick & Soll, 2006; Jose, et al, 2014; Prelec, et al, 2017; Palley & Soll, 2019) and crowd compositions (Mannes, et al, 2014; Budescu & Chen 2015). The effectiveness of different aggregation mechanisms and decision rules is contingent on the problem areas or application contexts (Kerr & Tindale, 2011). Different decision rules could impose different cognitive loads on the decision makers (Simon 1974, Hastie and Dawes 2001) and thereby impact individual judgments. The performance of such decision rules in collective judgment problem areas is potentially underexplored (King & Lakhani, 2013; Klein & Garcia, 2015). Decision rules employ different types of scales (eg., ordinal, cardinal, etc.) and recent research has found that the effectiveness of aggregation is contingent on the different forms of judgments (in the form of ranked preferences, numerical values, etc) elicited using these rules (Durbach & Montibeller, 2019). When we look at aggregation mechanisms, although the simple average is found to be effective in distilling collective wisdom both in theory and practice (Armstrong, 2001; Larrick & Soll, 2006), the presence of systematic biases in individual decisions can cause serious challenges in aggregation (Montibeller & Winterfeldt, 2018) and thereby limit the benefits of commonly used averaging techniques.

This dissertation aims to meet the call to study decision rules and aggregation mechanisms used in the collective judgment space. The main objective is to understand the impact of individual biases and/or cognitive limitations on collective judgments and propose prescriptive mechanisms to aggregate individual judgments. Therefore, my overarching research question is – 'How does decision rules and aggregation mechanisms affect collective judgment accuracy?'. To answer this question, I focus on two prominent application areas of collective wisdom - *innovation management* and *judgmental forecasting*.

My dissertation is composed of three essays. The first essay compares the performance of commonly used decision rules for idea evaluation & selection in the context of open innovation. The second essay aims to demonstrate systematic biases in judgmental forecasting of time series subject to structural breaks (focusing on structural mean shifts), to prescribe de-biasing trimming techniques for forecast aggregation and to propose a robust forecast ensemble mechanism. Finally, the third essay aims to test the proposed forecast ensemble mechanism further under extended break-frequency conditions and to also explore different approaches to elicit judgments about the occurrence of breaks in forecasting tasks. Together, all these three essays aim to unfold decision rules and aggregation heuristics by demonstrating potential cognitive limitations and biases of the decision maker and further, to provide insights into the use of decision rules, and to prescribe aggregation mechanisms to improve decision quality. Chapter 1 of this study aims to make descriptive and prescriptive contributions by highlighting the impact of different

decision rules used in idea evaluation on cognitive loads of the decision maker and by comparing the performance of these decision rules. The study further makes a prescriptive contribution to managers of open innovation contexts by providing insights on the design elements of an open innovation initiative. Chapter 2 of this dissertation aims to contribute to the descriptive research space by demonstrating systematic biases of the decision maker in judgmental time series forecasting contexts with structural breaks and by further demonstrating the implications of such biases on commonly used rules to extract crowd wisdom. Further, chapter 2 and chapter 3, also contribute to prescriptive research by prescribing potentially value adding de-biasing techniques for forecast aggregation and by proposing a robust forecast ensemble approach.

Chapter 1

Under this section, I focus on the application area of idea evaluation and selection in the context of open innovation, a context which has gained prominence in both research and practice (Terwiesch and Ulrich 2009; Bockstedt et al, 2016). There are a few results regarding the efficacies of decision rules used in idea evaluation processes. Using an online experiment, this study examines the efficacy of two idea evaluation decision rules: scoring vs. ranking. In the scoring rule, the evaluators were asked to rate the quality of each idea by assigning it a score (e.g., from 0 to 10), while in the ranking rule the evaluator simply ordered all ideas according to their perceived qualities. The results suggest that the crowd wisdom extracted using the scoring rule outperforms the crowd wisdom derived using ranking rule in terms of the

likelihood of selecting the highest-quality ideas. Moreover, the scoring rule attains higher accuracy with smaller crowd sizes. To better understand the efficacy of these decision rules, further analysis was carried out at the level of individual decision makers, and the results suggest that the scoring rule strictly outperforms the ranking rule in terms of the likelihood of selecting the highest-quality ideas. This result remains robust, irrespective of the possibility of allowing ties in the ranking rule. However, when the number of ideas to be evaluated is reduced from eight to three, the efficacies of the two rules do not differ. Based on the observations from the experimental data, an explanatory model in which the information becomes a cue that directs the participants' efforts to evaluate the ideas (i.e., the time taken for evaluation) is proposed and tested. From a managerial perspective, this study provides guidance for the choice of idea evaluation process and thereby the design of open innovation initiatives.

Chapter 2

Under this section, I focus on the application area of judgmental time series forecasting. Forecast combination literature has emphasized on the need for detailed studies on forecast aggregation and trimming rules in aggregation (Makridakis & Winkler, 1983; Clemen, 1989; Armstrong, 1989; Armstrong, 2001; Thomson, et al, 2019). While the extant literature explores aggregation mechanisms and trimming rules (Armstrong, 2001; Jose, et al, 2014) for forecasts in stable environments, the focus here is on relatively unstable environments characterized by fundamental structural shifts (Pesaran, et al, 2006; Aue & Horvath, 2013). A structural shift can

cause systematic biases and therefore, commonly used forecast combinations may not be as effective (Clements & Hendry, 1998; Hendry & Clements, 2004; Atiya, 2020) and this study focusses on this special case of time series under changes in the form of structural mean shifts. This study introduces a simple and a novel form of asymmetric trimming to aggregate judgmental time series forecasts. In an experimental study, it is found that forecasters are sensitive to structural shifts (mean shifts) in time series and are quite systematically biased in their forecasts based on the direction of the shifts, making simple averaging and static trimming approaches less applicable. This section proposes two ensemble forecasting methods incorporating asymmetrically trimmed forecasts that accounts for the systematic biases to aggregate judgmental forecasts under relatively unstable environments. Judgmental forecasting literature has often highlighted human judges' diagnostic ability to detect systematic patterns in time series environments (Simon, 1990; Lawrence, et al, 2006; Seifert et al. 2013). Therefore, in the first method, the 'Break Judgment Ensemble', exploits information about forecasters' subjective judgment regarding the occurrence of a structural break and these aggregated judgments are used as weights to form forecast ensembles. In the second ensemble method (referred to as the 'Past Performance Ensemble'), weights for forecast ensemble are derived from the ranked squared error performance of each trimmed forecasts in the time-period prior to the forecast period. The prescriptive Past Performance Ensemble method yields significant improvements in forecast accuracy measured in terms of Mean Absolute Percentage Error (MAPE). However, the performance of the Break Judgment Ensemble is not always the best. To further test the external validity,

the Past Performance Ensemble is tested on a set of time series from the real world and the findings support the robustness of the Past Performance Ensemble method. This study contributes to the forecast combination literature by bridging the two streams of judgmental forecasting and forecast aggregation. This essay, in particular, investigates time series environments characterized by structural breaks, an area, relatively unexplored by the judgmental forecast combination research. This essay contributes to descriptive research by demonstrating the presence of systematic biases associated with structural shifts and by further highlighting the potential effects of these biases in forecast aggregation. The value of asymmetric trimming under the influence of such systematic biases is emphasized and this essay also contributes to the prescriptive research on forecast aggregation by proposing the applicability of asymmetric trimming in forecast aggregation. In addition, this study contributes to prescriptive research by proposing a robust prescriptive mechanism of forecast ensemble for time series contexts characterized by structural shifts. The ensemble mechanism is found to be a prescriptively easy and an intuitively appealing approach for improving judgmental forecast performance.

Chapter 3

Under this section, I continue to study the area of judgmental forecast combination for time series characterized by structural shifts by testing the robustness of the Past Performance Ensemble method under extended frequency of structural breaks and by further exploring different break judgment elicitation approaches with an aim to improve the Break Judgment Ensemble method.

Judgmental forecasting literature has emphasized on the human judgement's overwhelming role in forecasting (O'Connor, 1993; Lawrence, et al, 2006; Fildes & Goodwin, 2007; Moritz, et al, 2014) primarily due to the diagnostic ability of human judgments to detect discontinuities and systematic patterns in data series (Simon, 1990; Lawrence, et al, 2006). However, the results from the experimental study in Chapter 2 found that break judgments were not highly accurate (similar to a study by O'Connor et al 1993). This section thus explores the question – 'how to improve judgmental break predictions and combine these judgmental break predictions into forecast ensembles to yield improved forecast performance?' This Chapter explores different approaches to elicit break judgments. Specifically, in the first approach the break judgment elicitation method used in Chapter 2 is extended by eliciting detailed probability judgments for break predictions from the forecaster. In the second approach, unlike the earlier break judgment aggregation method which was based on simple majority rule to derive weights, I explore a different aggregation rule to extract group wisdom. Recent research (for example, Prelec, et al, 2017; Palley & Soll, 2018) has highlighted the potential flaw in a majority-based approach in shared information contexts such as forecasting. The simple majority-based aggregation method of extracting the wisdom of crowd may not be as effective in such situations and I use a more recent approach known as the Surprisingly Popular Algorithm (Prelec, et al, 2017) in a different application area (judgmental forecasting of time series). Further, both these studies aim to prescribe forecast ensembles by leveraging these aggregated judgmental predictions.

The results in the studies in this chapter indicate that eliciting detailed probability judgments for break predictions (first approach) does not necessarily improve the break prediction accuracy and the performance of the Break Judgment Ensemble forecasts. However, the Break Judgment Ensemble using the surprisingly popular algorithm to aggregate break judgments (the second approach) yields improvement in forecast performance compared to both simple averaging and symmetric trimming methods of forecast aggregation. Nevertheless, the Past Performance Ensemble approach provides robust forecast performance improvements across all the treatments in both the studies. This chapter contributes to judgmental forecast aggregation by exploring different methods to aggregate judgments and further by testing the applicability of the Surprisingly Popular Algorithm in aggregation of judgments (regarding time series characteristics) in forecasting tasks. This chapter further contributes to prescriptive research by testing the robustness of the earlier proposed prescriptive forecast ensemble method under extended conditions.

A snapshot summarizing the overall dissertation is given below.

Main Research Question:
How does decision rules and aggregation mechanisms affect collective judgment accuracy?

Context: Innovation management

Essay 1: **Comparing Scoring and Ranking Decision Rules**

RQ*:* How do scoring and ranking decisions rules perform in terms of the likelihood of selecting the highest-quality ideas?

Method: Experimental studies involving industry experts and crowd-sourced participants

Finding*:* Scoring rule outperforms the ranking rule in terms of the likelihood of selecting the highest-quality ideas.

Context: Judgmental forecasting

Essay 2: **Forecast ensembles under uncertainty**

RQ*:* How does the occurrence of structural mean shifts affect judgmental forecasts? How does systematic biases in such forecasts affect forecast aggregation? *What de-biasing technique and forecast combination mechanism could yield improved forecast performance under such scenarios?*

Method*: Experimental studies involving crow-sourced participants and secondary data from real-world survey forecasts*

Finding*: Forecasts are systematically biased as an aftermath of a structural shift. Forecast aggregation using asymmetric trimming rules add value under systematic biases.Forecast ensemble involving asymmetric trim forecasts yields improved forecast performance.*

Essay 3: **Exploring different break judgment aggregation methods**

RQ*: How to elicit judgmental break predictions in time series forecasting characterised by structural breaks? How to leverage such aggregated break predictions in combining forecasts?*

Method*: Experimental studies involving crowd-sourced participants and laboratory setting*

Finding*: Detailed probability judgments does not necessarily improve break judgment accuracy and forecast performance. Aggregation of judgmental break predictions using the surprisingly popular algorithm improves the performance of forecast ensemble.*

Figure 1: Thesis Outline

Introducción

Mi objetivo académico se centra en comprender la eficacia de diferentes reglas de decisión y mecanismos de agregación utilizados para condensar la sabiduría colectiva en diferentes contextos de aplicación. Aprovechar el conocimiento colectivo confiando en la sabiduría de los grupos (Surowiecki, 2004) —un fenómeno demostrado por primera vez por Galton (1907)— se ha convertido en un enfoque cada vez más popular para encontrar soluciones a los problemas empresariales. Con los rápidos avances de la tecnología de la información, Internet y las redes sociales, podemos observar que, en los últimos tiempos, el "grupo" (por ejemplo, formado por clientes, usuarios, expertos del sector, etc.) está desempeñando un papel fundamental en la toma de decisiones complejas para las empresas y otras instituciones; algunos campos de aplicación emergentes destacados son: las encuestas de predicción y los pronósticos (Jose, et al, 2014; Tetlock y Gardner, 2016; Grushka-Cockayne, et al, 2017), los mercados de predicción (Atanasov, et al, 2017) y la innovación abierta y los torneos de innovación (von Hippel, 2005; Terwiesch y Ulrich, 2009). Los ejemplos del mundo real son abundantes: la selección de diseños de camisetas en Threadless (Malone, et al, 2010), el proyecto 10K de Google, la IBM Innovation Jam (Klien y García, 2015), la agregación de pronósticos macroeconómicos en los bancos centrales (Mannes, Soll y Larrick, 2014) y el Ideastorm de Dell (Bayus, 2013), entre muchos otros.

La investigación sobre la decisión conductual identifica la corriente de los juicios colectivos como un área de investigación importante y relevante (Durbach y

Montibeller, 2019). Los beneficios de los juicios colectivos (es decir, la sabiduría de los grupos) se atribuyen a la independencia y la base de conocimientos diversos de los participantes en la toma de decisiones (es decir, predictores, pronosticadores, etc.) (Surowiecki, 2004; Davis-Stober, et al, 2014). Con un creciente uso de la sabiduría de los grupos en varias aplicaciones de toma de decisiones, la investigación reciente ha insistido en la necesidad de desarrollar conocimientos sobre las reglas de decisión (King y Lakhani, 2013; Kornish, et al, 2017). La investigación se ha centrado también en extraer la sabiduría de los grupos mejorando los mecanismos de agregación (Larrick y Soll, 2006; Jose, et al, 2013; Prelec, et al, 2017; Palley y Soll, 2018) y las composiciones de los grupos (Mannes, et al, 2014; Budescu y Chen 2015). La eficacia de los diferentes mecanismos de agregación y reglas de decisión depende de las áreas de los problemas o los contextos de aplicación (Kerr y Tindale, 2011). La existencia de distintas reglas de decisión podría imponer diferentes cargas cognitivas a los responsables de la toma de decisiones (Simon 1974, Hastie y Dawes 2001) y, por tanto, afectar a los juicios individuales. El rendimiento de estas reglas de decisión en áreas de problemas de juicio colectivo está potencialmente poco explorado (King y Lakhani, 2013; Klein y García, 2015). Las reglas de decisión también emplean diferentes tipos de escalas (por ejemplo, ordinal, cardinal, etc.) y la investigación reciente ha comprobado que la eficacia de la agregación depende de las diferentes formas de juicios (en forma de preferencias clasificadas, valores numéricos, etc.) obtenidos utilizando estas reglas (Durbach y Montibeller, 2019). Cuando observamos los mecanismos de agregación, aunque el promedio simple se considere eficaz para condensar la

sabiduría colectiva tanto en la teoría como en la práctica (Armstrong, 2001; Larrick y Soll, 2006), la presencia de sesgos sistemáticos en los tomadores de decisiones individuales puede implicar desafíos considerables en la agregación (Montibeller y Winterfeldt, 2018) y limitar los beneficios de las técnicas de promedio comúnmente utilizadas.

Esta tesis pretende atender la necesidad de estudiar las reglas de decisión y los mecanismos de agregación utilizados en el contexto del juicio colectivo. El objetivo principal es comprender el impacto de los sesgos individuales y las limitaciones cognitivas en los juicios colectivos, y proponer mecanismos prescriptivos para agregar juicios individuales. Por lo tanto, mi pregunta de investigación general es: ¿cómo afectan las reglas de decisión y los mecanismos de agregación a la precisión de los juicios colectivos? Para responder a esta pregunta, me centro en dos destacadas áreas de aplicación de la sabiduría colectiva, a saber, la gestión de la innovación y el pronóstico de juicio.

Mi tesis se compone de tres ensayos. El primero compara el rendimiento de las reglas de decisión comúnmente utilizadas para la evaluación y selección de ideas en el contexto de la innovación abierta. El segundo pretende demostrar los sesgos sistemáticos en el pronóstico de juicio de las series temporales sujetas a rupturas estructurales (centrándose en los cambios de la media estructural), para prescribir técnicas de recorte para la reducción de sesgos en la agregación de pronósticos y proponer un mecanismo robusto de conjunto de pronósticos. Por último, el tercer ensayo tiene como objetivo probar el mecanismo de conjunto de pronósticos

propuesto bajo condiciones de frecuencia de ruptura extendida y también explorar diferentes enfoques para obtener juicios sobre la ocurrencia de rupturas en las tareas de pronóstico. En conjunto, estos tres ensayos pretenden desplegar las reglas de decisión y la heurística de agregación demostrando las posibles limitaciones cognitivas y los sesgos del responsable de la toma de decisiones y proporcionando más información sobre el uso de las reglas de decisión, así como prescribiendo mecanismos de agregación para mejorar la calidad de la decisión. El capítulo 1 de este estudio pretende hacer contribuciones descriptivas y prescriptivas, destacando el impacto de las diferentes reglas de decisión utilizadas en la evaluación de ideas sobre las cargas cognitivas del decisor y comparando el rendimiento de estas reglas de decisión. Además, el estudio hace una contribución prescriptiva a los gestores de contextos de innovación abierta en cuanto a los elementos de diseño del entorno de innovación abierta. El capítulo 2 de esta tesis tiene como objetivo contribuir al entorno de la investigación descriptiva demostrando los sesgos sistemáticos del tomador de decisiones en contextos de pronóstico de series temporales de juicio bajo rupturas estructurales y demuestra además las implicaciones de estos en diferentes enfoques a la hora de obtener la sabiduría de los grupos. Además, los capítulos 2 y 3 también contribuyen a la investigación prescriptiva al prescribir técnicas de reducción de sesgos con un potencial valor añadido para la agregación de pronósticos y proponiendo un enfoque robusto de conjunto de pronósticos.

Capítulo 1

En esta sección, me centro en el área de aplicación de la evaluación y selección de ideas en el contexto de la innovación abierta, un contexto que ha ganado importancia tanto en la investigación como en la práctica (Terwiesch y Ulrich 2009; Bockstedt et al, 2016,). Existen pocos resultados sobre la eficacia de las reglas de decisión utilizadas en los procesos de evaluación de ideas. Mediante un experimento en línea, el presente estudio examina la eficacia de dos reglas de decisión para la evaluación de ideas: puntuación frente a clasificación. En la regla de puntuación, se pidió a los evaluadores que calificaran la calidad de cada idea asignándole una puntuación (por ejemplo, de 0 a 10), mientras que en la regla de clasificación el evaluador simplemente ordenó todas las ideas según sus cualidades percibidas. Los resultados indican que la sabiduría de los grupos obtenida mediante la regla de puntuación supera a la sabiduría de los grupos derivada mediante la regla de clasificación en lo que respecta a la probabilidad de seleccionar las ideas de mayor calidad. Además, la regla de puntuación alcanza una mayor precisión con tamaños de público más pequeños. Para comprender mejor la eficacia de estas reglas de decisión, se lleva a cabo un análisis más profundo a nivel de los responsables de la toma de decisiones individuales, y se comprueba que la regla de puntuación supera estrictamente a la regla de clasificación en cuanto a la probabilidad de seleccionar las ideas de mayor calidad. Este resultado sigue siendo sólido, independientemente de la posibilidad de permitir los empates en la regla de clasificación. Sin embargo, cuando el número de ideas que evaluar se reduce de ocho a tres, las eficacias de las dos reglas no difieren. A partir de las observaciones de los datos experimentales, se propone y comprueba un modelo explicativo en el

que la información se convierte en una pista que dirige el esfuerzo de los participantes en la evaluación de las ideas (es decir, el tiempo que se tarda en evaluar). Desde el punto de vista de los gestores, este estudio proporciona orientación para la elección del proceso de evaluación de ideas y, por tanto, para el diseño de iniciativas de innovación abierta.

Capítulo 2

En esta sección, me centro en el área de aplicación del pronóstico de juicio de series temporales. La literatura sobre la combinación de pronósticos ha hecho hincapié en la necesidad de realizar estudios detallados sobre la agregación de pronósticos y las reglas de recorte en la agregación (Makridakis y Winkler, 1983; Clemen, 1989; Armstrong, 1989; Armstrong, 2001; Thomson, et al, 2019). Si bien la literatura existente explora los mecanismos de agregación y las reglas de recorte (Jose, et al, 2014; Armstrong, 2001) para los pronósticos en entornos estables, el foco se desplaza aquí a entornos relativamente inestables caracterizados por cambios estructurales fundamentales (Pesaran, et al, 2006; Aue y Horvath, 2012). Un cambio estructural puede causar sesgos sistemáticos y las combinaciones de pronóstico comúnmente utilizadas pueden ser ineficaces (Clements y Hendry, 1998; Hendry y Clements, 2004; Atiya, 2020), y este estudio se centra en este caso especial de series temporales con cambios en forma de cambios estructurales de la media. El presente estudio introduce un enfoque de recorte sencillo y novedoso (recorte asimétrico) para agregar pronósticos de juicio de series temporales. En un estudio experimental, se descubre que los pronosticadores son sensibles a dichos cambios

(desplazamientos de la media) en las series temporales y tienen un sesgo bastante sistemático en sus pronósticos en función de la dirección de los cambios, lo que hace que los enfoques de promediación simple y recorte estático sean menos aplicables. En esta sección se proponen dos métodos de pronóstico conjunto que incorporan pronósticos recortados asimétricamente y que tienen en cuenta los sesgos sistemáticos para la agregación de pronósticos de juicio en entornos relativamente inestables. La literatura sobre pronósticos basados en juicios ha destacado a menudo la capacidad de diagnóstico de los jueces humanos para detectar patrones sistemáticos en entornos de series temporales (Simon, 1990; Lawrence, et al, 2006; Seifert et al. 2013). Por lo tanto, en el primer método (el "Conjunto de Juicios de Ruptura") se explota la información sobre el juicio subjetivo de los pronosticadores con respecto a la ocurrencia y el tipo de ruptura estructural, y estos juicios agregados se usan como ponderaciones para formar conjuntos de pronósticos. En el segundo método de conjuntos (denominado "Conjunto de rendimiento pasado"), las ponderaciones para el conjunto de pronósticos se derivan del rendimiento del error cuadrático clasificado de cada pronóstico recortado en el periodo de tiempo anterior al periodo de pronóstico. El método prescriptivo de Conjunto de Rendimiento Pasado produce mejoras significativas en la precisión de los pronósticos, medida en términos de error porcentual absoluto medio (MAPE, por sus siglas en inglés). Sin embargo, el rendimiento del Conjunto de Juicios de Ruptura no es siempre el mejor. Para reforzar la comprobación de la validez externa, el método de Conjunto de Rendimiento pasado se pone a prueba con un conjunto de series temporales del mundo real y los resultados respaldan la solidez del método

de conjunto de rendimiento pasado. Este estudio contribuye a la literatura de la combinación de pronósticos al tender un puente entre dos corrientes: el pronóstico basado en el juicio y la agregación de pronósticos. Concretamente, el presente estudio investiga entornos de series temporales caracterizadas por rupturas estructurales, un área relativamente inexplorada en la investigación sobre la combinación de pronósticos de juicio. Este capítulo contribuye a la investigación descriptiva demostrando la presencia de sesgos sistemáticos asociados a los cambios estructurales y destacando además los efectos potenciales de estos sesgos en la agregación de pronósticos. Se destaca el valor del recorte asimétrico bajo la influencia de dichos sesgos sistemáticos y este ensayo contribuye a la investigación prescriptiva sobre la agregación de pronósticos al proponer la aplicabilidad del recorte asimétrico en la agregación de pronósticos. Además, este estudio contribuye a la investigación prescriptiva en el terreno de la agregación de pronósticos al proponer un mecanismo prescriptivo robusto de conjunto de pronósticos para contextos de series temporales caracterizados por cambios estructurales. Se advierte que el mecanismo de conjuntos es un enfoque prescriptivo fácil e intuitivamente atractivo para mejorar el rendimiento de los pronósticos basados en juicios.

Capítulo 3

En esta sección, continúo estudiando el área de la combinación de pronósticos de juicio para series temporales caracterizadas por cambios estructurales, probando la solidez del método de Conjunto de Rendimiento pasado bajo una frecuencia

ampliada de rupturas estructurales y, además, exploro diferentes enfoques de obtención de juicios de ruptura con el objetivo de mejorar el método de Conjunto de Juicios de Ruptura.

La literatura de pronósticos de juicio ha subrayado el papel dominante del juicio humano en el pronóstico (O'Connor, 1993; Lawrence, et al, 2006; Fildes y Goodwin, 2007; Moritz, et al, 2014), principalmente debido a la capacidad de diagnóstico de los juicios humanos para detectar discontinuidades y patrones sistemáticos en las series de datos (Simon, 1990; Lawrence, et al, 2006). Sin embargo, los resultados del estudio experimental del capítulo 2 revelaron que los juicios de ruptura no eran muy precisos (de forma similar a un estudio de O'Connor et al 1993). En esta sección se explora, pues, la cuestión de cómo mejorar las predicciones de juicios de ruptura y combinar estas predicciones de juicio de ruptura en conjuntos de pronósticos para obtener un mejor rendimiento del pronóstico. Este capítulo explora diferentes enfoques para obtener juicios de ruptura. En concreto, en el primer enfoque se amplía el método de obtención de juicios de ruptura utilizado en el Capítulo 2, obteniendo juicios de probabilidad detallados para las predicciones de ruptura del pronosticador. En el segundo enfoque, a diferencia del método anterior de agregación de juicios de ruptura que se basaba en una simple regla de mayoría para derivar las ponderaciones, este estudio explora una regla de agregación diferente para extraer la sabiduría del grupo. Investigaciones recientes (por ejemplo, Prelec, et al, 2017; Palley y Soll, 2018) han destacado el posible defecto de un enfoque basado en la mayoría en el caso de contextos de información compartida como el

pronóstico. El simple método de agregación basado en la mayoría para extraer la sabiduría de los grupos puede ser menos eficaz en tales situaciones, de modo que utilizo un enfoque más reciente conocido como el algoritmo sorprendentemente popular (Prelec, et al, 2017) en un área de aplicación diferente (pronóstico de juicio de series temporales). Además, estos dos estudios pretenden prescribir conjuntos de pronósticos aprovechando estas predicciones de juicio agregadas.

Los resultados de los estudios de este capítulo indican que la obtención de juicios de probabilidad detallados para las predicciones de ruptura (primer enfoque) no mejora necesariamente la precisión de la predicción de ruptura y el rendimiento de los pronósticos del Conjunto de Juicios de Ruptura. Sin embargo, el Conjunto de Juicios de Ruptura que utiliza el algoritmo sorprendentemente popular para agregar juicios de ruptura (el segundo enfoque) produce una mejora en el rendimiento del pronóstico en comparación con los métodos de agregación de pronósticos de promedio simple y recorte simétrico. No obstante, el enfoque del Conjunto de Rendimiento Pasado proporciona mejoras sólidas en el rendimiento de los pronósticos en todos los tratamientos de ambos estudios. Este capítulo contribuye a la agregación de pronósticos de juicio mediante la exploración de diferentes métodos de agregación de juicios y la comprobación de la aplicabilidad del algoritmo sorprendentemente popular en la agregación de juicios en lo tocante a ciertas características de las series temporales en las tareas de pronóstico. Este capítulo contribuye a la investigación prescriptiva probando la solidez del método de conjunto de pronósticos prescriptivo propuesto anteriormente en condiciones ampliadas.

A continuación, se presenta un resumen de la tesis general.

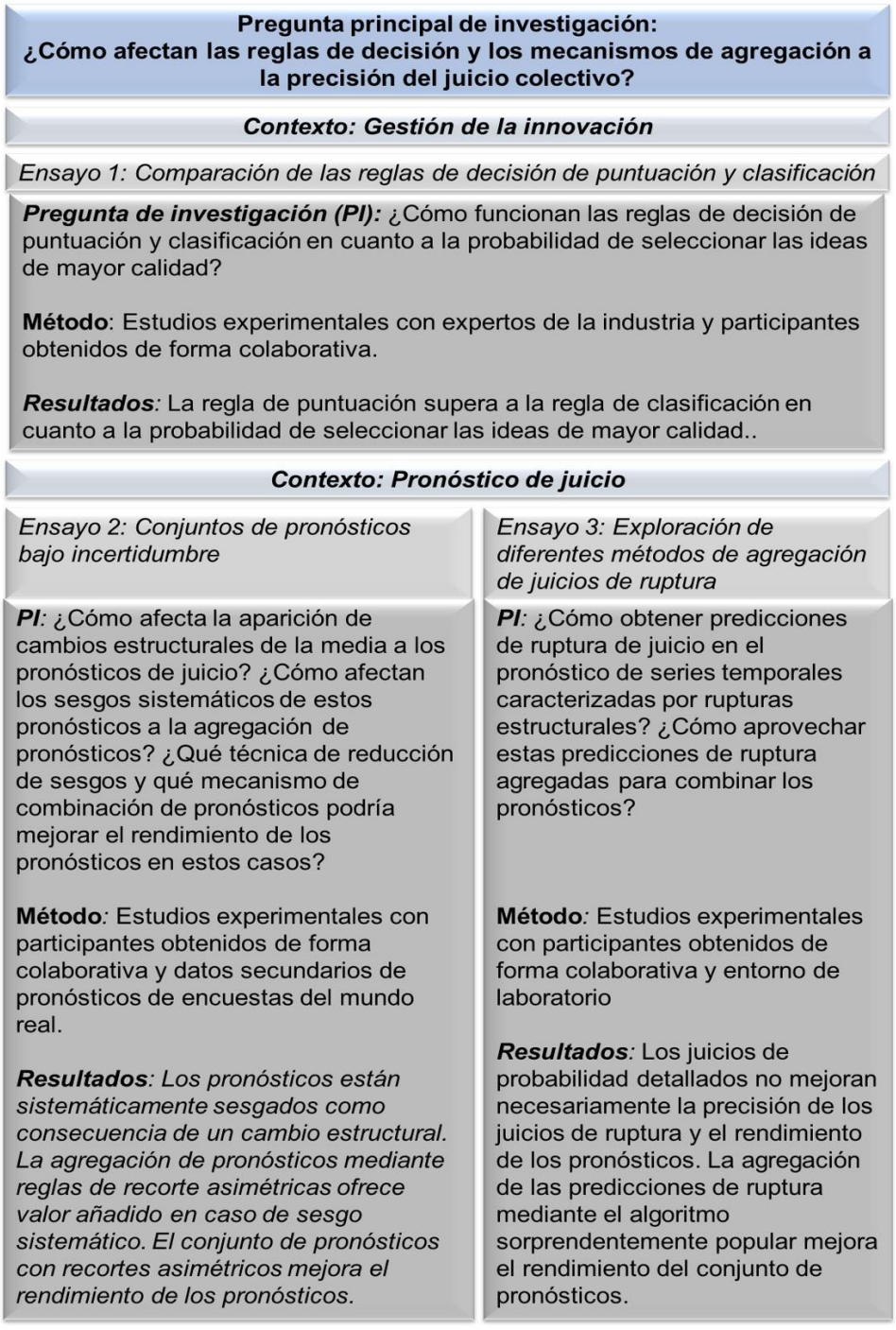


Figure 2: Esquema de la tesis

1 Chapter 1: Scoring vs. Ranking - An Experimental Study of Idea Evaluation Processes

1.1 Introduction

The concept of open innovation has gained prominence in both research and practice (Chesbrough 2006, Terwiesch and Xu 2008, Terwiesch and Ulrich 2009, Bockstedt et al, 2016) and has become the source for thousands of creative ideas (Girotra et. al. 2010, Poetz and Schreier 2012). However, firms cannot pursue all ideas generated. Instead, only a small portion of the most promising ideas are selected after being evaluated (Terwiesch and Xu 2008, Terwiesch and Ulrich 2009). For example, in the entertainment industry, of 300 “idea scratches”, only 5–6 survive and are commercialized into films (Terwiesch and Ulrich 2009). Other research finds that, at most, 10–30% of the ideas from open innovation engagements are eventually considered by firms (Klein and Garcia 2015).

In practice, traditionally the attention of idea evaluation has been directed towards accessing a larger pool of experts with relevant expertise, which could be quite costly. For example, to bring the “optimal” technological solutions to market, over 50 senior executives at IBM had to spend several weeks of their time to screen thousands of online entries by its open innovation participants (Bjelland & Wood, 2008). Similarly, Google had to engage “more than 3,000 employees to filter the unexpected surge of ideas for its 10 to the 100th project” (Klein & Convertino, 2014, P. 2; Buskirk, 2010). With the surge of online open platforms and communities, managers in different

industries have found that using their innovation communities to evaluate and select the best ideas may solve some of their problems associated with high costs of idea evaluation and selection (King & Lakhani, 2013). The online communities evaluate ideas using different decision rules such as rating or scoring, ranking, voting, etc. Under increasing use of collective wisdom in different decision-making applications, recent research has emphasized the need for developing insights on decision rules used for idea evaluation and selection processes (Kornish and Hutchison-Krupat, 2017; King & Lakhani, 2013). Different decision rules could impose different cognitive loads on the decision makers (Simon 1974, Hastie and Dawes 2001) and thereby potentially impact individual judgments. The performance of such decision rules in collective judgment problem areas is underexplored (King & Lakhani, 2013; Klein & Garcia, 2015). Recent behavioral decision research also indicates that collective wisdom is contingent on judgments elicited using different decision rules (such as ranked preferences, numerical ratings, etc.) (Durbach & Montibeller, 2019). Two decision rules are often observed in practice in the idea evaluation & selection application area, viz., scoring and ranking. Under the scoring rule, the evaluators are asked to rate the quality of each idea by assigning it a score (for example, a number from 0 to 10). Next, the scores provided by the evaluators are aggregated, and the ideas that receive the highest scores are selected. The scoring rule has been widely applied in project management and research proposal evaluation (Dahan and Hauser 2002, Toubia and Florès 2007, Dahan et al. 2010). In contrast with scoring, the ranking rule does not require the evaluator to assign a score to each specific idea. Instead, the evaluator simply orders all ideas according to their perceived

qualities. Then, the rankings made by the evaluators are aggregated, and the ideas ranked at the top of the list are selected. Both the decision rules are popular in practice. For example, in its weekly design contests, Threadless decides which T-shirt designs to produce by selecting the entries that receive the highest overall scores (Malone et al. 2010). In another open innovation contest called “Osram LED—Emotionalize Your Light”, users were asked to rank order a set of ideas or design entries (Cruz-Cunha 2012).

Unfortunately, most existing research on idea evaluation has thus far been directed toward achieving access to a larger pool of experts with relevant expertise (Kornish and Hutchison-Krupat, 2017), and few studies have investigated the efficacies of the different decision rules or processes being used to evaluate new ideas by online communities (King and Lakhani 2013). Consequently, firms often lack guidance on the use of decision rules for the design of idea evaluation processes. A process itself does not depend on individual experts and therefore could yield more robust evaluation accuracy in different situations. Understanding the efficacy of different evaluation rules is therefore crucial for extracting collective wisdom and thereby the success of open innovation.

Towards this end, an online experiment was designed to compare the efficacies of the scoring and ranking processes for idea evaluation. Specifically, the study began by building a pool of eight ideas drawn from the crowdfunding platform KickStarter.com. All eight ideas were new ideas and/or designs in the healthcare sector. With an approach similar to those of numerous existing studies (Ozer 2009, Girotra et. al. 2010, Poetz and Schreier 2012, Putman and Paulus 2009, Kornish and

Ulrich, 2014), this study used a panel of experts (from health care sector) to evaluate the qualities of these ideas and used the experts' evaluations as the "benchmark" for each idea's quality. Then, participants from Amazon Mechanical Turk were recruited to evaluate the ideas under different treatments (e.g., different evaluation processes and different manners of information presentation). Further, the efficacies of both idea evaluation processes were compared vis-à-vis the benchmarks.

Due to the limited results in the existing literature, this study is exploratory by nature. The experimental results suggest that the scoring process strictly outperforms the ranking process in terms of the likelihood of selecting the highest-quality ideas (both at an aggregated and individual level) regardless of how much information about each idea is presented to the participants. This result remains robust when the possibility of making a tie is allowed in the ranking process (this condition is henceforth referred to as "weak order ranking"). Further, it is seen that the scoring rule attains higher levels of accuracy with a relatively smaller crowd size when compared to the ranking rule. Additionally, it is also found that when the number of ideas to be evaluated is reduced from eight to three, the efficacies of the two idea evaluation processes become similar. In contrast to the prediction from the information processing theory that adding more information helps improve the efficacy of the scoring but not the ranking process, the experimental results suggest that the efficacy of the ranking process is improved when additional information is provided, yet the efficacy of the scoring process does not change. To explain these findings, this study proposes a path-dependent theoretical model in which the information does not directly influence the evaluation accuracy; instead, it becomes

a cognitive cue that directs the participants' efforts to evaluate the ideas (i.e., the time taken for evaluation). This model is empirically supported by the experimental data.

1.2 Literature Review and Theoretical Background

Gains from the parallel search efforts of a large number of participants have made innovation contests an effective approach to generate high quality solutions to innovation challenges (Terwiesch & Xu 2008; Terwiesch & Ulrich, 2009). Thus, various studies have paid attention to the process of innovation in the context of innovation tournaments or contests. Two streams of research can be observed. The first stream of study has focused on the innovation community per se. This stream studies the characteristics of the community participants (such as submission behavior, demographics, cultural background, etc) and participants' probability of success in the contests (Terwiesch and Xu, 2008; Bayus, 2013; Bockstedt, et al 2015; Bockstedt, et al 2016). Another stream of study has looked at contest design elements from a contest manager or organizer's perspective by analyzing the contest as a process (Boudreau, et al 2011; Erat & Krishnan, 2012; Wooten and Ulrich, 2015; Erat, 2017). The focus of these studies has been to demonstrate how different process elements influence expected contest outcomes such as degree of participation, quality of ideas, etc. Both these streams have focused on the contest's idea generation phase and this essay focuses on idea evaluation phase.

The extant innovation management literature has mainly viewed idea evaluation as a prediction task and aimed to answer two questions: (1) What should be asked? (2)

Who should be asked (Kornish & Hutchison-Krupat, 2017)? Specifically, rich results have been found concerning the design of idea evaluation criteria (Ulrich and Eppinger, 2015) and the evaluation performance of different types of participants (Kornish and Ulrich, 2014). Very few studies have examined how different methods of implementing the evaluation criteria influence evaluation efficacy. For example, Wilson and Schooler (1991) demonstrated that forcing the participants to rate ideas explicitly according to certain criteria may yield worse results than when the judgments are made holistically based on an aggregated feeling. The underlying mechanism is that when a subject is asked to decompose an idea into specific dimensions, their perception of an idea may change for the worse.

Many existing studies have extensively examined the idea generation process as well as the relationship between idea generation and evaluation under different team structures. With respect to the idea generation process, Wooten and Ulrich (2017) used a set of field experiments with online contests to examine the effect of in-process feedback on idea generation. Their results show that directed feedback is positively associated with agent participation as well as the average quality of entries submitted. In another recent study, Erat (2017) formulated a model of multistage development of idea pools and demonstrated that the dispersion of ideas in the pool has a positive impact on the value derived from the top ideas, especially when learning across stages is limited. With respect to the relationship between idea generation and evaluation, mixed results have been found. For example, in an experimental setting, Putman & Paulus (2009) examined the performance of two team structures (nominal vs. interactive) when the participants conducted both idea

generation and evaluation. In a nominal team, the members generated and evaluated ideas individually, whereas in interactive groups, the members created and evaluated ideas collectively. These authors found that the average originality of the selected ideas that received the best evaluations was higher for nominal groups than the interactive groups. However, as assessed by independent idea quality evaluators, participants under both team structures rarely selected their best ideas (Putman & Paulus 2009). Faure (2004) and Rietzschel et al. (2010) conducted similar studies and found that nominal teams are better than interactive teams in terms of idea generation. However, they did not find a significant difference between the interactive and nominal teams in terms of the quality of the selected ideas. These findings imply that the potential of selecting better ideas is not (fully) realized in the idea evaluation stage. In contrast with the aforementioned studies, this study does not aim to examine the idea generation process; instead, it focuses on the idea evaluation stage. Additionally, there are quite limited insights in the innovation management literature on the comparisons and efficacies of different idea evaluation processes or decision rules.

The existing literature in decision analysis has compared the rating (scoring) and ranking processes based on bounded rationality. Under the perfect rationality assumption in which each decision maker comprehends and processes complete information about the alternatives and selects the alternative that maximizes the expected utility, both processes are expected to produce similar outcomes, i.e., these two processes can be considered interchangeable (Klein and Garcia 2015).

However, many experimental studies have demonstrated that the consistency between these two processes is relatively low (Moore 1975, Russell and Gray 1994).

With respect to the efficacy of these two evaluation processes, some studies have argued that the ranking process should outperform the rating (scoring) process due to two reasons. First, despite its simplicity, the rating (scoring) process tends to elicit primarily average scores from evaluators, which is often called as the “close-to-average bias” of the evaluators (Klein and Garcia 2015). As a result, the scoring may tend to do a poor job of distinguishing between good and excellent ideas. For example, a “nice” evaluator may give rather close scores to several ideas that may have quite different intrinsic quality. In this case, the scoring process may blur the difference between a truly extraordinary idea and some of the middling alternatives. In contrast, the ranking process forces participants to provide relative rankings of ideas, creating a sharp distinction between “good” and “bad” ideas and alleviating the close-to-average bias and the rating lock. Second, the ranking process requires a respondent to pay attention to all items together while making the selection decision, whereas in the rating process, the respondent is focused on one item at a time. Consequently, the ranking process forces decision makers to consider more information, in a holistic sense and could lead to better evaluation accuracy (Alwin and Krosnick 1985). In a related study, Harzing et al. (2009) found the ranking mechanism to be superior in evaluating cross-cultural values. These authors found that the ranked responses (with the stimuli presented in a concise manner) conformed more closely to their benchmark than the scored or rated responses.

In contrast, another school of thought predicts better performance from the scoring process. First, the scoring process is designed in such a manner that an evaluator focuses her attention on appraising the quality of only one idea at a time. In contrast, in the ranking process, the evaluator's attention cannot be focused on any single idea. Consequently, the ranking process is a much more complex and demanding task than the scoring process, and the processing and transforming of information to arrive at an evaluation decision thus becomes more difficult and stressful in the ranking process (Baddeley 1992; Medin et al. 2004), which in turn leads to mistakes in the evaluation due to incomplete transformation of information (Bettman and Kakkar 1977). Second, the scoring process may not necessarily perform worse than the ranking process in terms of reducing the close-to-average bias. Several related studies have found that people who must choose from many alternatives tend to utilize simplifying strategies to reduce the cognitive complexity of the decision (e.g., Simon 1974, Hastie and Dawes 2001). With these simplifying strategies, people often engage in pre-choice screening of the available alternatives (Beach 1993) to make more cognitive resources available for careful consideration of the remaining options (Parks and Cowlin 1995). As a result, the ranking of ideas that the evaluators propose may simply represent the evaluators' pre-ranking preferences, which could undermine the efficacy of the ranking process. The fine-grained scores on the interval scale in a scoring process might overcome such possible mistakes that could potentially arise in a coarser ranking process.

This study also considers the potential impact of providing additional information about an idea to the evaluators under different evaluation processes. An idea can

be viewed as a piece or chunk of information represented as a single meaningful item (Simon 1974, Gobet & Simon, 1996), and the number of chunks (i.e., ideas) that can be processed within the constraints of working memory can influence the effectiveness of information processing (Miller 1956; Simon 1974). From this theoretical perspective, the inherent nature of the ranking process would inevitably increase the cognitive loads of the evaluators, which in turn reduces the "slack" resources available to process additional information (Russo et al. 1975, Medin et al. 2004). Consequently, the provision of additional information about ideas should be ineffective in the ranking process. In contrast, under the scoring process, the evaluator's attention is focused on one idea at a time. In this case, additional information does not increase the cognitive load as much as it does under the ranking process and therefore could be potentially useful in helping the participants improve their evaluation accuracy. However, the extant research has not examined the validity of this prediction.

Another factor that has been shown to be positively related with evaluation accuracy is the evaluators' efforts to evaluate the ideas (Einhorn and Hogarth 1981, Luce et al. 1997). Luce et al. (1997) highlighted the tradeoff between decision speed and accuracy in difficult decisions. In the context of idea evaluation, the overall effects of the induced efforts on the evaluation accuracy need to be jointly considered with the specific evaluation process. For example, the ranking process naturally forces participants to expend greater effort on digesting and processing the information, but additional effort must be used to process the additional complexity embedded in the ranking process. In contrast, the scoring process induces less effort

from participants but has a lower demand for effort to process information due to its simplicity and due to the focus on each specific idea. As a result, it remains unclear how the effort exerted by the evaluators interacts with the evaluation process and how they may jointly influence the evaluation accuracy.

Overall, some conflicting arguments and results have been found in the extant literature. However, with limited conclusive results, it is difficult to structure some hypotheses. Instead, this study aims to experimentally explore these questions. By nature, this study is data-driven, and it proposes a theoretical framework and tests it *ex post* while presenting the empirical findings.

1.3 Experimental Design & Measurement

1.3.1 Experimental Design

Participants from Amazon Mechanical Turk were recruited to participate in a web-based experiment using Qualtrics. This experiment was comprised of three stages: (1) idea evaluation, (2) collecting information regarding the participants' characteristics, and (3) collecting experts' opinions. The web-based design resembles the virtual community and open innovation setups. Mechanical Turk has been used as a reliable source of experimental data. Numerous existing studies (e.g., Paolacci et al. 2010, Berinsky et al. 2012, Rand 2012, Paolacci and Chandler 2014, Lee et al. 2018) have addressed concerns regarding the data quality provided by participants in web-based experiments and demonstrated that online platforms can be a viable alternative for experimental data collection. Moreover, the results derived

from such platforms are found to be quite robust compared with those from traditional laboratory experimental settings in these studies.

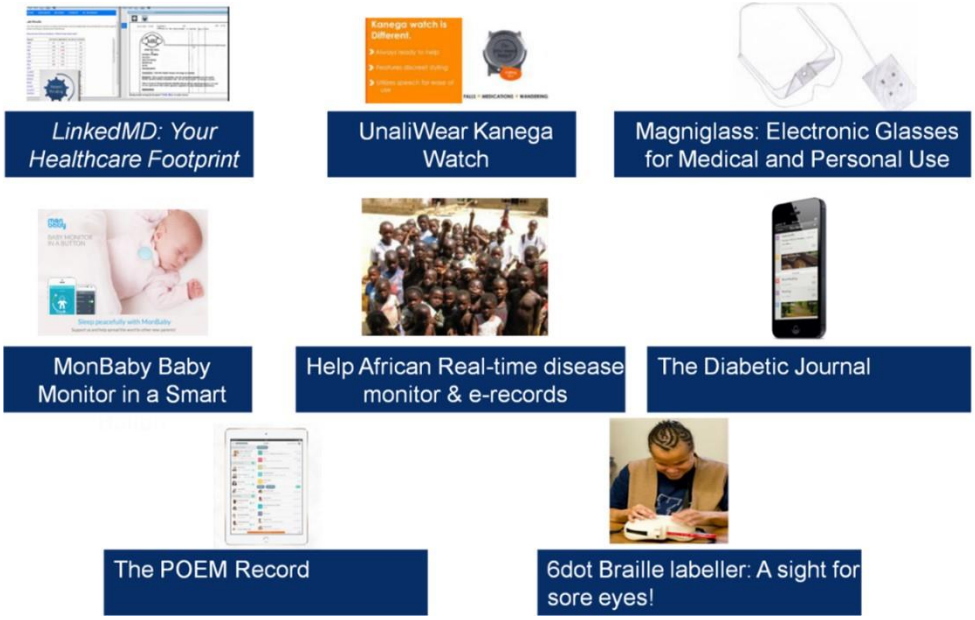


Figure 1-1: Snapshots of the ideas

In the first stage of the experiment, the study started by selecting eight ideas from the large crowdsourcing platform Kickstarter.com. All eight ideas belonged to the broad category of the intersection of information technology and healthcare, and they successfully obtained funding from Kickstarter. A snapshot of these ideas is presented in Figure 1-1.

Then, the eight ideas were presented to participants, and they were asked to evaluate the quality of these ideas under each of the four conditions: Ranking Short (RS), Ranking Long (RL), Scoring Short (SS), and Scoring Long (SL). In the *Short* conditions, the participants received only brief descriptions of the ideas, whereas in the *Long* conditions, they received extensive descriptions of the ideas. A typical brief description contained approximately 100 words, whereas an extensive description

typically contained 400–500 words. The participants were randomly assigned to one of the four conditions. In total, there were 423 participants in the experiment with at least 100 participants in each condition.

At the beginning of the experiment, a general description of the evaluation process was provided to the participants and were instructed that they should consider the originality or novelty of the ideas with respect to the existing market, the value of the ideas (i.e., the abilities of the ideas to solve relevant problems in healthcare) and the feasibility of the ideas in terms of the ease of translation of an idea into a commercial product. An idea is located at the fuzzy front end of new product development and represents only a vague business opportunity or direction to pursue, often lacking detailed or well-defined product concepts and attributes (Ulrich and Eppinger, 2015). As a result, unlike product or project evaluations (Kavadias and Chao, 2007), numerous existing studies regarding idea evaluation (e.g., Wilson and Schooler, 1991; Faure 2004; Girotra et al. 2010; Putman and Paulus 2009) have often asked participants to evaluate each idea holistically, instead of explicitly breaking down the ideas into attributes and evaluating each attribute. This study adopted a similar approach as these earlier studies to evaluate idea quality. To avoid any potential position or order bias (Bettman et al. 1991), the order in which the ideas were presented to each individual participant was randomized.

In the experiment, each participant received access to two panels of information. The first panel consisted of all eight ideas listed sequentially, and each idea was presented with either brief or detailed information. In the second panel, only the idea titles and the short summary of each idea were presented. In the two conditions of

the ranking process (RS and RL), the participants were asked to order the ideas in the second panel according to their assigned ranks. The participants were forced to strictly differentiate the ranks of these ideas by giving a unique order number to each idea (i.e., a *strong order* ranking process) from one to eight. In the two conditions of the scoring process (SS and SL), each idea in the second panel was accompanied with a rating slider scale below its summary. The participants were instructed to rate each idea on a Likert scale of 0 to 10, with 0 being a bad idea and 10 being an excellent idea. In contrast with the ranking process, the participants could give similar scores to multiple ideas.

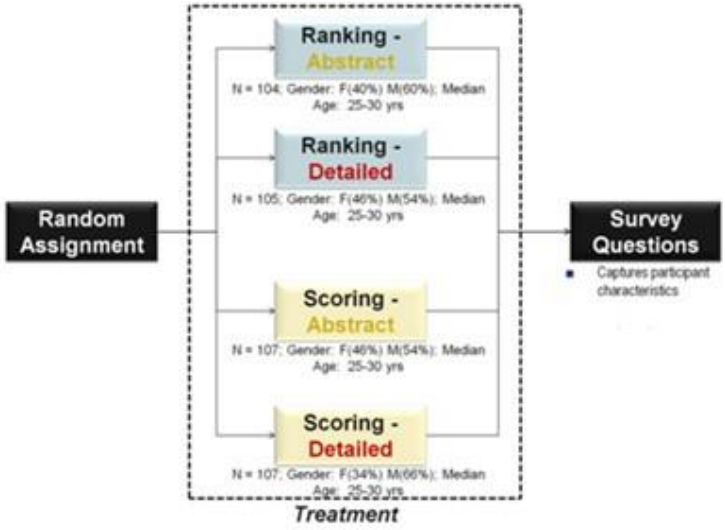


Figure 1-2: Design of the experiment

The participants were compensated with a fixed fee of $2 for their participation in the experiment. Additionally, it was clearly stated in the instructions at the beginning of the experiment that the participants' evaluations of ideas would be compared to a benchmark evaluation by a panel of experts, and the five participants under each treatment whose evaluations exhibited the highest percentage of

matches with the top three ideas from the experts' evaluations would be rewarded with an additional $10 each, which incentivized the participants to evaluate the ideas as accurately as possible. Figure 1-2 summarizes the details of the first stage of the experiment.

After the participants completed their idea evaluations, the experiment automatically proceeded to the second stage in which a survey was administered to each participant to collect information regarding their individual characteristics. Additionally, the IP address of each participant was checked to ensure the uniqueness of online participation. The time each participant spent on the evaluation was also recorded.

In the third stage of this experiment, a panel of healthcare-related experts was invited to evaluate the qualities of the eight ideas. These experts were participants in a management training program sponsored by the European Union for senior executives from the healthcare sector. On average, the experts in this study had over 10 years of healthcare-related experience. Nearly one-third of the experts had PhD degrees in relevant fields. The profile of the expert panel is summarized in Table 1-1.

No. of Experts	31
Industry	Healthcare Professionals
Average Experience	10.1 years
Gender	Female: 48%; Male: 52%
Representative Job Titles	Head of innovation. Researcher, Technology head/advisor, Project manager, Marketing manager
Education	PhD (32%); Masters (23%); Bachelors (45%)

Table 1-1: Profile of Experts

An expert panel is generally used in practice. For example, to bring the “optimal” technological solutions to market, IBM employs numerous senior industry experts to screen online entries by its open innovation participants (Bjelland and Wood, 2008). In the extant literature, the evaluations made by the experts are often used as the “benchmark” of the idea quality (e.g., Ozer, 2009, Girotra et al. 2010, Poetz and Schreier 2012, Kornish and Ulrich, 2014). In this study, the experts were given ideas with detailed descriptions and were requested to evaluate the ideas based on the previously mentioned criterion (novelty, usability, etc.) and provide their rating on a scale of 0-10. The experts’ evaluations were aggregated for each idea, thereby providing a benchmark score and associated ranks. Using the approach prescribed by Gwet (2002), AC2 measures were computed to verify the reliability of the evaluations provided by the experts. Gwet argues that any agreement by chance (say, a random assignment as the evaluator or rater is not certain about how to evaluate the item) can inflate the overall agreement probability. The reliability metric should thus ensure that such chance agreements do not influence the measure of actual agreement between raters. Gwet (2012, P. 121) introduces a measure of reliability (AC2, for ordinal or interval measures) between two or multiple raters, defined as the “conditional probability that two randomly selected raters will agree, given that no agreement will occur by chance.” These measures were statistically significant ($p<0.01$), which indicated a high level of overall reliability of the evaluations of the idea qualities. Some existing studies (e.g., Girotra et al. 2010) have demonstrated the relevance of using this method to examine reliability in the context of idea generation and selection - “Low p values suggest that the observed

inter-rater agreement is very unlikely to arise out of random chance, and that indeed the expert panel rated ideas in an internally consistent fashion" (Girotra, et al 2010, P. 598; Gwet, 2012).

1.3.2 Variables and Measures

Dependent Variable: Percentage of Matches

Because the generation of new ideas is highly uncertain, the chance of selecting a single "best" idea is quite low from a statistical point of view (Dahan and Mendelson 2001). Consequently, the focus of idea evaluation has been on maximizing the likelihood of identifying the several "best" ideas among a pool of ideas (Terwiesch and Ulrich 2009, Girotra et al. 2010). In practice, firms usually conduct multiple rounds of idea evaluation, and in each round, they select several promising ideas rather than a single idea. Then, the surviving ideas selected in one round are evaluated and screened again in the next round. This continuous process of idea selection is often referred to as an "innovation funnel" (Terwiesch and Ulrich 2009).

To operationalize the "innovation funnel" concept for idea evaluation, the accuracy of an evaluation process is quantified as the *idea match likelihood*, which is defined as the percentage of ideas that match between the crowd's (participants') and the expert panel's selections. Specifically, this study examines the match likelihood of each participant for the top two ideas (representing the top 25% of the ideas) as evaluated by the expert panel. To ensure the robustness of the results, the study also examined the match likelihood for the top three ideas (representing the top 37.5% of the ideas). Clearly, the more consistent the participants' and experts'

choices are, the higher is the accuracy of the idea evaluation process. This measure of accuracy also provides us with a comparable measure for two different types of data: interval (scoring) and ordinal (rank).

At an aggregate level of the crowd, in the ranking process, an average rank received by each idea is calculated. Then, these ideas are ordered according to the average rank each idea received, i.e., the idea with the lowest rank number is treated as the top one idea, and the idea with the highest rank number is treated as the worst idea as evaluated by all sample participants or the crowd. In the scoring process, the ideas are ordered according to the average score each idea receives. The idea that receives the highest average score is ordered as the top one idea. After completing this step of data aggregation, the order of ideas as evaluated by the crowd is compared with the benchmark order from the expert panel in order to identify the match percentage in the top two (three) ideas representing top 25 (37.5) percent of the ideas.

The match percentage calculation at the individual participant level is as follows. In the experiment, the participants could give similar scores to multiple ideas. This implies that ties could exist among multiple ideas, and such ties will blur the distinction between the top ideas and the remaining non-selected ideas. Should the top two or three ideas be selected from a pool of tied ideas, multiple selection outcomes with different match likelihoods will occur. For instance, in an extreme situation, a participant could give all ideas a score of 10. That would guarantee that he/she would get the top 2 or 3 ideas “right,” but the information value from his/her

scores would be lower than that from another evaluation in which only one of these top 2 or 3 ideas receives a score of 10 while all other remaining ideas would receive a score of 9. To address this issue, this study uses a three-step approach to measure each participant's match likelihood. First, based on the participant's evaluation, all possible outcomes when the top two (or three) ideas are selected were listed and the match likelihood associated with each outcome was calculated. Second, the probability of observing each outcome was calculated. Last, the probabilities derived from step two were used to arrive at the expected match likelihood of that participant. The following example illustrates the process of measuring the match likelihood for an evaluation with ties. In this example, the top two ideas evaluated by the experts are idea 8 and idea 6. Table 1-2 summarizes the evaluations made by the participant with ID 97293 in the *scoring* process.

In Table 1-2, the ideas are ordered according to the scores they received. Clearly, ideas 5, 6, 7, and 8 are top four ideas and are tied in scores. In this case, three outcomes can occur when the top two ideas are selected by this participant: (1) the selection of ideas 6 and 8, which results in a match likelihood of 100%; (2) the selection of only one of the top-two ideas (either 6 or 8), which results in a match likelihood of 50%; and (3) the selection of none of the top-two ideas (ideas 5 and 7 are selected), which results in a match likelihood of 0%. It follows that the probability of selecting both ideas 6 and 8 from the top four ideas is 1/6, the probability of selecting only one of the top-two ideas (either 6 or 8) is 4/6 or 2/3, and the probability of selecting none of the top-two ideas is 1/6. Thus, the expected match likelihood of the participant with ID 97293 is $100\% \times 1/6 + 50\% \times 4/6 + 0\% \times 1/6 = 50\%$.

Idea_5	10
Idea_6	10
Idea_7	10
Idea_8	10
Idea_2	8
Idea_3	8
Idea_1	7
Idea_4	7

Table 1-2: Sample evaluation with ties in the scoring process

Similar analyses were conducted for all the participants and the adjusted percentages of matches for both the top 2 and 3 ideas were calculated.

Treatments: Information

The participants in the experiment were randomly assigned to the four conditions, i.e., SS, SL, RS and RL. I included the level of information in the analysis with *Long* descriptions coded as treatment = 1 and *Short* descriptions coded as treatment = 0 to test for treatment effects within the scoring and ranking processes.

Explorative Variable: Duration

To examine the interaction between the participant's efforts and the evaluation process, the information concerning the evaluation duration of each participant in the experiments, i.e., the time spent by the participants on the task of idea evaluation was also captured. In the literature, time spent, or the variable *duration*, has often been used as a proxy for cognitive effort (Dhar and Nowlis 1999, Maule et. al., 2000).

1.4 Comparing the Ranking and the Scoring Decision Rules at an Aggregate Level

To examine at the aggregated level, how accuracy (i.e., the percentage of matches) varies across both the decision rules (Scoring and Ranking) as a function of the size of the crowd, a simulation was conducted by varying the number of participants under each treatment. By randomly choosing ‘n’ (n varied from 35 to 90) number of participants under each treatment the percentages of matches were calculated. This process was repeated over 1000 trials under two different match criteria (top 2 and top 3 idea matches). The accuracy values demonstrate the wisdom of the crowd (Surowiecki, 2004) effect, i.e., the accuracy improved as the size of the crowd increased (Figure 1-3). As the match criterion widens from the top 2 to the top 3, we observe that the performance of the ranking process improves; yet the scoring process clearly dominates. An interesting observation from Figure 1-3 is that the scoring process converges at higher levels of accuracy with smaller crowds compared with the ranking process. This finding indicates that the scoring process yields higher accuracy levels with relatively smaller crowds, which has direct managerial implications in the design of an open evaluation initiative.

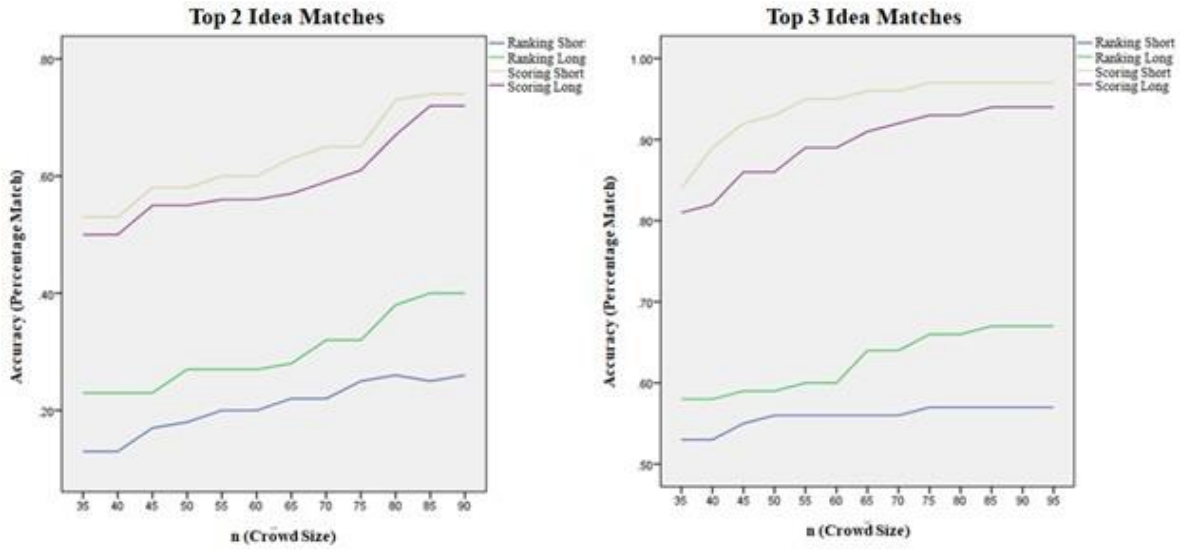


Figure 1-3: Accuracy as a function of crowd size under different match criterions

At an aggregate level, the scoring process performs better than the ranking process in identifying the top ideas. The effect of the treatment in terms of level of information also seems to be having some effect on the outcome based on the decision rules used. To further investigate the dynamics of the decision rules, the following sections investigate the performance of the decision rules in detail.

1.5 Comparing the Ranking and the Scoring Decision Rules at an Individual Level

The average percentages of matches for the top two ideas along two dimensions of treatments were calculated (by employing the steps described in the earlier section), i.e., information (short vs. long) and process (scoring vs. ranking). Table 1-3 summarizes the results.

		Information		Average
		Short	Long	
Process	Ranking	20%	28%	24%
	Scoring	39%	37%	38%
Average		29%	33%	31%

Table 1-3: The percentages of "matches" for the top two ideas (strong order ranking)

Table 1-3 illustrates that in both information conditions, the scoring process produced higher average percentages of matches than the ranking process. Additionally, the percentage of matches in the ranking process increases from 20% to 28% when the detailed information is provided.

Variable	F-ANOVA	P-value
Process	18.12	<0.01***
Information	0.08	0.31
Process × Information	0.17	0.02**

Table 1-4: ANOVA analysis (strong order ranking)

Next, an ANOVA analysis was conducted using process, information and their interaction term. Table 1-4 summarizes the results. The variable *process* was significant ($p<0.01$), and the variable *information* was nonsignificant ($p = 0.31$). Interestingly, a weakly significant interaction effect is found between the information and the process ($p < 0.05$). Specifically, in the ranking process, the provision of detailed information significantly increases the percentage of matches by 8 % ($p = 0.05$). However, providing more information does not significantly increase the

accuracy of the scoring process. A parallel analysis using the top three ideas criterion was also performed. All the results remained consistent.

In the experimental setting above in which the ideas were strictly ordered without ties, the experimental results show that the scoring process strictly outperforms the ranking process in terms of the likelihood of selecting the top two (three) ideas. One possible explanation of the relatively low evaluation accuracy of the ranking process is that, under the strong order ranking, the participants lose the freedom to assign similar ranks to two or multiple ideas in cases in which he/she perceives the ideas as equally "good". By contrast, the participants in the scoring process have the additional freedom to provide score ties among multiple ideas. As such, the discernibility of the ranking process may be undermined. To test this possibility, a new experimental study was conducted in which the participants could give similar rank numbers to multiple ideas (i.e., a *weak order* ranking process). The detailed experimental design and results are reported in Appendix A-1. Consistent with the previous results of the comparisons when the strong order ranking was employed, the new results in the weak order ranking experiment indicate that under both the detailed and short information conditions, weak order ranking produced lower percentages of matches than the scoring process in the main experiment.

1.6 Examining the Influence of Idea Quantity

As mentioned in the existing literature, the main limitation of the ranking process is that as the number of ideas increases, the number of required comparisons rapidly increases, which creates severe scaling challenges for large idea sets (Dahan and

Mendelson 2001, Toubia and Flores 2007, Dahan et al. 2010). This limitation implies that the difference in evaluation efficacy between the scoring and ranking processes may decrease as the number of ideas to be evaluated decreases. This section aims to explore this possibility. Clearly, randomly selecting some ideas from the *existing* experimental data cannot provide reliable results because the existing data were derived from processes in which eight ideas were evaluated and thus cannot reflect the underlying mechanism of evaluating fewer ideas[1]. Instead, three ideas were randomly picked out of eight available ideas and a *new* experimental study was conducted by letting the participants evaluate these three ideas under treatments similar to those of the previous experiments. Specifically, 299 new participants were randomly assigned to one of the four experimental conditions (69 in RL, 79 in RS, 73 in SL and 78 in SS evaluations). As the number of ideas in this experiment was reduced to three, the analysis examined the percentage of matches for the top one idea (top 33%) to make a fair comparison with the previous experiments (in which the top 25% of ideas were selected). Ties were allowed in the ranking process here. All other experimental settings remained the same.

Tables 1-5 and 1-6 summarize the average percentages of matches in the four treatments and present the results of an ANOVA analysis, respectively. When the number of ideas is reduced to three, no significant differences were found between the four treatments. The interaction effect between information and process is also not found. These results imply that when the number of ideas is sufficiently small (3

[1] However, a simulation study was conducted by selecting different number of ideas from the existing pool of ideas as a robustness check and the same is briefed in Appendix A-2.

in this experiment), the efficacies of the two processes (ranking and scoring) become equivalent.

		Information		Total
		Short	Long	
Process	Ranking	31%	33%	32%
	Scoring	36%	35%	35%
Total		33%	34%	33%

Table 1-5: ANOVA analysis (three ideas)

Variable	F-ANOVA	P-value
Process	0.49	0.485
Information	0.05	0.830
Process × Information	0.13	0.721

Table 1-6: ANOVA Analysis (three ideas)

1.7 A Path-dependent Model and an Empirical Test

The analysis thus far demonstrated that when eight ideas were evaluated by the participants, the scoring process achieved better evaluation accuracy than the ranking process. One potential mechanism causing this result is the high demand for cognitive “working memory” in the ranking process. From the perspective of information processing, the ranking process inherently requires evaluators to hold and process greater numbers of chunks of information than does the scoring process and therefore becomes a much more complex and demanding task. Consequently, the ineffective processing of information in the ranking process forces the decision makers to use heuristics that could lead to mistakes (Bettman and Kakkar 1977). As

argued in Section 2, this theoretical perspective predicts that additional information should improve the efficacy of the scoring process, instead of the ranking process. Although the statistical significance ($p = 0.05$) of the interaction effect between information and the evaluation process may not be very conclusive, the experimental findings presented in Section 5 do not support this prediction.

The experiments recorded the time spent by the participants on the idea evaluation task as an explorative variable. Table 1-7 summarizes the average evaluation durations under the different conditions. To exclude some extreme values and obtain robust results, a winsorization approach was adopted by replacing the outliers on both sides with 2.5 percentile value and 97.5 percentile value, respectively. Table 1-7 shows that the imposition of information has significant impacts on the duration in the ranking process. In particular, providing additional information will increase the duration from 7.08 to 9.79 minutes (2.71 minutes longer or a 38.2% increase, statistically significant at $p<0.01$) under the strong order ranking (similar increase is observed under the weak order ranking as well). By contrast, additional information leads to an increase of only 1.06 minutes (from 7.94 to 9.00 minutes or a 13% increase) in the scoring process. In the ranking process, the detailed (long) ranking requires the longest average duration followed by the detailed (long) scoring, the brief (short) scoring, and the brief (short) ranking. This consistent pattern indicates that the information exerts a greater influence on the evaluation time in the ranking process than in the scoring process, implying a potential moderating role of the information on the relationship between the process and the evaluation duration.

Duration			
Process	Information		Total
	Short	Long	
Ranking	7.08	9.79	8.44
Scoring	7.94	9.00	8.47
Total	7.51	9.40	8.50

Table 1-7:Average durations of evaluation (unit: minutes)

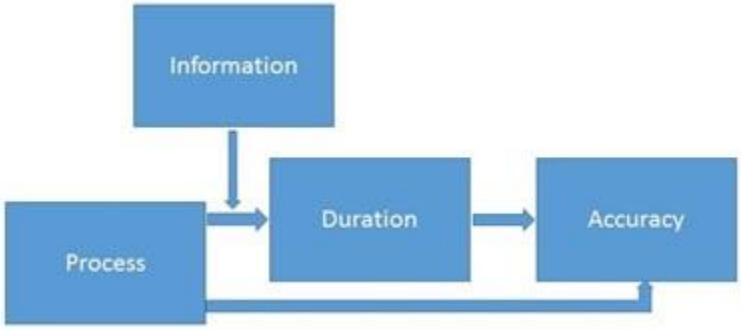


Figure 1-4: A path-dependent model of evaluation accuracy

Motivated by these observations, this study proposes a path-dependent model in which the evaluation accuracy (percentage of matches) is *directly* determined by both the evaluation process and the time (duration) that the participants spend evaluating the ideas, and the duration of evaluation in turn depends on both the evaluation process and the information provided. The experimental results summarized in Table 1-4 show that the information exerts no significant direct impact on the evaluation accuracy[2]. Therefore, the direct impact of the information is

[2] A model with information having a direct impact on accuracy was tested but was found to be not significant and thus it is not included here.

excluded. On the other hand, information may interact with process such that it will produce a greater increase in duration in ranking than in scoring processes. Figure 1-4 describes the model.

To examine the validity of this model, a structured two-stage empirical model was adopted to analyze the experimental data. In the first stage, the *duration* is treated as the dependent variable, and *information*, *process,* and their interaction term are independent variables. In the second stage, the predicted value of *duration* from the first stage and the *process* are used as the predictors of the evaluation accuracy. The two stages of analysis were conducted simultaneously as a system of equations. The *duration* variable was normalized by taking its natural logarithm. The evaluation accuracy is again measured by the percentage of matches for the top two ideas. Table 1-8 summarizes the results of our empirical analysis.

	First-stage eq:		Second-stage eq:	
	Log Duration		*Percentage of Matches*	
	b.	s.e.	b.	s.e.
Independent variables: second stage				
Log Duration			0.066**	0.025
Process			0.153***	0.026
Independent variables: first stage				
Process	0.163**	0.075		
Information	0.355***	0.075		
Process × Information	-0.231**	0.106		
Constant	1.755***	0.053	0.106**	0.049
N	423		423	
R-sq	0.058		0.091	
F-value	8.5572		21.136	
p-value	0.000		0.000	

*** p-value<0.01, **p-value<0.05, *p-value<0.1. s.e: robust standard errors.

Table 1-8:Empirical tests for the path-dependent model

In the first stage of analysis, significant effects of process, information and their interaction term are found. The results demonstrate that providing additional information increases the duration of idea evaluation. Additionally, information negatively moderates the direct effect of evaluation process on duration. Specifically, when detailed information is provided (information = 1), the ranking process takes longer than the scoring process; however, when only brief information is provided (information = 0), this effect is reversed. The results of the second-stage analysis demonstrate the significant and positive influences of duration and process on the evaluation accuracy. Overall, the results indicate a more intricate mechanism that explains the experimental findings than the conventional perspective of information processing theory. That is, information does not directly influence the evaluation accuracy. Instead, increasing information load could increase the cognitive effort, which in turn jointly influences the evaluation accuracy in combination with the evaluation process.

1.8 Discussions and Implications

The results demonstrated that when eight ideas are evaluated, the scoring process outperforms the ranking process in terms of the percentages of matches at both aggregate and individual levels. However, when the number of ideas is reduced to three, no significant difference was found between the ranking and scoring processes. At an aggregated level, the scoring rule is found to extract better wisdom of crowd from a relatively smaller group size while using the scoring rule to evaluate ideas. The experimental results also suggest that providing additional information

improves the efficacy of the ranking process but not that of the scoring process. These results negate the prediction from the theoretical perspective of information processing. To explain these experimental findings, this study proposed a path-dependent model in which the duration of the evaluation plays a critical role, and the data fit well with the theoretical model.

The results of this study have broad implications for both research and practice. The success of open innovation depends on both generating and identifying ideas with the greatest potential for the organization. Existing studies have extensively examined the idea generation process, but there are few results regarding the efficacies of idea evaluation processes. This study fills this academic void by comparing two commonly employed idea evaluation processes or decision rules and provides insights into how the efficacies of these processes are influenced by the number of ideas to be evaluated and the amount of available information. From a managerial perspective, this study provides guidance for the choice of idea evaluation process. The results demonstrate that when there are many ideas to be evaluated, letting reviewers rate each individual idea could lead to higher evaluation accuracy than letting them rank all ideas together. However, when the number of ideas is limited, for example, fewer than five, the process chosen for idea evaluation may not matter. Further, at an aggregated level, the scoring process can enable higher levels of accuracy with fewer participants as the scoring process is more effective for evaluating ideas when there are several ideas to be evaluated.

The results provide experimental evidence that information may not exert a direct effect on evaluation accuracy. Instead, information exerts an *indirect* effect by

moderating the relationship between the evaluation process and the evaluation outcome. When only a small amount of information is provided, the participants in the scoring process spend longer time on average evaluating ideas than do those in the ranking process. In this case, the participants in the ranking process are likely using some simplifying strategies to denote each single piece of brief information without fully incorporating them into the evaluation (e.g., Beach 1993). However, this study suggests that when additional information about each idea is provided, the participants find it difficult to simplify the information and increase their span of attention to consider and incorporate such information into their evaluations, which in turn increases the evaluation duration and accuracy.

Like any study, this study is subject to some limitations that create opportunities for future research. The general shortcomings associated with controlled experiments apply to this work. It could be interesting to empirically compare the efficacies of different idea evaluation processes in a more “realistic” setting. Specifically, this study uses a benchmark based on the idea evaluation of an expert panel which rated the ideas using a scoring scale (similar to the past studies in the innovation and new product development space, for eg., Ozer, 2009, Girotra et al. 2010, Poetz and Schreier 2012, Kornish and Ulrich, 2014). Since this study is comparing different decision rules, a valid argument could be that the superior performance of the scoring rule might be due to some correlations between the scoring scale used by both the expert panel and the subjects. This study in this regard followed the methods of obtaining and validating the benchmark evaluation as per extant research in the new product development space. Past studies have

indicated that though novices have been consistently shown to fall prey to logical fallacies, experts typically use heuristics that are designed to take advantage of relevant information and information structures along with their experience to make accurate decisions under environments characterized by uncertainty and limited information (Montgomery & Lee, 2021). The argument here is that the experts with their wide knowledge base and rich experience (domain, market, industry, etc) align themselves with the decision-making environment (Gigerenzer, 1996; Hoffrage, 2019) by relying on the most valid cues and relevant information. Shanteau (1992, p. 82) argues that "what novices lack is the experience or ability to separate relevant information from irrelevant information sources. Thus, it is the type of information used — relevant vs. irrelevant — that distinguishes between experts and others." Therefore, experts might be able to apply more accurate heuristics and outperform other decision strategies (Gigerenzer et al., 1999; Todd & Gigerenzer, 2000). The expert panel used in this study belong exclusively to the health-care industry, recognized by the European Union, and are highly qualified and experienced (please refer to Table 1-1) in the health-care domain. Thus, following the past studies in new product development and the regulatory-view (Montgomery & Lee, 2021), I believe that the chances of the results being fully driven by any correlations of scales is limited, especially when experts are aligned and sensitive to the environmental structures and the domain cues. However, an alternative test has not been separately conducted to test this belief and hence this remains to be a limitation of this study and turns out to be an important aspect to be tested for future studies. Nevertheless, within the resources and framework of the study it is found that a major

driver for the results here as indicated earlier is the cognitive load imposed by the number of ideas. In addition to the additional experiment with fewer number of ideas, a simple robustness analysis (shown in appendix A-2) indicate that the scoring rule is likely to outperform the ranking rule when the number of ideas is greater than four. However, when the number of ideas is reduced to four or below, the efficacies of the two rules are likely to become equivalent. Also, this study only uses the top-most ideas as per the rules of the extremes applicable to the innovation literature (Dahan & Mendelson, 2001; Terwiesch and Ulrich, 2009) and does not use the other ideas as such. A separate study could focus on the overall effectiveness of different scales in an applicable context using different measures (for example, refer to Langville & Meyer, 2012).

Theoretically, the results of this study suggest that in the idea evaluation processes, the information is likely to be used as a cue to inducing cognitive efforts. However, this effect is relatively weak in these experiments ($p = 0.05$), and thus we must be cautiously aware of the limitations regarding the generalizability of this result. More in-depth research could be conducted in the future to examine the relationship between information and the evaluator's efforts as well as the generalizability of this relationship in different settings. Additionally, it could be quite relevant and interesting to explore the relative performance of additional evaluation processes across multiple dimensions and categories of ideas. For example, how the format in which the information is displayed (e.g., a structured table vs. unstructured text) affects idea evaluations could be an interesting research question.

2 Chapter 2: The Effectiveness of Ensembles of Judgmental Forecasts in Times of change

2.1 Introduction

One of the critical challenges in managerial forecasting relates to the question of how to efficiently adapt to structural breaks in time series environments (Giacomini & Rossi, 2010). Structural breaks represent sudden, discontinuous changes (see for instance Tian & Anderson, 2014; O'Connor, Remus & Griggs, 1993), which cause upward or downward shifts in the mean of the data series. Such shifts can result from a wide range of contextual factors such as technological disruptions, regulatory changes, legislative interventions, natural disasters, etc (Pesaran, et al, 2006). For example, the implications of the recent spread of the COVID-19 on economy and business has been a topic of interest. A study conducted by McKinsey[3] has stressed upon the sudden drop in economic growth predictions due to the pandemic and further predicts the recovery of economic factors to follow a staggered muted pattern and, on similar lines, OECD has seen the necessity to substantially downward-adjust economic growth predictions for the year 2021 in order to account for sudden drops in economic productivity, unexpected increases in unemployment as well as lack of investor confidence (Boone, 2020; Boone, et al, 2020). Similarly, studies have focused on the implications of changes in technology on market shifts (for example,

[3] https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/the-coronavirus-effect-on-global-economic-sentiment#

see National Research Council. Persistent Forecasting of Disruptive Technologies" Report National Academies Press, 2010). These studies highlight that technology changes are a ‘source of surprise’ and predicting the exact times of shift in demand or forecasting demand under such scenarios remains a great challenge as there is uncertainty in how a technology per se advances, how the regulatory bodies influence the technology advancement, or how consumer preferences shape around these changes, or factors which are ‘difficult to anticipate’.

From a scholarly perspective, structural breaks have frequently been studied in the context of forecasting stock market returns (Paye & Timmermann, 2006), exchange rates (Rossi, 2006), and other macroeconomic variables (Rossi, 2013). However, their detection appears to represent a complex cognitive challenge as individuals systematically under- and overreact to information signals. Forecasters are likely to systematically underreact to forecast errors in relatively unstable environments but overreact to errors in relatively stable environment (Kremer et al. 2011). Massey and Wu (2005) argue that individuals primarily react to signals that are indicative of change and only secondarily to the environmental system generating the signal. It is also shown that there is an important systematic bias associated with individuals’ forecasts when time series undergo a structural change (O’Connor et al, 1997 and 1993; Lawrence & Makridakis, 1989). In particular, individuals under-forecast in response to sudden upward shifts and over-forecast in response to sudden downward shifts. For example, during the 1970s, when sudden oil price shocks led to an unexpected increase in inflation, judgmental forecasts obtained through the

Survey of Professional Forecasters (SPF) and Blue-Chip Economic Indicators (BCEI) have been found to be systematically biased in that they consistently under-forecasted inflation rates (Rossi & Sekhposyan 2016). During the 1980s and 1990s, however, when a series of political interventions reduced inflation beyond expectation, forecasting judgments from the same surveys were found to consistently overestimate inflation rates.

Considering the notion that structural breaks are of importance in a variety of different forecasting contexts as well as the increasing body of descriptive research demonstrating that forecasting judgments are systematically biased in these environments, prescriptive approaches that could potentially improve forecasting performance remain largely unexplored. Specifically, while recent research has proposed a few prescriptive aggregation rules that are intended to (at least partially) compensate for judgmental biases in individual forecasts (e.g., Grushka-Cockayne et al. 2017, Jose et al. 2014), these studies have so far focused on aggregating forecasts in stationary environments. Addressing this limitation by studying the effectiveness of aggregation rules in nonstationary forecasting environments lies at the heart of this study. The objective of this study is threefold: First, to extend the existing literature on forecast combinations by systematically studying the performance of various aggregation rules in environments containing structural breaks and demonstrate the limitations of commonly used aggregation rules under the presence of systematic biases in such environments. Second, to propose an aggregation rule (a novel form of asymmetric trimming) to minimize the impact of

systematic biases on forecast aggregation. Finally, third, to prescribe a forecast ensemble method to combine forecasts obtained from different forecast aggregation rules to improve judgmental time series forecast performance in SB environments. The purpose of the ensemble methods is to efficiently extract wisdom of the crowd from a group of forecasters in environments where structural breaks can occur at any point of time in the data series. This study shows that the ensemble method is likely to compensate for the systematic biases frequently observed in judgmental forecasts. The contribution of this essay therefore lies at the intersection of behavioral and prescriptive decision theories.

In the past literature on judgment aggregation, simple averaging has frequently been found to represent an effective means of combining forecasts (Makridakis & Winkler, 1983; Clemen, 1989; Armstrong, 1989; Armstrong, 2001; Thomson, et al, 2019). However, if judgmental forecasts suffer from systematic biases, the benefits of simple averaging may be dampened. This is because forecasting judgments that are systematically biased towards same direction are less likely to bracket the true time series value (Larrick & Soll, 2006). Recent studies have also proposed a symmetric trimming methodology for improving the accuracy of aggregated forecasts (Armstrong, 2001, Jose et al 2014). Symmetric trimming refers to the process of truncating upper and lower extreme judgments within a group before averaging, which has been shown to improve forecasting performance especially when the variance of forecast judgments is large (Jose et al. 2014). However, in nonstationary environments symmetric trimming may not prove to be as effective, because

truncating a systematically biased opinion pool, potentially accurate judgments may get unintentionally lost, which would subsequently result in a lower bracketing rate. The extant literature does not address the issue of how to effectively aggregate forecasting judgments in unstable time series environments. This essay aims to propose an asymmetric trimming rule (analogical to the rule proposed by some studies which focus on combine probability forecasts, for example, Gaba et al, 2017) for coping with forecasting environments that are subject to mean shifts. Specifically, the rule begins by ranking individual forecasts from smallest to largest and then eliminates judgments from one side only to retain the most valuable instances of bracketing. When judgmental forecasts are positively biased (i.e., when they under-forecast the true value) due to an upward mean shift in the series, it is easy to see that the lower extreme values will be less predictive in the aggregation process. In this situation the trimming rule therefore suggests to asymmetrically truncate the opinion pool on the left side. Throughout this essay, such a rule is referred to as the *left trim*. In contrast, when judgmental forecasts are negatively biased (i.e., when they over-forecast the true value) due to a downward mean shift in the series, it would be valuable to asymmetrically trim only the highest values in the pool. Hence, such a rule is referred to as the *right trim*. In sum, this study demonstrates that symmetric trimming is preferable in stable environments, whereas left/right trimming is preferable in unstable environments that contain upward/downward shifts in the mean of the data series.

In the context of these symmetric and asymmetric trimming rules, one critical challenge is to decide on *how* the most beneficial trimming method should be chosen in each period of the time series, particularly if the timing and type of structural break is not known a prior. Armstrong (2001) argues that in environments that are characterized by high levels of uncertainty, it is recommended to ensemble various forecast methods. Consistent with this view, this essay proposes an ensemble method that combines the three trimming rules by relying on a weighted sum (see Figure 2-1). This study examines the efficiency of two alternative approaches to deriving weighing schemes for such an ensemble method.

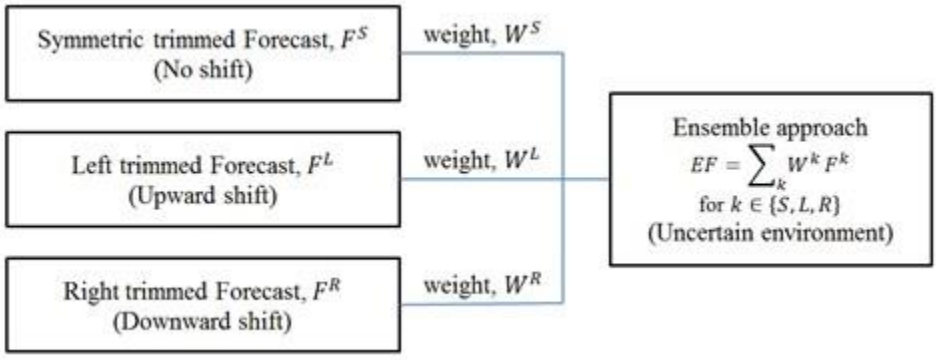


Figure 2-1: The Ensemble Method

The first approach (which is referred to as the “Break Judgment Ensemble”) exploits information about forecasters’ subjective judgment regarding the occurrence of a structural break. The judgmental forecasting literature has often highlighted human judges’ diagnostic ability to detect systematic patterns in time series environments (Lawrence, et al, 2006; Simon, 1990; Seifert et al. 2013). Here, the task of generating a judgmental forecast is decomposed into two sub-tasks: (i) a diagnostic judgment if a structural break has occurred in the most recent part of the time series (e.g. an

upward mean shift, a downward mean shift or no mean shift) and (ii) a forecasting judgment for the next period in the time series. A collective break perception of judgments in the opinion pool can then be calculated as the proportion of forecasters who share the same belief about the occurrence of a structural break. The Break Judgment Ensemble method hence combines trimming rules by using collectively held break perceptions as weights. In the second approach to derive weights for the ensemble method (which is referred to as the “Past Performance Ensemble”), the weights are derived from the ranked squared error performance of each trimming rule in the prior period of the time series. In other words, the Past Performance Ensemble assigns larger weights to those trimming rules, which led to smaller errors in the past time-period.

To test the performance of the Past Performance and Break Judgment Ensemble methods, an experimental study was conducted in which subjects provided forecasting judgments for a wide range of artificially generated time series environments containing structural breaks. In addition, real world data obtained from the Survey of Professional Forecasters were used for testing the external validity of the ensemble method.

In summary, this study finds that the proposed asymmetric trimming rule for forecast aggregation is likely to compensate for systematic biases frequently observed in the presence of structural breaks. Importantly, the proposed ensemble method combining different forecast aggregation rules yields robust forecast performance improvements in both, experimental set-up (in the range of 7-9 percent compare to

commonly used forecast aggregation rules) and real-world time series (in the range of 5-7 percent compare to commonly used forecast aggregation rules).

This study mainly aims to contribute to the forecast combination literature by analyzing the effectiveness of different forecast combination methods under an environment characterized by uncertainty which has been relatively unexplored in the judgmental forecast combination research. This study aims to add to the body of knowledge of the two streams of judgmental forecasting and forecast aggregation. This study aims to bridge these two streams by demonstrating the impact of systematic biases associated with time series characteristics contributing to descriptive research and by proposing a valuable prescriptive rule (asymmetric trimming) to combine forecasts by accounting for these systematic biases, contributing to prescriptive research. This study further contributes to prescriptive research by prescribing a forecast ensemble method to combine forecasts obtained from different forecast aggregation rules to improve judgmental time series forecast performance in environments characterized by structural shifts. Task decomposition (in terms of obtaining break predictions) could be helpful in obtaining cues about the direction of the series; however, a simple majority-based rule to aggregate the judgments may not necessarily be that effective. From a methodological perspective, this study incorporates sequential breaks or discontinuities in different possibilities, and this has not been incorporated in earlier experimental designs. The proposed prescriptive rule can overcome the practical difficulty of identifying super performers

in a group based on post sample data and could be a potential alternative to aggregate forecasts in practice.

2.2 Asymmetric Trimming and Ensemble Methodology

2.2.1 Judgmental Biases and Trimming Rules

Prior research suggests that forecasters frequently suffer from systematic biases under unstable time series environments that contain structural breaks, (O'Connor, et al, 1997; O'Connor, et al, 2001). Specifically, several studies have found that forecasters respond to sudden upward shifts in the mean of a time series with a positive judgmental bias. That is, they are likely to under-forecast by providing point estimates that are lower than the true value. In contrast, forecasters are likely to exhibit a negative judgmental bias when the mean of the time series shifts downwards. That is, individuals are likely to over-forecast the true value in response to such shifts (Lawrence & Makridakis, 1989; O'Connor et al, 1993; O'Connor et al 1997). Furthermore, research (for example, Kim, et al, 2001; Rossi, et al, 2016) also suggests that structural breaks tend to have a sustained effect on forecasting performance as the judgmental bias remains observable even several periods after the occurrence of the break.

One of the widely discussed reasons for systematic biases are found to be in anchoring and adjustment heuristics (Tversky & Kahnemen, 1974). In practice, forecasters may underweight new information, be over influenced by their past forecasts and/or actual realized values or most recent consensus forecasts, overweight random variations, etc. (Batchelor & Dua, 1992; Campbell & Sharpe,

2009; Ichiue & Yuyama, 2009; Fujiwara, et al, 2013). It is also found that forecasters often do not like to make large and sudden adjustments to their past forecasts in order to maintain consistency of reputation and therefore tend to make slow and smoothened adjustments to meet their rational expectations (Campbell & Sharpe, 2009; Nakazono, 2012; Chang & Chou, 2016).

In stationary time series environments, a large body of literature has focused on the role of forecast combinations to improve predictive performance. Previous studies have often highlighted the crucial role of diversity when aggregating judgments (Makridakis & Winkler, 1983; Clemen, 1989; Armstrong, 1989; Armstrong, 2001; Lichtendahl et al, 2013). The notion of diversity is based on the idea that diverse opinions held by group members will result in an increase in the likelihood that true values will be bracketed (Larrick and Soll, 2006; Hong & Page, 2012).

While simple averaging has often been shown to be effective in stable environments, Budescu and Chen (2015) propose a model that distributes weights proportional to the individual performance of forecasters such that it assigns larger weights to group members whose judgment result in small forecast errors. Other studies have proposed forecast trimming of opinion pools as viable approaches for improving predictive accuracy. Trimming is particularly valuable when the variance of judgments in an opinion pool is high, as it reduces the effect of extreme views held by individual forecasters in the group (Jose, et al, 2014; Grushka-Cockayne, et al, 2017).

While the aggregation methods mentioned above have been proven to work efficiently in stationary, stable time series environments, it remains unclear to what extent they can improve the predictive accuracy of judgmental forecasts in unstable environments containing structural breaks. Since individuals are likely to suffer from systematic biases when detecting the occurrence of structural breaks, a simple average or a symmetrically trimmed average would yield limited benefits as it may lead to a potential loss of less biased individual forecasts that would prove useful in the aggregation process. In other words, systematic biases in forecasts are likely to reduce the effectiveness of bracketing (in forecast aggregation) and hence impede the benefits of prescriptive models suggested by the extant literature.

2.2.2 Overview of Trimming Rules

This section provides a general intuition regarding the underlying mechanics of trimming. Let us consider a time series with a structural break that shifts the mean of the series upwards. Moreover, consider that forecasters observing this time series exhibit a positive bias in their judgmental predictions (i.e. under-forecast values). Now suppose the true value in period t is $A_t = 120$, and the opinion pool X_t consists of five individual judgments regarding A_t , i.e. $= [70,90,105,120,125]$ (Figure 2-2). A simple average of the five individual forecasts is $\bar{X}_t = 102$ and exhibits a judgmental bias of b_t =18 ($i.e., A_t - \bar{X}_t$). When applying a symmetric trimming rule by excluding one value from both extremes and averaging the remaining values in the opinion pool, the trimmed mean would be now elevated to $F_t^S = 105$, while exhibiting a smaller bias of $b_t^S = 15$. Note that in this chosen under-forecasting scenario,

symmetric trimming reduces the number of bracketing instances by eliminating the highest judgment in the opinion pool, even though the symmetrically trimmed mean appears closer to the true value A_t than the simple average X_t. In other words, when forecasts are systematically biased, symmetric trimming may lead to the loss of valuable information contained in the opinion pool. In consequence, this study proposes an asymmetric trimming rule, which truncates the forecast aggregated by eliminating the least valuable forecast judgments from one side while retaining the valuable instances of bracketing on the other side. In the above-mentioned example, extreme values at the lower end of the range are less valuable in the aggregation process and thus the two lowest judgments in the opinion pool should be trimmed[4] to yield an asymmetrically trimmed forecast of $F_t^L = 117$; which results in a smaller error than the symmetric trimming method.

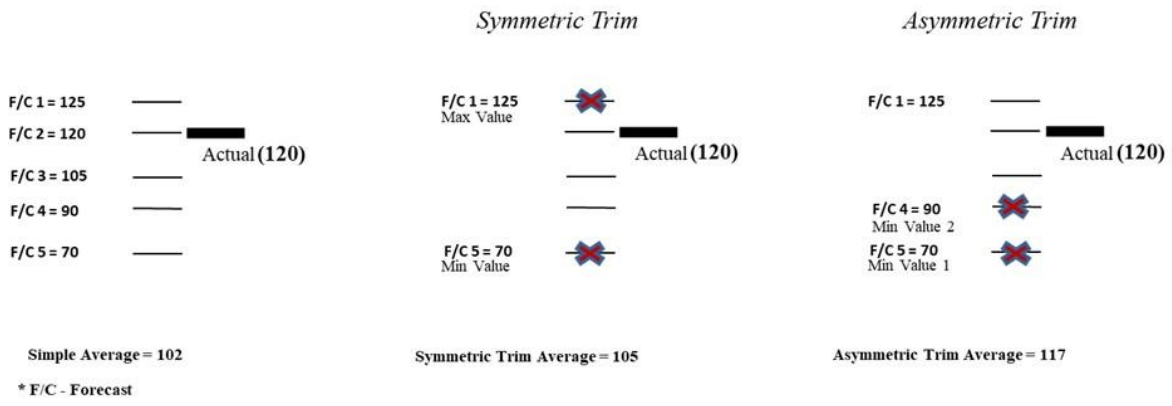


Figure 2-2: Comparing trimming rules in a positively biased forecasting scenario (under-forecast)

[4] two values are chosen in order to be consistent with the trim-size used in the earlier case of symmetric trimming.

Similarly, suppose the time series contains a structural break, which shifts the mean of the series downwards as well as judgments in the opinion pool exhibit a negative bias (i.e., over-forecast). Furthermore, consider that $A_t = 80$ and $X_t^R = [75,80,90,105,120]$ (Figure 2-3). Applying symmetric trimming by excluding one value from both extremes (i.e., the minimum and the maximum) would result in a trimmed estimate F_t^S =92. It could be seen that in this over-forecasting scenario, the process of symmetric trimming reduces the number of bracketing instances by eliminating the lowest extreme forecast. Instead, an asymmetric trimming rule can be used to retain these lowest forecasts. Thus, asymmetric trimming would truncate the opinion pool by eliminating the two highest extreme forecasts and yield a forecast of $F_t^R = 82$, which results in a smaller squared error than the symmetrically trimmed estimate.

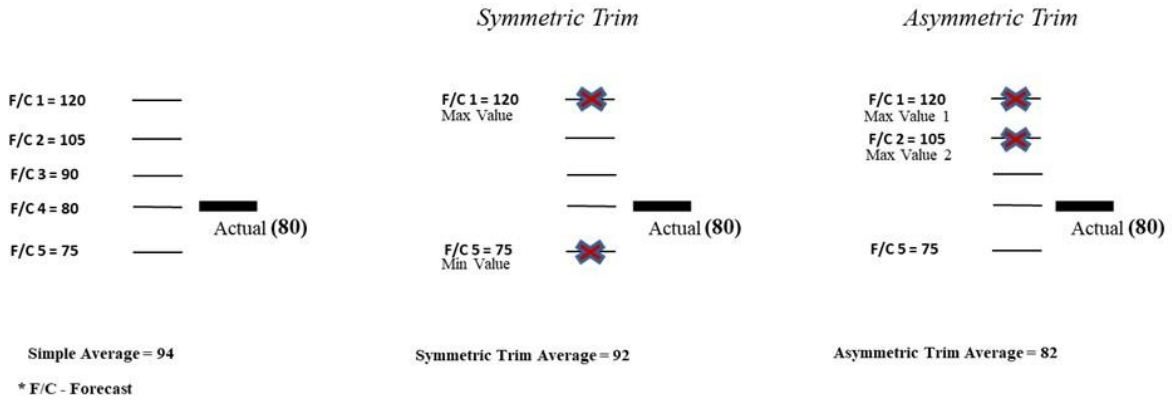


Figure 2-3: Comparing trimming rules in a negatively biased forecasting scenario (over-forecasting)

The two examples demonstrate the value of asymmetric trimming when judgmental forecasts suffer from systematic biases. This study proposes that in situations where

under-forecasting is prevalent, asymmetrically trimming the lowest judgments in the opinion pool (*left trim)* will improve forecasting performance. Likewise, in situations where over-forecasting is prevalent, asymmetrically trimming the highest judgments in the opinion pool (*right trim*) will improve predictive performance. While earlier studies have highlighted the benefits of symmetric trimming in situations where the variance of individual judgments in a group is high (Armstrong, 2001), this study argues that this will hold true only in stable, stationary time series environments. However, in unstable environments that are likely to contain structural breaks and in which judgmental forecasts are systematically biased, asymmetric trimming improves forecasting performance of aggregated forecasts.

2.2.3 Ensemble Forecast

Prior research has indicated that judgmental forecasting biases are directly related to the characteristics of the time series (O'Connor, et al, 1997; O'Connor, et al, 2001). The bias is likely to be positive for upward shift in the mean of the time series and negative for downward shifts. Moreover, there appears to be no systematic bias when the time series is stationary and does not contain any structural breaks. Thus, the use of the *right* aggregation rule (symmetric, left or right trim) at each period in the time series plays a critical role for improving judgmental forecast performance. However, when the time series contains structural discontinuities, it becomes extremely difficult to predict the occurrence and type of shifts. Various studies (Hendry & Clements, 2004; Diebold & Shin, 2019; Atiya, 2020) argue that in environments surrounded by high uncertainty, it is recommended to combine

different forecasts. An ensemble forecast is likely to increase the instances of bracketing of the true value (Larrick & Soll, 2006) so that the combined error will invariably be lower than the forecast error resulting from individual judgments alone.

As mentioned earlier, when there is evidence of systematic under- or over-forecasting, asymmetric trimming is likely to be beneficial, whereas symmetric trimming is likely to lead to superior predictive performance in stable segments of the time series. This essay proposes a methodology to ensemble forecasts that dynamically adapts to the characteristics of the data series at each time period.

Consider an opinion pool consisting of n judgmental forecasts. Denote the *i*-th order statistic of the forecasts for time period *t* as $X(i)_t$. The symmetric trimming method estimates the simple average of the experts' forecasts after trimming away $\mathrm{m}/2$ of the lowest and $\mathrm{m}/2$ of the highest judgments, where $1 \leq m \leq n$ represents an arbitrarily predetermined trim size. The symmetrically trimmed forecast is therefore given by $F_t^S = \frac{1}{n-m}\sum_{i=m/2+1}^{n-m/2} X(i)_t$. The left trim forecast is denoted by the simple average of the expert's forecasts after trimming away the m lowest forecast values, i.e., $F_t^L = \frac{1}{n-m}\sum_{i=m+1}^{n} X(i)_t$. Likewise, the right trim forecast represents the mean judgment after removing the *m* highest judgments in the opinion pool, i.e., $F_t^R = \frac{1}{n-m}\sum_{i=1}^{n-m} X(i)_t$. The ensemble forecast then combines these three forecasts F_t^kderived from different trimming rules $\mathrm{k} = \{\mathrm{S}, \mathrm{L}, \mathrm{R}\}$ with their respective weights W_t^k

$$EF_t = \sum_k W_t^k F_t^k,$$

where $\sum_k W_t^k = 1$ for a given t. This section will further explain the two proposed alternative approaches for assigning weights W_t^k.: Break Judgment Ensemble and Past Performance Ensemble. Break Judgment Ensemble relies on the use of empirically elicited judgments regarding the occurrence of a structural break in the time series in period t, whereas Past Performance Ensemble derives W_t^k deterministically from past forecast error performance in time-period $t-1$.

Break Judgment Ensemble

Decomposing the judgmental forecasting task into a set of simple and less cognitively demanding tasks can improve the accuracy of forecasts (Lawerence, at al 2006; Lee & Siemsen, 2017). Extant studies suggest the benefits of task decomposition and have found decomposed judgmental forecasting tasks to be more accurate compared to obtaining forecast from a single direct step (Edmundson, 1990; Armstrong & Collopy, 1993; Webby, et al, 2005). Detecting the occurrence of structural breaks represents an important task in forecasting time series characterized by structural breaks (Rossi, 2012) and it is found that human judgments possess some diagnostic skills to effectively anticipate change (Lawrence, 1983; Edmundson, et al, 1988; Goodwin, 2007; Seifert & Hadida, 2013; Mortiz, et al, 2014). Therefore, it is proposed here to decompose the task of generating judgmental forecasts into two sub-tasks: (i) a subjective assessment regarding the perceived occurrence of a structural break, and (ii) the judgmental forecast of the target event in time-period $t+1$.

For the elicitation of subjective break predictions, suppose that forecasters observe the time series including the most recent value A_{t-1} and provide a judgment regarding the occurrence of three possible types of movements, i.e., a stationary continuation of the previous series (“no shift”), an upward movement of the mean value underlying the time series (“upward shift”) or a downward movement of the mean value underlying the time series (“downward shift”). A collective break perception of the opinion pool can then be calculated as the proportion of forecasters who choose each movement option. This aggregated break prediction is used as an informative cue for inferring the directionality of the bias in the aggregated forecast. Based on the earlier discussion, we know that upward shifts are associated with under-forecasting and downward shifts are associated with over-forecasting. If the aggregated break prediction is indicative of a perceived upward shift in the time series (i.e., the proportion of subjects who indicate an upward shift is high), then it is likely that the judgments in the opinion pool suffer from a positive bias and, hence, left trimming would be preferred (upward shift → left trim). Analogously, if the aggregated break predictions are indicative of a perceived downward shift in the time series, it is likely that judgments in the opinion pool exhibit a negative bias and, thus, right trimming would be preferable (downward shift → left trim). Lastly, if the aggregated break prediction does not indicate any structural break, symmetric trimming would yield superior performance (no shift → symmetric trim). Therefore, this study proposes the use of an ensemble of different trimming rules $\mathrm{k} = \{\mathrm{S}, \mathrm{L}, \mathrm{R}\}$

employing the empirical break predictions as weights for the respective trimming rule. The Break Judgment Ensemble forecast is given by:

$EF_t^{BP} = \sum_k W_t^k F_t^k$,

where W_t^S, W_t^{L} and W_t^{R} represent the proportion of forecasters who indicate no shift, upward shift, and downward shift, respectively.

Past Performance Ensemble

In the second approach, an ensemble mechanism based on the performance of different trimming rules in the most recent time-period is proposed. Previous studies have relied on past predictive performance to aggregate different forecasting models (Hall and Mitchell, 2007; Jore et al., 2010; Wallis, 2011). In particular, forecast aggregation weights have been derived by measuring the squared forecast errors of different forecast models (Bates & Granger, 1969; Stock & Watson, 2004; Aiolfi & Timmermann, 2006). In the context of judgmental forecasts under structural breaks, this study proposes the Past Performance Ensemble method to ensemble different trimming rules $\mathrm{k} = \{\mathrm{S}, \mathrm{L}, \mathrm{R}\}$ by using the 'prior' ranked squared error performance of the trimming rules as weights. The ensemble forecast is therefore given by:

$EF_t^{FE} = \sum_k W_t^k F_t^k$

Where,

$$W_t^k = \frac{\left[\frac{(A_{t-1}-F_{t-1}^k)^2}{\sum(A_{t-1}-F_{t-1}^k)^2}\right]^{-1}}{\sum\left[\frac{(A_{t-1}-F_{t-1}^k)^2}{\sum(A_{t-1}-F_{t-1}^k)^2}\right]^{-1}}$$

Here A_{t-1} denotes the actual observed value in period $t-1$. The Past Performance Ensemble aims at minimizing the overall error of the aggregated forecast and thus assigns the highest weight to the trimming rule with the minimum squared error in $t-1$.

2.3 Experimental Study

To test the performance of the Break Judgment Ensemble and Past Performance Ensemble methods, an experimental study was designed and conducted in which subjects provided forecasting judgments for a wide range of artificially generated time series environments containing structural breaks. The following section provides a detailed description of the study.

2.3.1 Data

The experimental design was inspired by the forecasting tasks in O'Connor et al. (1997) and O'Connor et al. (1993). The study began by generating time series with four different base levels (85, 90, 100 and 105) and a standard deviation of sd=5. Each time series contained up to two structural breaks, which shifted the base value of the series either upwards or downwards by 20 units. Each time series was decomposed into 4 segments with different sequence of shifts in the series (e.g., "flat" – "up" – "flat" – "down"). Following this logic, the study consisted of six different sequences of time series, which contained either two upward shifts, two downward

shifts, an upward shift followed by a downward shift, a downward shift followed by an upward shift, only one upward shift or only one downward shift. Figure 2-4 illustrates two examples of the types of time series employed. Consistent with the randomization strategies used in earlier studies (O'Connor, et al, 1997; O'Connor, et al, 1993), one of the four different base levels was selected as the starting point for each subject (Ver. 1 to 4 in Figure 2-5). Each subject then continued with a time-series sequence as described in the four versions in Figure 2-5.

Each time series segment can be described as follows:

Segment 1: A stationary time series for the initial 20 periods was displayed in the segment 1, which did not contain any structural breaks and was generated with one of the four predetermined base values. Segment 1 was employed to enable subjects to learn about the general characteristics of the series and no forecasts were elicited in this segment.

Segment 2: This segment represented a stationary continuation of the same mean level, in which subjects were required to provide forecasts in every fourth period during periods 21 to 32.

Segment 3: The first structural break was introduced in segment 3, which lasted from period 33 to 48. The break was positioned in various, randomly selected time periods such that the forecaster began by providing forecasts for the series for at least two time periods and then observed a mean shift of the series. The occurrences of shifts were randomized in different series that a forecaster would work on so that the

occurrences of the breaks are not similar and predictable across all the series. Forecasters were required to provide forecasts every fourth period.

Segment 4: Depending on the type of the time series, this segment (periods 52 to 70) did or did not exhibit a second structural break in the continuation of the series. Thus, there were either two similar types of structural breaks in segment 3 and segment 4 (either two sequential upward or two sequential downward breaks) or two different structural breaks in segment three and segment four (upward-downward or downward-upward). In the case where no additional structural break was introduced in segment 4, the series from the end of segment 3 remained stationary until the end of segment 4. Forecasters continued to provide forecasts every fourth period.

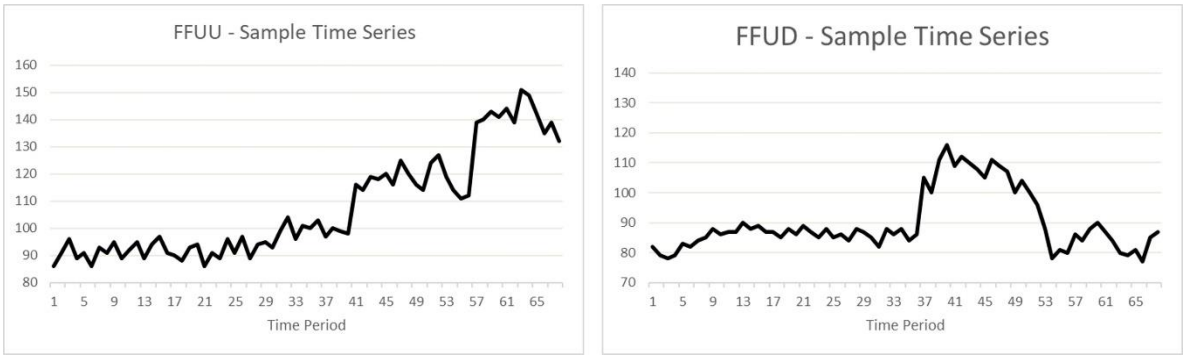


Figure 2-4: Example time series employed in the experiment

	Series 1	Series 2	Series 3	Series 4	Series 5	Series 6
Ver. 1	F – F – U – U Base -100 units	F – F – U – D Base-85 units	F – F – U – F Base -90 units	F – F – D – D Base -105 units	F – F – D– U Base -100 units	F – F – D– F Base -85 units
Ver. 2	F – F – U – D Base-90 units	F – F – D– U Base -85units	F – F – U – U Base -85 units	F – F – U – F Base -105 units	F – F – D– F Base -90 units	F – F – D – D Base -100 units
Ver. 3	F – F – U – F Base-100 units	F – F – U– D Base -105units	F – F – D – F Base -105 units	F – F – D – U Base -90 units	F – F – U– U Base -90 units	F – F – D – D Base -85units
Ver. 4	F – F – D – D Base-100 units	F – F – U– F Base -105units	F – F – U– D Base -105 units	F – F – U – U Base -90 units	F – F – D– F Base -90 units	F – F – D – U Base -85units

Legend: **F – Flat; U – Up; D - Down**

Figure 2-5: Overview of data simulation versions used in the experimental study

2.3.2 Forecasting Task

All series were displayed graphically to the subjects. After presenting the first segment, subjects observed the series progressing over time and provided the following judgments:

a) Break prediction - subjects' perception regarding the occurrence of a structural break during the last 4 observations, which was measured as a categorical variable with three possible options (no structural break, upward mean shift, downward mean shift).
b) Point forecast – subjects provided a single point estimate for t+1.
c) Prediction interval –subjects estimated lower and upper bounds of a 90 percent prediction interval.

In total, each subject provided forecasting judgments for all six time series sequences. Subjects were instructed that each series may or may not contain any number of structural breaks at unknown time periods and that one of the important tasks for them was to pay attention to such shifts and detect them while they perform

the forecasting task. The average time taken to provide all judgments was about 40 minutes. At the end of each time series sequence, subjects were provided with feedback on their performance. In particular, they observed the actual time series values in both graphical and tabular form as the series developed over time and were provided information on their performance in terms of mean absolute percentage error (MAPE).

2.3.3 Subjects

The forecasting task was implemented in Qualtrics using java-script and the subjects for the study were recruited on Amazon Mechanical Turk. Eighty-four (84) subjects (53 percent female and the average age was 34 years) were recruited for this experimental study. A few studies (e.g., Paolacci et al. 2010; Rand 2012; Paolacci and Chandler 2014; Lee et al. 2018) have addressed potential concerns regarding the data quality provided by participants in web-based experiments and have recently demonstrated that empirical findings obtained from such online platforms generally tend to be reliable. The subjects were provided with detailed instructions including examples and illustrations of the forecasting task[5], the process of providing forecasting judgments as well as descriptions of the terminologies used. Furthermore, subjects began by going through a practice exercise with a sample series lasting for two periods, which was used to familiarize themselves with the forecasting task. The participants were also able to see their performance in terms of MAPE at the end of the practice session. Afterwards, the actual forecasting task

[5]Study materials can be found in the appendix B-1

was initiated. Subjects were remunerated for the participation, and they were also incentivized for their performance based on their forecast accuracy. The participants earned $0.50 for their participation and could earn up to $2.00 extra based on their performance. The average pay-off per subject was $ 1.85

2.4 Data Analysis

To start with, a behavioral analysis of judgmental forecasting performance in response to structural breaks in time series data was conducted by looking into the following:

a) Average bias, representing the mean value of the difference between the actual value of the time series and the point forecast.
b) A classification matrix, showing the frequencies of correct versus incorrect break perception judgments.

I next analyze the prescriptive performance of the proposed ensemble methods outlined in section 2. The analysis uses the Mean Absolute Percentage Error (MAPE), which represents the average absolute difference between the true value and the forecast as a percentage of the actual value, to measure the forecast performance. In terms of group size, extant research has recommended to aggregate forecasts of between five to nine judges as larger opinion pools only marginally improve forecast performance (Gaba, et al, 2017; Jose & Winkler, 2008; Lawrence et al. 2006). The opinion pool used in the main analysis consists of $n = 9$ judges. Consistent with the trimming percentage recommended by extant studies

(Yaniv 1997; Armstrong, 2001; Hendry & Clements, 2004; Yaniv 2004), this study utilizes a trim size of $m = 2$ prior to aggregating forecast judgments.[6] Further, 1000 opinion pools were randomly sampled from the data set and the next section reports the MAPE for all the aggregation rules and the proposed ensemble methods. The performance of the Break Judgment and Past Performance Ensembles were then compared to the following alternative aggregation rules: (i) simple averaging, (ii) symmetric trimming, and (iii) the two asymmetric trimming rules separately.

2.5 Results

2.5.1 Behavioral Biases

As expected, individual forecasts were positively biased in segments subsequent to upward mean shifts with a mean bias of $\bar{b} = 12.27$ units (se= 0.33) and negatively biased in segments subsequent to observing a downward mean shift, ($\bar{b} = -13.59, se = 0.36$). Hence, the data from this study confirms to a forecasters' tendency to under-forecast in response to an upward shift and over-forecast in response to a downward shift in the mean of the time series. An example of the bias observed in upward and downward shifts from one of our series is illustrated in Figure 2-6, which shows mean forecasts across all subjects. Here, the bias is depicted by the difference in the actual (solid line) and the average forecast (dotted line). Consistent with extant research, this study shows that prior instabilities in the time series systematically influences judgmental performance (O'Connor, et al,

[6] Different sizes for the opinion pool (n = 7 and n = 11) as well as different trim sizes (m = 1 and 2) were also used, and qualitatively consistent results were obtained

1993; O'Connor, et al, 1997). The findings are also consistent with empirical studies in the domain of macroeconomic forecasting (see, for instance, Rossi, et al., 2016) as we see prior instabilities systematically influencing the bias of future forecasts. While judgmental forecasts generated in segments preceding structural breaks were also suffering from judgmental biases, they did so to a much smaller magnitude with a mean of -1.31 (*se* = 0.33). Overall, the presence of systematic judgmental biases in the data seems to support the notion that asymmetric trimming may prove to be a valuable tool for improving combined forecast performance.

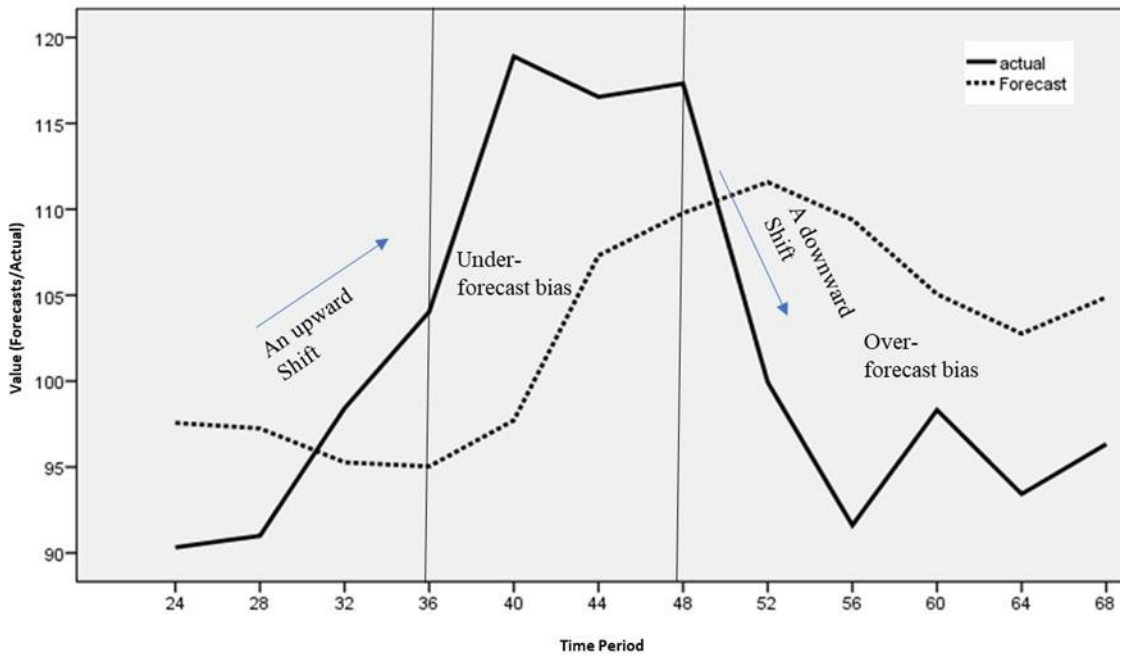


Figure 2-6: Bias in the form of under forecast and over forecast in a series with one upward shift and a downward shift from our study

Table 2-1 shows the accuracy of break perception judgments. Given that the response scale consisted of three categorical levels, by chance we would expect the classification accuracy displayed on the diagonal of Table 1 to be 1/3. A Chi-square

test indicates that judgments for upward and downward shifts are significantly different from 1/3 ($p<0.01$), suggesting that subjects systematically detected structural breaks better than by chance. Furthermore, it is observed that forecasters show a tendency of being hyper-vigilant since the rate at which subjects incorrectly judged the occurrence of a structural break was relatively high (66%). A considerable number of forecasters detected either upward or downward shift when there actually was none. This is in line with prior studies suggesting that forecasters are likely to overreact in their judgments (hyper vigilant behavior) as they have difficulties disentangling randomness from systematic change (O'Connor, et al, 1993). Also, a considerable number of forecasters did not detect a shift when there actually was one (29% for upward shift and 28% for downward shift) or even detected a shift in the opposite direction (29% for upward shift and 31% for downward shift). However, the majority of subjects (41% -42%) detected the correct direction when there is a shift.

		Actual Event		
		No Shift	**Upward Shift**	**Downward Shift**
Predicted Event	**No Shift**	1694 (34%)	123 (29%)	116 (28%)
	Upward Shift	1654 (33%)	175 (42%)	132 (31%)
	Downward Shift	1670 (33%)	122 (29%)	172 (41%)
Total		5018	420	420

Table 2-1: Empirical Break Detection Accuracy

2.5.2 Prescriptive Performance

Next, the MAPE for the two proposed ensemble methods as well as simple averaging, symmetric trim method, and each of the trimming methods alone were compared. Table 2-2 provides an overview of the forecast performance measure (MAPE) for all the six different series. Paired t-test comparisons between all forecasting methods reveal that the Past Performance Ensemble outperforms all other aggregation rules. The Past Performance Ensemble yields a nine percent lower MAPE compared to that of the simple averaging ($p<0.01$), an eight percent lower MAPE relative to the symmetric trimming ($p<0.01$), and finally a seven percent improvement ($p<0.01$) in MAPE relative to the Break Judgment Ensemble approach.

Forecast Method	**Pre-Break Segment**	**Post-Break Segment**	**Overall**
Simple Averaging	0.1223 (0.0043)	0.1959 (0.0043)	0.1692 (0.0038)
Symmetric Trimming	0.1304 (0.0046)	0.1880 (0.0041)	0.1667 (0.0038)
(Asymmetric) Left Trim	0.1400 (0.0052)	0.2195 (0.0059)	0.1921 (0.0051)
(Asymmetric) Right Trim	0.1346 (0.0040)	0.1873 (0.0032)	0.1672 (0.0030)
Break Perception Ensemble	0.1292 (0.0043)	0.1853 (0.0039)	0.1648 (0.0037)
Past Performance Ensemble	0.1022 (0.0038)	0.1620 (0.0041)	0.1533 (0.0035)

Table 2-2: Mean values of MAPE for different aggregation rules and Forecast Ensemble

Table 2-2 also illustrates that in the segment prior to the structural break, symmetrically trimmed forecasts outperformed left trim forecasts ($p<0.01$). However, there is no significant difference between symmetrically trimmed forecasts and the right trim approach, which may be related to the fact that judgments in the pre-break segment exhibited on average a marginally negative bias. In sum, these findings demonstrate the benefits of symmetric trimming for time series which are relatively stable. In the pre-break segment, the Past Performance Ensemble outperforms simple averaging ($p<0.01$), symmetrically trimming ($p<0.01$) and the Break Perception Ensemble ($p<0.01$). In post-break segments, the Past Performance Ensemble outperforms simple averaging ($p<0.01$), symmetric trimming ($p<0.01$) and the break perception ensemble ($p<0.01$). The Break Perception Ensemble, in turn, also performs better than simple averaging ($p<0.01$) (2 percent improvement) and symmetric trimming ($p<0.01$) (1 percent improvement). To sum up the key findings, the results indicate the Past Performance Ensemble to outperform all other aggregated forecasts in terms of MAPE across segments as it successfully leverages the benefits of symmetric trimming during stable time-period segments and the benefits of asymmetric trimming during post-break segments.

Further analysis looked at differences in forecast accuracy across different mean-shift patterns (Table 2-3).

Forecast Aggregation Approach	FFUU	FFUD	FFUF	FFDD	FFDU	FFDF
Simple Averaging	0.1017 (0.0037)	0.1430 (0.0031)	0.1086 (0.0044)	0.2782 (0.0145)	0.1787 (0.0019)	0.2052 (0.0073)
Symmetric trimming	0.0984 (0.0035)	0.1388 (0.0027)	0.1033 (0.0041)	0.2781 (0.0141)	0.1882 (0.0022)	0.1934 (0.0072)
Left Trim	0.0836 (0.0029)	0.1649 (0.0049)	0.0907 (0.0029)	0.3330 (0.0172)	0.1888 (0.0021)	0.2914 (0.0093)
Right Trim	0.1325 (0.0041)	0.1371 (0.0039)	0.1477 (0.0063)	0.2390 (0.0112)	0.2037 (0.0020)	0.1434 (0.0055)
Break Perception Ensemble	0.0952 (0.0036)	0.1475 (0.0033)	0.1051 (0.0042)	0.2737 (0.0135)	0.1853 (0.0021)	0.1821 (0.0066)
Past Performance Ensemble	0.0981 (0.0030)	0.1191 (0.0024)	0.0869 (0.0024)	0.2589 (0.0135)	0.1825 (0.0026)	0.1745 (0.0063)

Table 2-3: Mean values of MAPE for different forecast methods – series-wise results

Consecutive similar mean shifts: Table 2-3 indicates that for the series with two consecutive similar shifts (i.e., FFUU and FFDD), the respective asymmetric trimming rules (left and right) provides value. The appropriate asymmetric trimming rule (left trim for consecutive upward shifts and right trim for the consecutive

downward shifts) works the best ($p<0.01$) as the forecast bias is systematically one-sided (positive for upward shifts and negative for downward shifts).

Single mean shift: For the series with a single mean upward shift (FFUF), the left trim rule performs better than both simple averaging ($p<0.01$) and symmetric trimming ($p<0.01$). The Past Performance Ensemble outperforms the left trim approach. This can be attributed to the longer flat segments in this type of series and the appropriate use of the benefits of symmetric trimming. For the series with single mean downward shift (FFDF), the right trim rule performs the best compared to all other rules and ensembles ($p<0.01$). This could be attributed to the large negative bias (a mean value of -12.61 and a std. error of 0.93) associated with this series. This series had a negative bias across both the pre break and post break segments. Hence the right trim rule remains to be superior. The Past Performance Ensemble outperforms all the other aggregation rules (simple averaging, symmetric trimming, and the left trim) and the Break Judgment Ensemble ($p<0.01$). Overall, the benefits of both left and right trimming approaches in the respective cases is evident.

Consecutive different mean shifts: While forecasts for the series FFUD are on an average negatively biased for both the pre-break segment ($\bar{b} = -3.08$, se = 0.88) and for the fourth segment, i.e., the segment post the downward mean shift ($\bar{b} = -11.26, se = 0.83$), the forecasts for the time series segment post the upward break exhibit a positive bias ($\bar{b} = 10.60, se = 0.77$). We can observe an overall marginally better performance of the right trimming over simple averaging ($p<0.05$) and no significant difference between right trimming and the symmetric trimming rules.

However, the Past Performance Ensemble again clearly outperforms all other aggregation rules (p<0.01). When considering forecasts for the FFDU series, a negative bias is observed for the segment post the downward mean shift ($\bar{b} = -13.66, se = 0.87$) and a positive bias is observed for the segment post the upward mean shift ($\bar{b} = 10.00, se = 0.77$). However, the segment before the occurrence of a shift is relatively unbiased ($\bar{b} = -0.61, se = 0.72$). Though both ensemble methods perform better than the symmetric trimming rule (p<0.01), there is no significant difference with respect to the simple averaging rule. To sum up, the Past Performance Ensemble offers a robust mechanism for improving forecast performance in various time series contexts that are characterized by structural discontinuities, particularly when the direction of structural breaks is mixed and reversed.

Overall, the experimental study demonstrates the systematic under-forecasting and over-forecasting biases associated with upward and downward structural shifts in time series. Though the asymmetric trimming rule (left and right) would yield improvements in forecast performance under biased conditions (positive and negative), using these trimming rules appropriately is a challenge as the break occurrence is uncertain in terms of its timing and frequency. The proposed Past Performance Ensemble addresses this challenge by offering an approach that is intuitively appealing and simple to implement. Also, the Break Perception Ensemble does not perform as well as the Past Performance Ensemble and this could possibly be attributed to the subjects' tendency to overreact to perceived breaks and make

inaccurate break perception judgments, affecting the weighting-scheme. I explore other approaches to elicit break occurrence judgments in the third essay of this dissertation.

2.5.3 Prediction Interval Accuracy

Though this study focusses on the forecast performance in terms of MAPE, since the subjects had provided interval forecasts as well along with the point forecasts, prediction interval forecast performance was computed based on the Hit Ratio. Hit Ratio was computed as follows: when the predicted judgmental interval includes the actual time series value, it is considered as a hit and then the mean hit rate for the different heuristic methods was calculated. Higher hit rates are desirable. The analysis revealed that the Past Performance Ensemble consistently performs better than the simple averaging and the symmetric trimming rules by an average of 12 percent ($p<0.01$). The Break Judgment Ensemble model also performs better than both the simple averaging and the symmetric trimming models by 4 percent ($p<0.01$). Here again, the Past Performance Ensemble approach yields better improvement in the interval forecast performance compared to all other prescriptive rules.

2.6 Robustness Check with Real-World Time Series

In order to test the performance of the Past Performance Ensemble method beyond artificially generated time series data, this section aims to employ the forecast aggregation rules and ensemble on publicly available real world data sets from the U.S Federal Reserve Bank of Philadelphia's survey of Professional Forecasters (US SPF).

Past studies (for example Stock & Watson, 1996; Pesaran, et al, 2006; Pesaran, et al 2013; Rossi, 2013) have studied the impact of structural breaks on different macro-economic variables spanning different time series properties. In this study, on similar lines, I selected a diverse set of target variables to capture a broad range of different time series properties that could differ in terms of the frequency of structural shifts, trends, etc. Further, the time series of each of the variable selected was subject to the Bai-Perron test (Bai & Perron, 1998; Bai & Perron, 2003) to test the presence of structural breaks. The results of the Bai-Perron test indicated occurrence of multiple structural breaks across all the time series used here. This led to using 16 different macro-economic variables[7]. It can be observed that the time series are characterized by multiple upward and downward breaks and are also characterized by different time series properties such as trends and variance[8].

This study uses the US SPF survey quarterly forecast data which contains individual forecasting judgments for different quarters with number of forecasters participating in the forecasting activity varying in each quarter. Forecasters in SPF provide judgments regarding four consecutive quarters in each forecast-period and the

[7] Real GDP (RGDP), Real Personal Consumption Expenditures (RCONSUM), Real Federal Government Consumption and Gross Investment (RFEDGOV), Real Non-residential Fixed Investment (RNRESIN), Nominal Corporate Profit (CPROFIT), Real Net Export (REXP), Real Change in Private Inventories (RCBI), three-month Treasury bill rate (T-bill), 10-year Treasury Bond Rate (TBOND), Unemployment Rate (UNEMP), Nonfarm Payroll Employment (EMP) , Nominal GDP (NGDP), Real Residential Fixed Investment (RRESINV), GDP Price Index (PGDP), Average Level of Housing Starts (HOUSING), and Average Level of the Index of Industrial Production (INDPROD). Time series graphs with an indication of the occurrence of structural breaks for all variables are included in the appendix B-2.

[8] Tests for homogeneity of variance across segments were conducted for all time series using Brown-Forsythe tests on the residuals obtained from the respective time series regression models. This indicated that five variables (RGDP, RCONSUM, RFEDGOV, RNRESIN, and CPROFIT) had homogenous variance across segments (similar to the experimental set-up). These variables also had at least two or more segments with mean shifts without any trends. The other variables had varying time series properties (upward and downward trends) quite different from the experimental set-up and were also subject to non-homogeneous variances across segment (time series analysis for RCBI, TBOND, TBILL, UNEMP, EMP and HOUSING variables indicated significant ARCH effect indicating conditional heteroskedasticity and high volatility).

analysis here uses the forecast provided for the next immediate quarter from the forecast period for the analysis. Thus, each time-period t refers to a quarter during a given year. For the analyses, I use forecasting judgments provided for the next immediate quarter during a given time period. For each variable, 100 opinion pools were randomly formed, with the pool size of $n = 9$ judgments and compared the performance of the Past Performance Ensemble with the previously used benchmark methods. The chosen trim size was 2, as in the experimental study. Note that the Break Perception Ensemble is not relevant here because the SPF data does not include break perception judgments.

Results

Table 2-4 reports the MAPE for different aggregation methods for all the variables. Paired-samples t-test shows that the Past Performance Ensemble consistently outperforms ($p<0.01$) simple averaging, symmetric trimming, and the asymmetric trimming rules[9] (In the case of PGDP, the Past Performance Ensemble outperforms the symmetric trim, $p<0.05$). The proposed Past Performance Ensemble[10] yields an improvement in the range of 2 to 7 percent over the simple averaging and symmetric trimming rules. Table 2-4 also includes a time series statistical model forecast[11].

[9] There is no significant difference between Past Performance Ensemble and left trimming for RCONSUM, a series which contains sequential upward shifts over time and the asymmetric right trim rule yields the lower MAPE ($p<0.01$) for four variables (PGDP, RCBI, T-Bill and T-Bond) than the Past Performance Ensemble forecasts; these four time-series are characterized by multiple sequential down-shifts and this may lead to consistent under-forecasting and thereby the right trim may have an advantage .

[10] The Past Performance Ensemble forecasts were also computed by varying the time-windows (2 and 3 quarters) for the past forecast squared error weight calculations. The results remain qualitatively similar.

[11] The SPSS Expert Modeler (which choses the best model among different exponential smoothening and ARIMA models) was used to generate the time series model forecasts for each time-period progressively to generate forecast for every

Here it is observed that the performance of the Past Performance Ensemble is not significantly different from that of the statistical model forecast for eight variables (RFEDGOV RCONSUM, REXP, PGDP, UNEMP, TBILL, RCBI and HOUSING). However, the statistical model forecast yields better MAPE performance for the other remaining variables.

Variables	**Simple Averaging**	**Symmetric Trimming**	**Left Trim**	**Right Trim**	**Past Performance Ensemble**	**Time Series Model Forecast**
RGDP	0.0305 (0.0004)	0.0305 (0.0004)	0.0304 (0.0004)	0.0309 (0.0004)	0.0295 (0.0004)	0.0120 (0.0025)
RCONSUM	0.0151 (0.0003)	0.0152 (0.0002)	0.0149 (0.0002)	0.0157 (0.0003)	0.0148 (0.0002)	0.0128 (0.0022)
RFEDGOV	0.0265 (0.0004)	0.0269 (0.0004)	0.0271 (0.0004)	0.0283 (0.0005)	0.0261 (0.0004)	0.0248 (0.0040)
RNRESIN	0.0368 (0.0004)	0.0368 (0.0004)	0.0377 (0.0004)	0.0372 (0.0004)	0.0359 (0.0004)	0.0286 (0.0041)
CPROFIT	0.0844 (0.0007)	0.0839 (0.0007)	0.0919 (0.0009)	0.0851 (0.0007)	0.0777 (0.0007)	0.0567 (0.0084)
REXP	0.1548 (0.0018)	0.1578 (0.0018)	0.1622 (0.0016)	0.1637 (0.0022)	0.1520 (0.0018)	0.0728 (0.0071)

subsequent period. Since statistical models needs some data to initiate the model, the first 15 observations in each time series were used to initiate the model.

NGDP	0.0075 (0.00007)	0.0075 (0.00007)	0.0079 (0.00007)	0.0082 (0.00008)	0.0074 (0.00007)	0.0566 (0.0090)
RRESINV	0.0487 (0.0005)	0.0479 (0.0005)	0.0507 (0.0005)	0.0489 (0.0005)	0.0453 (0.0004)	0.03516 (0.0053)
PGDP	0.0149 (0.0003)	0.0148 (0.0003)	0.0155 (0.0003)	0.0146 (0.0003)	0.0147 (0.0003)	0.0123 (0.0029)
INDPROD	0.0477 (0.0007)	0.0480 (0.0007)	0.0492 (0.0007)	0.0476 (0.0007)	0.0468 (0.0007)	0.0188 (0.0030)
UNEMP	0.0414 (0.0002)	0.0409 (0.0002)	0.0512 (0.0002)	0.0389 (0.0003)	0.0372 (0.0002)	0.0307 (0.0017)
T-Bill	0.6247 (0.0135)	0.5238 (0.0113)	0.8965 (0.0196)	0.3671 (0.0075)	0.4373 (0.0096)	0.3741 (0.0798)
RCBI	1.5351 (0.0309)	1.5423 (0.0305)	2.031 (0.0437)	1.1342 (0.0196)	1.3527 (0.0253)	2.2516 (0.7361)
EMP	0.0047 (0.00005)	0.0045 (0.00005)	0.0045 (0.00006)	0.0049 (0.00005)	0.0042 (0.00005)	0.0015 (0.0003)
TBOND	0.1425 (0.0017)	0.1406 (0.0017)	0.1693 (0.0019)	0.1264 (0.0016)	0.1371 (0.0016)	0.1194 (0.0099)
HOUSING	0.0916 (0.0006)	0.0903 (0.0006)	0.1025 (0.0007)	0.0912 (0.0005)	0.0841 (0.0005)	0.0784 (0.0053)

Table 2-4: Mean values of MAPE for different aggregation rules and Forecast Ensemble (Real-world time-series)

Overall, it is observed that the Past Performance Ensemble consistently outperforms the symmetric trimming and simple averaging rules in the real-world time series characterized by structural breaks. These findings, therefore, generally demonstrate the robust performance of the Past Performance Ensemble and further highlights that the proposed ensemble method may indeed prove to be a valuable in real world applications.

2.7 Discussion

This chapter studies the performance of judgmental forecasts in time series environments involving structural breaks. To start with, an experimental study relying on artificially generated data was conducted. This study revealed those behavioral biases that forecasters are likely to exhibit subsequent to the occurrence of upward or downward shifts in the mean of the data series. To overcome such biases, I then proposed a novel judgment aggregation rule (asymmetric trimming) and further proposed two forecast ensembles that differed with regards to the criterion used for dynamically adapting aggregation rules to the properties of the forecasting environment. Finally, a series of real-world data contexts was used in order to test the robustness of the key findings. Forecasters systematically under-forecasted after observing structural breaks shifting the mean of the series upwards and over-forecasted when the structural break shifted the time series mean downwards. Asymmetric trimming rule is a useful means to counteract this bias while aggregating forecasts. Further the asymmetric trimming rule is particularly useful when being used as part of a Past Performance Ensemble, which identifies the most effective

trimming rule based on minimizing trimming errors on the most recent observations in the time series.

This study contributes to the descriptive research in judgmental forecasting literature to point out potential biases in forecasting time series characterized by uncertainty and structural shifts and demonstrate the presence of systematic biases in the form of under and over forecasting. Such systematic bias reduces the number of bracketing instances in forecast combination and thereby affects the benefits of aggregation. It could be seen that symmetric trimming in such biased situations often leads to the loss of valuable forecast values. This study also contributes to prescriptive research by proposing a novel forecast aggregation rule – the asymmetric trimming rule by highlighting the usefulness of the asymmetric trimming rule in retaining the most valuable forecasts while aggregating forecasts which are subject to systematic biases. This study further makes contribution to the prescriptive research in the area of forecast combination of judgmental forecasting by proposing a robust forecast ensemble approach to derive the benefits of different aggregation-trim rules. Though the use of judgmental break predictions obtained from the forecasters to combine different trimmed forecasts yielded improvements in forecast performance, the accuracy of the judgmental break predictions was found to be relatively low, and this may be the reason for only marginal improvements in the forecast performance of the Break Judgment Ensemble. Nevertheless, the Past Performance Ensemble method employing past forecast errors as weights to

combine the forecasts consistently outperforms both symmetric trimming and simple mean approach to forecast aggregation.

This study mainly contributes to the forecast combination literature by analyzing the effectiveness of different forecast combination methods under an environment characterized by unknown mean shifts which has been relatively unexplored in the forecast combination research. This study aims to bridge the two streams of judgmental forecasting and forecast aggregation by demonstrating the impact of systematic biases associated with time series characteristics on the aggregation process and prescribe a valuable ensemble mechanism to combine forecasts by addressing the potential pitfalls of the existing aggregation rules under structural shifts. The proposed forecast ensemble provides a means to enhance the use of existing symmetric trimming heuristic. It is also seen that task decomposition (in terms of obtaining break predictions) does not necessarily help in reducing the hyper-vigilant behaviour observed in earlier judgmental forecasting studies. From a methodological perspective, this study incorporates sequential breaks or discontinuities in different possibilities, and this has not been incorporated in earlier experimental designs. From a managerial perspective, the prescriptive forecast ensemble can overcome the practical difficulty of identifying super performers in a group based on post sample data and thus could be a potential valuable mechanism to aggregate forecasts in practice. Further, the Past Performance Ensemble lays out a mechanism that could be used as the basis of a machine learning-based algorithm for generating adaptive forecasts in environments where judgments are common.

This study is also subject to some limitations. The break prediction accuracy in the experimental study was quite low. Some studies in the past (for example, Matyas & Greenwood, 1990; Harvey, et al, 2017; Speekenbrik, et al, 2012) have observed high levels of false alarms (like the hyper-vigilant behavior observed in this study) in the prediction of structural shifts in time series. These studies point out that when the time series is difficult (with high autocorrelation), forecasters tend to relax their decision criteria in favour of change. Further, here, a simple majority-based aggregation approach is employed to extract 'wisdom of crowd' while using the break predictions. Recent research (for example, Prelec, 2017; Palley & Soll, 2018) has highlighted the potential flaw in such an approach in shared information contexts. In this study it was seen that some forecasters were exceptionally good, and some were extremely bad in predicting breaks along with some moderately performing individuals. The context of forecasting is characterised by information being shared by many people and in such contexts the underperformance of majority opinion is not necessarily because of a completely uniformed panel (Prelec, 2017). The simple majority method of extracting the wisdom of crowd may not be effective in this situation and there is some potential to use recent approaches such as surprisingly popular algorithm (Prelec, 2017) and pivoting (Palley & Soll, 2018). The next chapter aims to investigate this topic further.

In conclusion, in a world with constant change, understanding how to efficiently generate forecasts is of critical importance to organizations. This study offers insights on some of the behavioral biases and associated pitfalls, and further,

proposes a potential prescriptive mechanism to overcome such pitfalls and improve forecast performance over time.

3 Chapter 3: Leveraging Task Decomposition in Judgmental Time Series Forecast Aggregation

3.1 Introduction

This essay continues to draw upon the judgmental forecasting literature to study the relatively unexplored area of aggregation of time series forecasts under an environment characterized by structural change and uncertainty. The experimental study in chapter two has demonstrated the presence of systematic biases under the influence of structural breaks in time series, and chapter two has further detailed the ramifications of such biases in forecast aggregation leading to the underperformance of commonly used aggregation rules. The chapter 2 further highlights the benefits of asymmetric trimming by demonstrating how such a trimming approach retains the most valuable forecasts in the aggregation process. However, how to make use of the asymmetrically trimmed forecasts remains an interesting question. As indicated in the earlier study, eliciting human judgments about the qualitative nature of the time series, and aggregating those judgments could be a useful method to effectively use the asymmetrically trimmed forecasts.

This chapter further expands on the use of human judgments in forecast aggregation by conducting two studies to elicit human judgments about the qualitative nature of the time series. The first study elicits probability judgments about the occurrence of a break during a forecasting period from the subjects. Specifically, the task of eliciting the break judgments was decomposed by asking the subjects to estimate the

chances that the series in the most recent period to their forecasting period had shifted upward, downward, or not shifted and these probability judgments were averaged to form the weights to ensemble different trimmed forecasts. This study also tests all the prescriptive aggregation rules used in the earlier study under a time series forecasting scenario characterized by three consecutive breaks. In the second study, unlike the majority-based aggregation as in the first approach (and, also the break judgment ensemble approach used in chapter 2), a different aggregation approach, known as the Surprisingly Popular Algorithm (SPA) (Prelec, et al, 2017) is employed to develop an ensemble of the different trimmed forecasts. Although the aggregated qualitative judgments about the series in terms of the forecasters' perception about break or shift occurrence were used as a cue to combine different trimmed forecasts, it was observed that a majority-based aggregation rule might be less effective and just lead to low to moderate prediction accuracies. The forecasters show a tendency of being hyper-vigilant and hunting for discontinuities and patterns evident by the high rate (about 66 percent) of misclassification of non-breaks as breaks. Such behaviors have been found in earlier studies (for eg., O'Connor et al, 1993) where forecasters seem to find it difficult to differentiate between randomness and structural shifts. It is interesting to note that in the study under chapter two, while a few forecasters have been found to be exceptionally good at predicting the time series behavior many are quite inaccurate at the task. Such variances have existed in some of the earlier studies (eg., O' Connor, 1993) as well reflective of shared information contexts such as forecasting. The benefit of aggregation is limited by correlation of judgment errors (Clemen & Winkler, 1985) and the benefits of

averaging informative shared judgments is limited because the individuals in the group tend to repeat the same information. Under a systematically biased situation such as the one caused by the structural shifts in time series as shown in chapter two, we plausibly face a similar 'shared information problem' (Palley & Soll, 2019). The context of forecasting is characterized by information being shared by many people and in such contexts the underperformance of majority opinion is not necessarily because of a completely uniformed panel but due to a flawed method of majority-based aggregation method (Prelec, et al, 2017). The second study in this chapter thus proposes to apply a judgment aggregation method recently proposed by Prelec, et al (2017) known as the 'Surprisingly Popular Algorithm' (SPA) to the judgmental forecasting context.

Extant research has emphasized that breaking down the judgmental forecasting task structure into a set of simple tasks can improve forecast performance (Lawerence, et al, 2006; Seifert & Hadida, 2013). Armstrong (2001) suggests the benefits of task decomposition and a few other studies (such as Webby, et al, 2005; Armstrong & Collopy, 1993; Edmundson, 1990) have found decomposed judgmental forecasting tasks to be more accurate relative to holistic forecasting tasks. Further, it is seen that forecasting performance is improved in time series characterized by disturbances and breaks when the task is decomposed to gather judgments separately about special events or the causal factors (Edmundson et al, 1988; Webby, et al, 2005). Therefore, collecting information about shifts in the form of the forecaster's perception about the movement of the time series might be valuable especially, in

time series environments characterized by structural breaks. In the study in chapter two, it was seen that task decomposition and using the additional information on break perceptions were useful, although the accuracy of the aggregated judgments was relatively low. Therefore, study 1 in this essay details the break elicitation process to improve the performance of judgmental break ensemble forecasts. The break judgment elicitation process is further detailed by seeking probability associated with the different choices associated with a mean shift. This study tests if these detailed probability judgments could improve the forecast performance of the judgmental ensemble forecasts.

Study 2 in this essay, introduces an alternative elicitation procedure as suggested by Prelec (2004) wherein, in addition to providing their own judgments, the subject also makes a prediction of others' response. In the context of forecasting under discontinuities, the forecasting task is decomposed by gathering information on the forecaster's perception about the movement of the time series (similar to the earlier study in chapter 2), and further by supplementing this question by seeking the forecaster's guess about how other forecasters might respond to the same question. Prelec, et al, (2017) introduced an algorithm, (which they termed, the Surprisingly Popular Algorithm) for aggregating judgments after eliciting judgments using this alternative method. The surprisingly popular algorithm selects the choice that gains more support (in the actual world) than the predicted. By employing this alternative elicitation and aggregation method, this study aims to improve the collective wisdom

about the time series characteristic. Such aggregated break predictions are further used as a cue to combine different trimmed forecasts.

Results from the study 1 here further adds on to the robustness of the Past Performance Ensemble approach. In support of the findings in chapter 2, the Past Performance Ensemble yields consistent performance improvements compared to all other aggregation rules in a time series environment characterized by three consecutive breaks as well. Though one of the objectives of study 1 was to improve the performance of Break Judgment Ensemble by using the detailed probability judgments, the results from the study 1 here indicates that the detailed probability elicitations with respect to the break judgments do not necessarily improve the performance of the Break Judgment Ensemble. Similar to the results in the study in chapter 2, the performance of the Break Judgment Ensemble continues to be inferior to the proposed Past Performance Ensemble. The Past Performance Ensemble continues to deliver robust forecast performance. The task of eliciting detailed probability judgements associated with different break choices might have been cognitively more demanding.

Results from the study 2 show that the proposed ensemble based on the Surprisingly Popular Algorithm (Break Judgment Ensemble-SPA), improves the ensemble forecast accuracy - yielding 14 percent improvement in MAPE than simple average forecasts and 4 percent improvement in MAPE than the symmetrically trimmed forecasts. The Past Performance Ensemble continues to yield the lowest MAPE and remains consistent and robust in its performance across all the studies. Study 2

improves the Break Judgment Ensemble (an ensemble approach using the break perceptions as a cue to weigh different trimmed forecasts) by using the SPA to extract collective wisdom of forecasters about break occurrence.

Overall, this chapter contributes to prescriptive research by improving the proposed Break Judgment Ensemble and also by demonstrating the robust performance of the proposed Past Performance Ensemble method under forecasting environments subject to structural mean shifts. The application of the SPA so far has been mainly in contexts in which the correct answers are already established (factual propositions such as state capitals, etc.) and there is a need to test the potential application of the SPA in situations such as forecasting and predictions where the true answer (or true value) is in principle not known at the time of the prediction or when the decision is being made (Lee, Daneilko & Vi 2018). A few studies (for example, Lee, Daneilko & Vi 2018; Rutchick, et al, 2020) have found the SPA to be effective in sports predictions and emphasizes that more studies need to explore the performance of the SPA under different conditions and prediction contexts. In line with this call, this chapter tests the application of SPA in the area of judgmental forecasting and contributes to descriptive research by demonstrating the use of SPA in distilling crowd wisdom pertaining to prediction of unknown time series patterns.

3.2 Study 1: Detailed Break-Probability Elicitation

From the experimental study in chapter 2 it was observed that Past Performance Ensemble yields consistent performance improvements relative to the other aggregation methods. However, the Break Judgment Ensemble method is never the

best in terms of MAPE performance. This study aims to further test the performance of the ensemble methods under three consecutive structural breaks and further, aims to improve the Break Judgment Ensemble method by eliciting detailed probability judgments about any potential shifts in the time series.

Different to the earlier study, in this study probability judgments about the occurrence of a break during a forecasting period were elicited from the subjects. Specifically, the task of eliciting the break judgments is decomposed by asking the subjects to estimate the chances that the series in the most recent period to their forecasting period had shifted upward, downward, or not shifted. Therefore, average judgmental probabilities for these three options (no shift, upward shift, and downward shift) are used as the weights -W_t^k for the different trim forecast in the Break Judgment Ensemble.

An experimental study was designed in which subjects provided forecasting judgments for a wide range of artificially generated time series environments containing structural breaks. The following section provides a detailed description of this study.

3.2.1 Data

This study began by generating time series with three different base levels (100, 105, and 110) and a standard deviation of $sd = 5$. Each time series contained three structural breaks which shifted the base value of the series either upwards or downwards. In this study, the base value of the series was shifted by different

magnitudes of 25 units, 20 units and 15 units and the participants were randomly assigned to one of these three mean shifts. Each time series was decomposed into four segments, which were used to manipulate the sequence of movements in the series (e.g., “flat” - “up” – “down” – “up”). Figure 3-1 illustrates two examples of the types of time series employed. Three structural breaks were introduced in each series with the first segment being the flat segment. Following this logic, eight different sequences of time series were created, which contained three sequential upward shifts, three sequential downward shifts, two upward shifts followed by a downward shift, two downward shifts followed by an upward shift, one upward shift followed by two downward shifts, one downward shift followed by two upward shifts, or two different alternating upward and downward shifts. Each of these eight series were generated with the three different base values mentioned above. These 24-time series (eight combinations of shifts, each with three different base values) were randomly arranged in six different sequences (versions) with each sequence encompassing of four series, as shown in figure 3-2. All the three magnitudes of mean shifts had a similar setup.

Each time series segment can be described as follows:

Segment 1: A stationary time series for the initial 20 periods was displayed in the segment 1, which did not contain any structural breaks and was generated with one of the three predetermined base values. Segment 1 was employed to enable subjects to learn about the general characteristics of the series and no forecasting task was required in this segment.

Segment 2, 3 and 4: The first structural break was introduced in segment 2. The break was positioned in various, randomly selected time periods such that the forecaster began by providing forecasts for the series for at least one time-period and then observed a mean shift of the series. The occurrences of shifts were randomized in different series so that the occurrences of the breaks are not similar and predictable across all the series. Forecasters were required to provide forecasts every fourth period. The segments 3 and 4 consisted of the second and the third structural breaks, respectively. The occurrence of the shift was randomized in these segments as well.

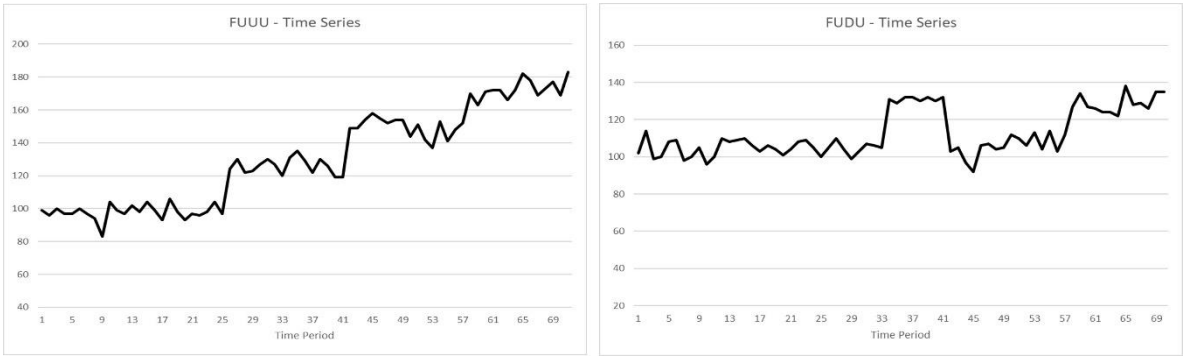


Figure 3-1: Example time series employed in study 1

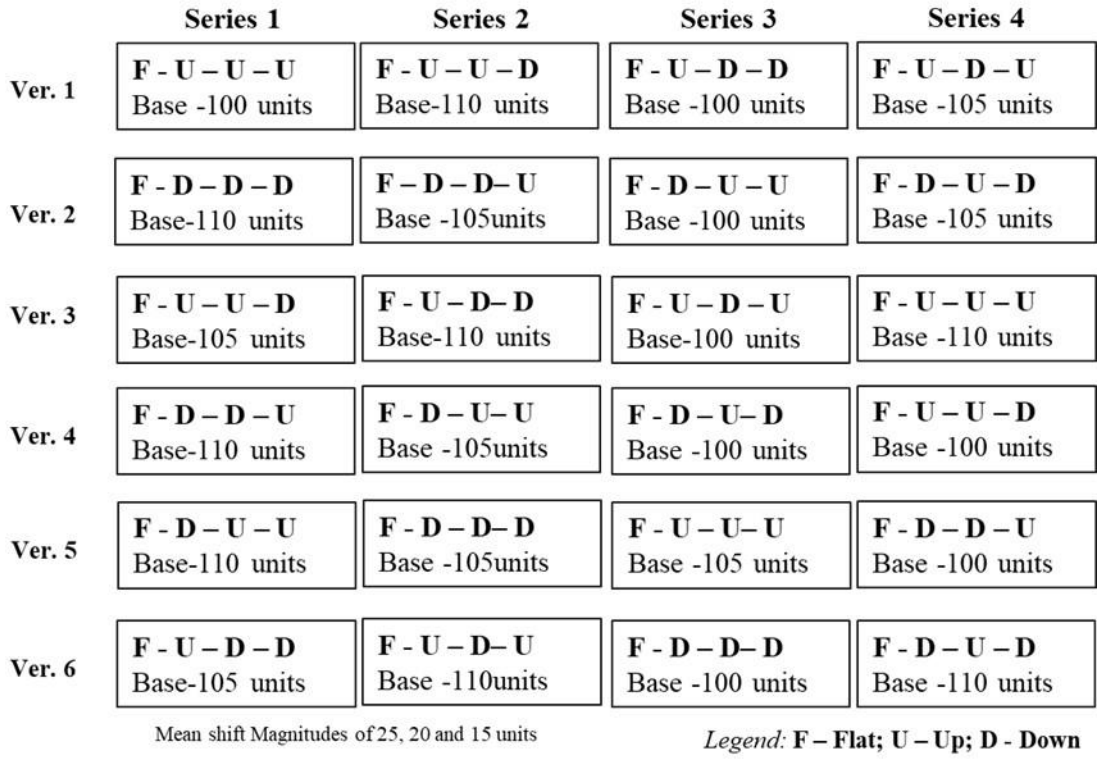


Figure 3-2: Time series sequences used in study 1

3.2.2 Forecasting Task

All time-series were displayed graphically to the subjects. After presenting the first segment, subjects observed the series progressing over time and provided the following judgments:

a) Break prediction - subjects' perception regarding the occurrence of a structural break during the most recent observations. This was captured as the probability of occurrence of an upward shift, a downward shift or a no shift. The subjects used a slider bar or entered the values between 0 and 100 for the three options (upward / Downward / no – shift). The probability values for the three options were forced to a sum of 100 percent.

b) Point forecast – subjects provided a single point estimate for t+1.

In total, each subject provided forecasting judgments for all four time series sequences. Subjects were instructed that each series may or may not contain any number of structural breaks at unknown time periods and that one of the important tasks for them was to pay attention to such shifts and detect them while they perform the forecasting task. The average time taken to provide all judgments was about 37 minutes. At the end of each time series sequence, subjects were provided with feedback on their performance. In particular, they observed the actual time series values in both graphical and tabular form as the series developed over time and were provided information on their performance in terms of mean absolute percentage error (MAPE).

3.2.3 Subjects

The forecasting task was implemented in Qualtrics using java-script and subjects were recruited on Amazon Mechanical Turk. A total of 539 subjects (41 percent female and the average age was 40 years) were recruited for this study across all the three mean shifts (magnitude 25, 20 and 15 units). The subjects were provided with detailed instructions including examples and illustrations of the forecasting task, the process of providing forecasting judgments as well as descriptions of the terminologies used. This was followed up with the subjects going through a set of questions to ensure that they understood the terminologies used in the task. Furthermore, the subjects also completed a practice exercise with a sample series lasting for two forecasting periods to familiarize themselves with the task. Subjects

were remunerated for their participation, and they were also incentivized for their performance based on their forecast accuracy. The participants earned $1.00 for their participation and could earn up to $3.00 extra based on their performance. The average pay-off per subject was $ 2.10.

3.2.4 Results

This section provides an overview of the behavioral biases observed in this study and further compares the MAPE performance for the different prescriptive rules and ensemble methods. The opinion pool used in the main analysis here consists of $n = 9$ judges. Consistent with the trimming percentage utilized in the past studies, a trim size of $m = 2$ was utilized prior to aggregating forecast judgments.[12] A random sample of 1000 opinion pools from the data set was used for MAPE comparisons.

3.2.4.1 Behavioral Biases

Similar to the findings in the study in chapter 2, the individual forecasts were found to be positively biased in segments subsequent to upward mean shifts and negatively biased in segments subsequent to observing a downward mean shift for all the three different magnitudes of mean shifts. A mean bias of $\bar{b} = 11.58$ units (se= 0.51), $\bar{b} = 8.85$ units (se= 0.48), and $\bar{b} = 9.71$ units (se= 0.39) was observed in segments subsequent to an upward shift in the three different treatments of mean shift of 25 units, 20 units and 15 units respectively. A mean bias of $\bar{b} = -11.37$ units (se= 0.50), $\bar{b} = -4.28$ units (se= 0.40), and $\bar{b} = -1.83$ units (se= 0.31) was

[12] The analysis was carried out using different sizes for the opinion pool (n = 7 and n = 11) as well as for different trim sizes (m = 1 and 2) and qualitatively consistent results were obtained.

observed in segments subsequent to a downward shift for the three different magnitudinal shifts of 25 units, 20 units and 15 units respectively. Hence, consistent evidence on forecasters' tendency to under-forecast in response to an upward shift and over-forecast in response to a downward shift in the mean of the time series were found in this study as well. The segments preceding the structural breaks had a bias close to zero (mean bias of $\bar{b} = 0.35$ units; se= 0.55; $\bar{b} = -1.05$ units; se= 0.59; $\bar{b} = 1.11$ units; se= 0.58). Overall, the presence of systematic judgmental biases in the data supports the arguments presented in chapter 2 that asymmetric trimming may prove to be a valuable tool for improving aggregated forecast performance.

In this experimental setup, subjects provided break detection judgments by providing the percentage chances associated with the occurrence of the three possible events in the most recent periods of the time series: no mean shift, upward shift, or downward shift.

Table 3-1 summarizes the accuracy of break perception judgments by displaying mean probability judgments provided by subjects). Though the break (both upward and downward) prediction accuracy is better than randomness ($p<0.01$), forecasters again show a tendency of being hyper-vigilant since the rate at which subjects incorrectly judged the occurrence of a structural break was relatively high (above 70 percent across all the three treatments).

Mean Shift Magnitude - 25 units		Actual Event		
		No Shift	**Upward Shift**	**Downward Shift**
Predicted Event	**No Shift**	28.21%	21.83%	25.85%
	Upward Shift	37.38%	47.30%	30.25%
	Downward Shift	34.41%	30.87%	43.90%
	Total	100.00%	100.00%	100.00%

Table 3-1: Break Detection Accuracy[13]

3.2.4.2 Prescriptive Performance

Next, this section compares the performance of the two ensemble methods (Break Judgment Ensemble and Past Performance Ensemble) to simple averaging and each of the trimming rules. Table 3-2 provides an overview of forecast performance for all the eight series. Paired t-test comparisons reveal that the Past Performance Ensemble yet again outperforms all other rules. Consistent with the study in chapter 2, the Past Performance Ensemble yields lower MAPE compared to simple averaging ($p<0.01$) (32 percent improvement), and a lower MAPE relative to symmetric trimming ($p<0.01$) (11 percent improvement). The Past Performance Ensemble also yields a lower MAPE relative to the Break Perception Ensemble ($p<0.01$). Table 3-1 also illustrates that in the pre-break segment, symmetric trimming outperformed both the left and right trim forecasts ($p<0.01$). Overall, the Past Performance Ensemble outperforms simple averaging ($p<0.01$), symmetric

[13] The break prediction accuracy is qualitatively similar in both the other treatments with 20- and 15-units magnitude-shifts and the same is provided in Appendix C.

trimming ($p<0.01$) and the Break Perception Ensemble ($p<0.01$). The Past Performance Ensemble yields robust performance improvements under different magnitudes of mean shifts. However, while the Break Perception Ensemble performs better than simple averaging ($p<0.01$), it yields higher MAPE in comparison to the symmetric trimming rule.

Mean Shift Magnitude - 25 units

Forecast Aggregation Approach	**Pre-Break Segment**	**Post-Break Segment**	**Overall**
Simple Averaging	0.0550 (0.0012)	0.1482 (0.0030)	0.1507 (0.0037)
Symmetric trimming	0.0479 (0.0010)	0.1148 (0.0020)	0.1155 (0.0025)
Left Trim	0.0769 (0.0018)	0.2063 (0.0058)	0.1962 (0.0069)
Right Trim	0.0791 (0.0017)	0.1348 (0.0019)	0.1417 (0.0019)
Break Perception Ensemble	0.0545 (0.0012)	0.1312 (0.0025)	0.1322 (0.0030)
Past Performance Ensemble	0.0374 (0.0012)	0.1027 (0.0014)	0.1025 (0.0015)

Table 3-2: Mean values of MAPE for different aggregation approaches[14]

To sum up the key findings, the detailed probability elicitations does not improve the Break Judgment Ensemble approach. The performance of the Break Judgment Ensemble is inferior to both Past Performance Ensemble and symmetric trim

[14] These results are for trim size of 2 and the results for trim size 1 is qualitatively similar. Results are qualitatively similar for mean-shifts of 20 and 15 and these are provided in Appendix C.

approaches. This study finds that the Past Performance Ensemble appears to be robust across segments as this approach leverages the benefits of symmetric trimming during relatively stable time-period segments as well as the benefits of asymmetric trimming to overcome potential biases found in the periods subsequent to the occurrence of structural breaks.

Mean Shift Magnitude - 25 units

Forecast Aggregation Approach	**FUUU**	**FUUD**	**FUDD**	**FUDU**	**FDDD**	**FDDU**	**FDUU**	**FDUD**
Simple Averaging	0.1070 (0.0047)	0.0921 (0.0036)	0.1484 (0.0048)	0.0850 (0.0019)	0.3840 (0.0150)	0.1757 (0.0047)	0.0971 (0.0019)	0.1165 (0.0023)
Symmetric trimming	0.0830 (0.0045)	0.0801 (0.0033)	0.1307 (0.0061)	0.0764 (0.0023)	0.2995 (0.1360)	0.1268 (0.0026)	0.0954 (0.0018)	0.0989 (0.0021)
Left Trim	0.0469 (0.0016)	0.0706 (0.0015)	0.1063 (0.0034)	0.0913 (0.0019))	0.6483 (0.0257)	0.3047 (0.0105)	0.1167 (0.0041)	0.1848 (0.0047)
Right Trim	0.1915 (0.0076)	0.1624 (0.0069)	0.1895 (0.0062)	0.1128 (0.0034)	0.1323 (0.0027)	0.1190 (0.0027)	0.1327 (0.0029)	0.0886 (0.0019)
Break Perception Ensemble	0.0954 (0.0046)	0.0798 (0.0028)	0.1365 (0.0049)	0.0813 (0.0020)	0.3181 (0.0123)	0.1493 (0.0033)	0.0902 (0.0015)	0.1071 (0.0025)
Past Performance Ensemble	0.0627 (0.0023)	0.0752 (0.0029)	0.1084 (0.0041)	0.0726 (0.0021)	0.1817 (0.0037)	0.1235 (0.0022)	0.0905 (0.0019)	0.1001 (0.0017)

Table 3-3: Mean values of MAPE for different aggregation approaches across different series

Table 3-3 offers a summary decomposed into the eight different time series segments. When examining forecasting behavior in consecutive time series

segments that are associated with the same mean shift (i.e., FUUU and FDDD), asymmetric trimming yields the lowest MAPE ($p<0.01$) and the Past Performance Ensemble outperforms the other aggregation methods by leveraging the respective asymmetric trims appropriately. We can observe that in series with alternating upward and downward shifts (i.e., FUDU and FDUD) the Past Performance Ensemble takes advantage of both left and right trim approaches. As a result, it yields better performance ($p<0.01$) than other aggregation rules for FUDU and FDUD (there is no significant difference between the Past Performance Ensemble and symmetric trimming rule for FDUD). The sequence of shifts has an impact on the bias – similar consecutive shifts tend to sustain and increase the bias in the same direction. For example, for FUUD, we observe a mean bias of $\bar{b} = 13.2$ (se = 1.41) post the first break and a mean bias of $\bar{b} = 21.8$ (se =2.51) post the second break and finally after the third break in the opposite direction, the bias tends to move in the opposite direction (with a mean bias not significantly different from zero, $\bar{b} = 3.06$; se = 1.58). A similar pattern is observed in the reverse direction with the FDDU series and also with series with different order of break sequences like FUDD and FDUU. We can therefore see that the left trim rule (as an example for series FUUD) and the right trim rule (as an example, for series FDDU) yield value in these respective cases. We can further note that the Past Performance Ensemble[15] quite aptly leverages these asymmetric trim forecasts to account for such variations in

[15] Past Performance Ensemble performs the best ($P<0.01$) compared to the simple averaging, symmetric and both the asymmetric trimming rules for FDUU and yields the second-best performance to the right trim ($p<0.05$) for FUDD.

break sequences to yield better forecast performance compared to that of the commonly used simple averaging and symmetric trimming aggregation rules.

Overall, the study concurs with the observations from the study in chapter 2. This study re-emphasises the benefits associated with all the three trimming approaches (left, right and symmetric) depending on the shifts and the bias associated with the forecasts. Though the Break Judgment Ensemble provides consistent improvement in forecast performance compared to the simple averaging approach, its performance is inferior to both symmetric trimming and Past Performance Ensemble method. The Past Performance Ensemble yields robust forecast performance improvements across all the treatments in this study. This study, similar to the study in chapter 2 finds high levels of false alarms in that a break is predicted when there was none. This reduced break prediction accuracy affects the forecast ensemble's weighing scheme. Some studies in the past (for example, Matyas & Greenwood, 1990; Harvey, et al, 2017; Speekenbrik, et al, 2012) have observed high levels of false alarms in the prediction of structural shifts in time series. These studies point out that when the time series is difficult, forecasters tend to relax their decision criteria in favor of change. There could be different cognitive processes leading to the high rates of false alarms (Furlong & Wampold, 1981). One of the possible reasons for forecasters to read too much into random variations is that they often make local assessments and judge candidate changes in isolation from the rest of the series (Brown & Steyvers, 2005). Another explanation could be that when forecasters assess a recognized change in the time series relative to the overall

variability in the series by using a threshold criterion, they may ignore the positive autocorrelation in the data series leading to an underestimation of the overall variability and thereby an over-estimation of the likelihood of the change (Speekenbrink, et al, 2012).

The next study in this chapter proposes an alternative approach to elicit and combine judgmental break predictions from forecasters.

3.3 Study 2: Applying the Surprisingly Popular Algorithm

This study focusses on employing an alternative elicitation and break judgment aggregation procedure. The break judgment aggregation approaches used earlier have been based on a majority rule with higher weight being given to majority opinion in a group. Under a systematically biased situation such as the one caused by the structural shifts in time series as shown in chapter two and study 1 in this essay, we plausibly face a 'shared information problem' (Palley & Soll, 2019). The context of forecasting is characterized by information being shared by many people and in such contexts the underperformance of majority opinion is not necessarily because of a completely uniformed panel but due to a flawed method of majority-based aggregation method (Prelec, et al, 2017). In such contexts the benefits of averaging are limited because a significant part of the average judgment contains the same information being repeated (Palley & Soll, 2019). This study aims to apply a judgment aggregation method proposed by Prelec, et al (2017) known as the 'Surprisingly Popular Algorithm' (SPA) to the judgmental forecasting context. Further, this study

prescribes a forecast ensemble approach by leveraging these aggregated judgmental predictions.

The SPA combines the cognitive judgment of the decision maker and the meta-cognitive judgment in the form of an estimate of others' judgments in a shared information context. In the alternative elicitation procedure used in this study, as suggested by Prelec (2004), in addition to providing their own judgments, the subject also makes a prediction of others' break perception. In the context of forecasting under discontinuities, the forecasting task is decomposed by gathering information on the forecaster's perception about the movement of the time series, and further supplement this question by seeking the forecaster's guess about how other forecasters might respond to the same question. Prelec, et al, (2017) introduced the surprisingly popular algorithm (SPA), for aggregating judgments after eliciting judgments using this alternative method. The surprisingly popular algorithm selects the choice that gains more support (in the actual world) than the predicted. By employing this alternative elicitation and aggregation method, this study aims to improve the accuracy of aggregated break predictions or the collective wisdom about the time series characteristic. Such aggregated break predictions shall be used as a cue to aggregate different trimmed forecasts.

3.3.1 Revised forecast aggregation method

In the revised forecast aggregation method, both symmetrically trimmed forecast and asymmetrically trimmed forecasts are used. In the first step, we shall decide on the direction of the trim for asymmetrically trimming the forecast, i.e., whether to use a

left trim forecast or a right trim forecast. As seen in the earlier studies in this chapter and chapter 2, the under-forecasting and over-forecasting behavior of a forecaster leads to systematic biases post a break. The earlier results and bias patterns have clearly indicated that the mean biases of the forecasters persist over certain time periods even after the occurrence of a break. The bias reduces over time as the series stabilizes. Earlier studies have also seen that mean forecasts typically underreacts to new information and the direction of the bias continues over a few time periods in the time series (for ex., Kim et al, 2000). The bias persists over a transition phase involving several periods post a break (Becker, et al, 2009). Prior mean forecast typically serves as a (imperfect) signal of the information common to all forecasters and the difference between prior mean forecasts and subsequent mean forecasts acts as a reasonable indicator of the bias in the forecasts (Kim et al, 2000). Thus, bias in the past can be used as a cue to decide between the left and the right trimmed forecasts. As a second step we shall use the selected asymmetrically trimmed forecast (from step 1) or the symmetrically trimmed forecast based on the aggregated break predictions from the surprisingly popular algorithm – i.e. if the aggregated prediction derived from the SP method indicates a break scenario as the extracted collective wisdom, we shall use the asymmetrically trimmed forecast or if the extracted collective wisdom indicates a no-break scenario, we shall use the symmetrically trimmed forecast. So basically, this study proposes a simple choice model as follows:

Step 1:

$AS_t^+ = F_t^L$, if $\bar{b}_{t-1} > 0$ or

$AS_t^+ = F_t^R$, if $\bar{b}_{t-1} < 0$

Step 2:

$EF_t^{SP} = AS_t^+$, if $P_i = 1$,

Else, $EF_t^{SP} = F_t^S$

Where,

EF_t^{SP} denotes the aggregated forecast for the time-period t,

AS_t^+ denotes the asymmetrically trimmed forecast for time period t; + denotes the direction of the asymmetrically trimmed forecast, i.e., left trim or right trim arrived based on step 1 described above.

F_t^L , F_t^R, and F_t^S denotes left trimmed, right trimmed and symmetrically trimmed forecasts respectively.

P_i is the aggregated break prediction derived based on the SP Algorithm taking values of 1 (a break scenario) or 0 (a no break scenario)

To illustrate the SP Algorithm, let us consider the following scenario:

y_t = aggregate percentage of subjects predicting occurrence of a break (yes, a break occurred) while making the forecast for time-period t.

E_t^y = aggregate percentage of subjects' expectation that others also predict the occurrence of a break, while making the forecast for time-period t.

As per the SP Algorithm,

If [y_t - E_t^y] > [(1 - y_t) – (1- E_t^y)], P_i = 1 (break occurred), else P_i = 0 (no break)

3.3.2 Data

The experiment design is similar to the study employed in the chapter 2 with time series with structural mean shifts – both upward and downward shifts. The study incorporates six different types of time series with different structural break combinations - series with two upward shifts, series with two downward shifts, series with both upward and downward shifts in both the possible combinations, series with one upward shift and finally series with one downward shift. Four different base levels (viz., 85, 90, 100 and 105) with a standard deviation of five units were used. Each structural break is a mean shift of 20 units in the upward or downward direction. Each series is divided into four segments:

Segment 1: The time series for the initial 20 periods was displayed in the segment. The time series does not consist of any structural breaks and is generated with one of the four base values. This segment enables the forecaster to observe the time series and assess its characteristics.

Segment 2: This segment was basically a continuation of the series displayed in the first segment, again without any breaks. However, now in this segment, the forecaster made forecasts for every fourth period during the periods 21-32.

Segment 3: The first structural break was introduced in segment 3, which lasted from period 33 to 48. The break was positioned in various, randomly selected time periods such that the forecaster began by providing forecasts for the series for at least two time periods and then observed a mean shift of the series. The occurrences of shifts were randomized in different series and forecasters were required to provide forecasts every fourth period.

Segment 4: Depending on the type of the time series, this segment (periods 52 to 70) did or did not exhibit a second structural break in the continuation of the series. Thus, there were either two similar types of structural breaks in segment 3 and segment 4 (either two sequential upward or two sequential downward breaks) or two opposing structural breaks in segment three and segment four (upward-downward or downward-upward). In the case where no additional structural break was introduced in segment 4, the series from the end of segment 3 remained stationary until the end of segment 4. Forecasters were required to provide forecasts every fourth period.

Each forecaster provided forecasts for three types of series. The combination of all the series was randomized and six random combinations of the series were formed. This provides us with six versions, or three different series combinations and the forecasters were assigned to these versions randomly. As shown below in Figure 3-3, a forecaster was randomly assigned to one of the six versions, with each version comprising of three different series. Each series consisted of four segments described above in the form of flat segments (segments without any structural break)

or segments with the break (upward and downward mean shifts). The different base values used for each time series are also provided.

	Series 1	Series 2	Series 3
Ver. 1	F – F – U – U Base -100 units	F – F – U – D Base-85 units	F – F – U – F Base -90 units
Ver. 2	F – F – D – D Base-100 units	F – F – D– U Base -85units	F – F – U – U Base -85 units
Ver. 3	F – F – U – F Base-100 units	F – F – U– D Base-105 units	F – F – D – F Base-105 units
Ver. 4	F – F – D – D Base-100 units	F – F – U– F Base -105units	F – F – U– D Base -105 units
Ver. 5	F – F – D – D Base-105 units	F – F – D– U Base -100units	F – F – D– F Base -85 units
Ver. 6	F – F – U – U Base-105 units	F – F – D– F Base -100units	F – F – D– U Base -105 units

Legend: **F – Flat; U – Up; D - Down**

Figure 3-3: Time-series sequences used in study 2

3.3.3 The Forecasting Task

All series were displayed graphically to the subjects. After presenting the first segment, subjects observed the series progressing over time and provided the following judgments:

a) Break prediction - their perception about the occurrence of a significant shift or discontinuity in the series – Yes or No

b) Confidence - their confidence about their perception in part a (50 – 100%)

c) Guess about other's response – what percentage of other people thought if there was a shift or discontinuity in part a (0-100%)
d) Point forecast – the forecasters provide an estimate of the point forecast for the next time-period t+1.

In total, each subject provided forecasting judgments for all three time series sequences. Subjects were instructed that each series may or may not contain any number of structural breaks at unknown time periods and that one of the important tasks for them was to pay attention to such shifts and detect them while they perform the forecasting task. The average time taken to provide all judgments was about 34 minutes. At the end of each time series sequence, subjects were provided with feedback on their performance in terms of mean absolute percentage error (MAPE).

3.3.4 Subjects

The forecasting task was implemented in Qualtrics using java-script and international MBA students were recruited from the international MBA program at IE Business School. These students were participants of the quantitative methods course and were trained in basic statistics and forecasting techniques. The students had an option to volunteer and participate in this study. A total of 153 subjects (35 percent Female, an average age of 28 years and an average work experience of 6.5 years) were recruited for this study. The subjects were provided with detailed instructions including examples and illustrations of the forecasting task, the process of providing forecasting judgments as well as descriptions of the terminologies used. This was followed up with the subjects going through a set of questions to ensure that they

understand the terminologies used in the task. Furthermore, the subjects had a practice exercise with a sample series lasting for two forecasting periods, which they used to familiarize themselves with the task. The participants had an incentive to earn 5 bonus points in their Quantitative Methods course. In addition, top five participants (participants with the lowest MAPE) also had a chance to win Amazon gift vouchers worth $ 50.

3.3.5 Results

This section provides an overview of the behavioral biases observed in this study and further compare the MAPE performance for the different prescriptive rules. The opinion pool used in the main analysis here consists of $n = 9$ judges. Consistent with the trimming percentage utilized in the past studies, a trim size of $m = 2$ was utilized prior to aggregating forecast judgments.[16] A random sample of 1000 opinion pools from the data set was used and MAPE and standard error for each tested aggregation rule is reported below.

3.3.5.1 Behavioral Biases

Similar to the findings in the earlier studies, this study also found consistent evidence on the forecaster's tendency to systematically under-forecast in response to an upward mean shift and to systematically over-forecast in response to a downward shift in the mean value of the time series. The forecasts post an upward mean shift was positively biased with a mean bias of $\bar{b} = 8.53$ units (se= 0.42), and the

[16] The analysis was carried out using different sizes for the opinion pool (n = 7 and n = 11) as well as for different trim sizes (m = 1 and 2) and qualitatively consistent results were obtained.

forecasts post a downward mean shift was negatively biased with a mean $\bar{b} = -6.54$ units (se= 0.35). The forecast during the pre-break period remained to be relatively un-biased (mean $\bar{b} = 0.58$ units, se= 0.29).

In this experimental setup, subjects provided break detection judgments by assessing whether a break occurred in the most recent periods of the time series: no mean shift, or a mean shift[17]. The classification matrix in Table 3-4 summarizes the accuracy of break detection judgments. If subjects chose options by chance, we would expect the classification accuracy displayed on the diagonal of Table 3-4 to be 1/3. Considering this, the data shows that empirically elicited judgments lie consistently above chance (Chi-square test significant at $p<0.01$), suggesting that subjects were capable of correctly detecting a fair number of structural breaks. The break judgment accuracy is higher than all the earlier studies (in chapter 2 and chapter 3). Specifically, we could see that unlike the earlier studies, here we do not see the hyper-vigilant tendency of the forecaster – indicated by the high accuracy of predictions when there was no shift or break in the time series (accuracy of 68 percent). A possible reason for this could be the simplified 'break-question' (binary in nature – break or no break). Decomposing the judgmental forecasting task into simple steps can improve accuracy (Lawerence, 2006) and the break prediction task here in this study is cognitively less demanding (Lee & Siemsen, 2017) compared to the task in the earlier studies.

[17] In this study, the direction of the break (if the forecaster predicts the occurrence of a shift) is determined by the change in the point forecast made in the forecasting period with respect to the forecast made in the prior period.

		Actual Event		
		No Shift	**Upward Shift**	**Downward Shift**
Predicted Event	**No Shift**	68.00%	34.40%	37.80%
	Upward Shift	15.5%	56.80%	7.30%
	Downward Shift	16.5%	8.80%	54.90%
	Total	100.00%	100.00%	100.00%

Table 3-4: Break Prediction Accuracy (Study 2)

3.3.5.2 Prescriptive Performance

This section compares the forecast performance of the different aggregation rules and (simple averaging, symmetric trimming) and the prescriptive forecast ensemble methods.

Table 3-5 provides an overview of the forecast performance measure (MAPE) for all the six different series. Paired t-test comparisons between all forecasting methods reveal that the Past Performance Ensemble outperforms all other models. It can be sees that consistent with the results in the earlier studies in chapter 2 and chapter 3, the Past Performance Ensemble yields lower MAPE compared to that of simple averaging ($p<0.01$) (19 percent improvement averaged across all the series) and yields a lower MAPE relative to the symmetric trimming rule ($p<0.01$) (9 percent improvement averaged across all the series). Though the Break Judgment Ensemble yields a lower MAPE compared to simple averaging ($p<0.01$) (10 percent improvement averaged across all the series), there is no significant difference in MAPE compared to the symmetric trimming rule.

This study has aimed to improve the Break Judgment Ensemble by eliciting break occurrence judgments and aggregating the same using the Surprisingly Popular Algorithm. This new proposed ensemble forecast method is referred to as the Break Judgment Ensemble-SPA here. The new proposed ensemble yields improvement in forecast performance. Break Judgment Ensemble-SPA yields lower MAPE compared to that of simple averaging ($p<0.01$) (14 percent improvement averaged across all the series) and yields a lower MAPE relative to the symmetric trimming rule ($p<0.01$) (4 percent improvement averaged across all the series). The Past Performance Ensemble, however, is the best and yields a five percent improvement than the Break Judgment Ensemble-SPA.

Forecast Aggregation Approach	**Pre-Break Segment**	**Post-Break Segment**	**Overall**
Simple Averaging	0.0643 (0.0009)	0.1001 (0.0010)	0.0898 (0.0007)
Symmetric Trimming	0.0564 (0.0007)	0.0863 (0.0009)	0.0797 (0.0007)
Left Trim	0.0622 (0.0011)	0.1155 (0.0013)	0.1002 (0.0010)
Right Trim	0.0905 (0.0012)	0.1170 (0.0012)	0.1094 (0.0009)
Break Judgment Ensemble	0.0597 (0.0008)	0.0861 (0.0009)	0.0801 (0.0007)
Break Judgment Ensemble-SPA	0.0589 (0.0008)	0.0839 (0.0008)	0.0767 (0.0006)
Empirical Performance Ensemble	0.0598 (0.0007)	0.0781 (0.0008)	0.0728 (0.0006)

Table 3-5: Mean values of MAPE for different aggregation approaches

As per expectations, Table 3-5 also illustrates that in the segment prior to the structural break, symmetrically trimmed forecasts outperformed both the left and right asymmetrically trimmed forecast ($p<0.01$). The benefits of symmetric trimming for time series which are relatively stable can be observed here. Also, the proposed forecast ensemble methods (both Past Performance Ensemble and Break Judgment Ensemble-SPA) outperform the simple averaging, symmetric trimming, asymmetric trimming rules and the basic Break Judgment Ensemble.

Table 3-6 offers a summary of the MAPE metric decomposed into the six different time series segments.

Consecutive similar mean shifts: Table 3-6 indicates that for the series with two consecutive similar shifts (i.e., FFUU and FFDD), the respective asymmetric trimming approach (left and right) provides value. The appropriate asymmetric trimming method (left trim for consecutive upward shifts and right trim for the consecutive downward shifts) works the best ($p<0.01$) as the forecast bias is systematically one-sided (positive for upward shifts and negative for downward shifts).

Consecutive different mean shifts: For both the series here (FFUD and FFDU), both the ensemble methods (Break Judgment Ensemble-SPA and Past Performance Ensemble) outperform simple averaging, symmetric trimming and both the asymmetric trimming rules ($p<0.01$).

Forecast Aggregation Approach	FFUU	FFUD	FFUF	FFDD	FFDU	FFDF
Simple Averaging	0.0813 (0.0015)	0.0953 (0.0018)	0.0804 (0.0017)	0.1022 (0.0021)	0.0980 (0.0017)	0.1142 (0.0027)
Symmetric Trimming	0.0678 (0.0012)	0.0818 (0.0016)	0.0636 (0.0013)	0.0834 (0.0019)	0.0943 (0.0018)	0.0922 (0.0025)
Left Trim	0.0476 (0.0008)	0.1215 (0.0028)	0.0457 (0.0008)	0.0918 (0.0018)	0.1133 (0.0023)	0.1879 (0.0037)
Right Trim	0.1259 (0.0025)	0.1007 (0.0020)	0.1367 (0.0028)	0.0700 (0.0014)	0.1159 (0.0019)	0.0792 (0.0019)
Break Judgment Ensemble	0.0691 (0.0012)	0.0825 (0.0016)	0.0680 (0.0015)	0.0810 (0.0017)	0.0969 (0.0017)	0.1031 (0.0027)
Break Judgment Ensemble-SPA	0.0595 (0.0009)	0.0798 (0.0016)	0.0619 (0.0012)	0.0702 (0.0014)	0.0908 (0.0017)	0.0917 (0.0024)
Empirical Performance Ensemble	0.0588 (0.0009)	0.0738 (0.0014)	0.0504 (0.0008)	0.0733 (0.0015)	0.0867 (0.0016)	0.0885 (0.0022)

Table 3-6: Mean values of MAPE for different aggregation approaches across different series

Single mean shift: For the series with a single mean shift (FFUF and FFDF), the left trim rule (for FFUF) and the right trim rule (for FFDF) perform better than both simple mean ($p<0.01$) and symmetric trimming ($p<0.01$) aggregation rules. The Past Performance Ensemble outperforms all the three aggregation rules – simple

averaging, symmetric trimming and both the Break Judgment Ensemble methods ($p<0.01$). The Break Judgment Ensemble-SPA outperforms simple averaging rule, symmetric trimming rule and the Break Judgment Ensemble ($p<0.01$); however, it does not yield a significant difference in MAPE compared to the symmetric trimming for the FFDF series.

To sum up, it can be observed that the proposed ensemble based on the Surprisingly Popular Algorithm, improves the ensemble forecast using break judgments as cue to combine forecasts from different aggregation rules. The novel approach to extract crowd wisdom has been found to be useful in judgmental forecasting applications. The Past Performance Ensemble has remained quite consistent and robust in its MAPE performance.

3.3.6 Discussions

This chapter aimed at improving the Break Judgment Ensemble method, one of the proposed forecast ensemble methods in this dissertation primarily. In order to achieve this objective, two studies were conducted to explore two different approaches to elicit judgments about occurrence of structural breaks in time series forecasting. Study 1 used a break elicitation procedure with detailed probability choices to obtain perceptions on the movement of the times series and these detailed probability break judgments were used as the weighing-scheme for the forecast ensemble. Study 2 elicited break judgments as a combination of cognitive and meta-cognitive judgments and applied the Surprisingly Popular Algorithm to the combine forecasts in the ensemble forecast. While the detailed probability judgments

did not significantly improve the performance of the Break Judgment Ensemble forecasts, the application of the SPA yielded significant improvement to the Break Judgment Ensemble forecasts. This study also demonstrates the robust performance of the proposed Past Performance Ensemble method under different extended break conditions. The Past Performance Ensemble method yields the lowest MAPE consistently across all the studies. The improved Break Judgment Ensemble-SPA method yields consistent lower MAPE than all other commonly used aggregation rules. Overall, the ensemble approach with prior forecast error as weights combining the three trimmed forecasts appropriately offers a robust mechanism to improve forecast performance in time-series environments characterized by structural mean shifts. In addition, the proposed improvement to the break judgment ensemble method by leveraging the Surprisingly Popular Algorithm has yielded consistent improvement in forecast performance and provides an approach to use judgmental predictions as a cue to combine forecasts effectively under systematically biased environments.

This chapter primarily contributes to prescriptive research by refining and fine tuning the proposed Break Judgment Ensemble mechanism and by demonstrating the consistent performance of the proposed Past Performance Ensemble mechanism. Both these forecast ensemble approaches make effective use of asymmetrically trimmed forecasts to overcome systematic biases commonly observed in forecasting environments characterized by structural breaks. This study contributes to the judgmental forecasting literature by testing the application of a novel judgment-

aggregation method, the SPA, to elicit dynamic time series characteristics in forecasting tasks. The application of the SPA so far has mainly been limited to contexts with factual questions in extant research (Lee, Daneilko & Vi 2018; Rutchick, et al, 2020) and this study has extended its application to judgmental time series forecasting aiming to address the call for testing the SPA in real world forecasting and prediction applications.

This study has been limited to the prediction of one characteristic of the time series (that of, the occurrence of breaks) and has been limited to a single experimental study. I hope that the results in this study would encourage future studies to investigate the prediction of other time series characteristics using the SPA and further test the application of the same in different contexts. The studies conducted here have been purely on artificially generated time series and therefore, it would be prudent to extend these studies to real-time environments to improve the external validity.

In conclusion, in a world of constant change, it is important to generate more accurate forecasts and this chapter along with chapter 2 offers novel methods that are intuitively appealing and easy to implement in real-world forecasting environments.

Conclusion

This dissertation aims to meet the call for growing need to study decision rules and aggregation mechanisms used in extracting collective wisdom in wisdom of crowd applications. The process of distilling collective wisdom is often challenged by the cognitive limitations and systematic biases of the decision maker imposed by decision rules and aggregation mechanisms. The main objective of my dissertation was to understand the impact of individual biases and/or cognitive limitations on collective judgments and to propose prescriptive mechanisms to aggregate individual judgments. I focused on two prominent areas of wisdom of crowd applications from both research and practice perspectives – open innovation management and judgmental forecasting.

In the first chapter, I focused on idea evaluation and selection in the context of open innovation. The study compared two commonly employed idea evaluation decision rules (scoring and ranking) and provides insights into how the efficacies of these decision rules are influenced by the number of ideas to be evaluated and the amount of available information. At an aggregated level, the scoring decision rule can enable higher levels of accuracy with relatively smaller crowd size as the scoring process is more effective for evaluating ideas when there are several ideas to be evaluated. This study contributes to descriptive research by highlighting the potential cognitive limitations associated with the use of different decision rules used in idea evaluation & selection, a relatively unexplored space in new product development and open innovation. The essay further makes prescriptive contribution by providing insights

into the design elements of open innovation environments (for example, number of ideas to be presented, crowd size estimations, etc).

The latter two chapters focus on a relatively unexplored, yet an important area of judgmental forecasting of time series characterized by structural breaks. These essays bridge the two streams of judgmental time series forecasting and wisdom of aggregated forecasts. These studies analyzed the performance of different forecast aggregation methods and finds that commonly used aggregation methods such as simple averages and symmetric trimming may be less effective in forecast environments characterized by structural breaks. Further, chapter two and three proposed a novel form of asymmetric trimming aggregation rule (for point forecasts) to overcome the limitations of traditional aggregation methods. These chapters collectively contribute to descriptive research by demonstrating the presence of systematic biases in forecasts that are caused by time series discontinuities such as structural breaks under judgmental forecasting tasks and by highlighting how the presence of such biases can affect commonly used forecast aggregation rules. Further, the essays contribute to prescriptive research by proposing robust forecast ensemble methods to aggregate judgmental forecasts of time series that are characterized by structural shifts. One of the prescribed forecast ensemble methods was also tested on real time macroeconomic forecast variables and the results (in terms of forecast performance) were found to be robust.

Collectively the essays in the dissertation have aimed to improve our understanding on mechanisms and decision rules used to distil collective wisdom in different

application contexts. Together, the studies carried out in this dissertation have demonstrated the impact of systematic biases and cognitive limitations of the decision maker in different contexts on the process of extracting collective wisdom. Further the dissertation prescribes guidelines and methods to minimize the impact of such biases (and/or cognitive limitations) while distilling collective wisdom. Overall, the studies in this dissertation have aimed to contribute to both descriptive and prescriptive research in aggregation of individual judgments to improve decision quality.

Conclusión

Esta tesis pretende responder a la creciente necesidad de estudiar las reglas de decisión y los mecanismos de agregación utilizados en la extracción de la sabiduría colectiva en las aplicaciones de la sabiduría de los grupos. El proceso de condensación de la sabiduría colectiva se ve a menudo dificultado por las limitaciones cognitivas y los sesgos sistemáticos del decisor impuestos por las reglas de decisión y los mecanismos de agregación. El objetivo principal de mi tesis era comprender el impacto de los sesgos individuales y las limitaciones cognitivas en los juicios colectivos y proponer mecanismos prescriptivos para agregar los juicios individuales. Me he centrado en dos áreas destacadas de las aplicaciones de la sabiduría de los grupos, tanto desde el punto de vista de la investigación como de la práctica: la gestión de la innovación abierta y el pronóstico de juicio.

En el primer capítulo, me he centrado en la evaluación y selección de ideas en el contexto de la innovación abierta. El estudio compara dos reglas de decisión para la evaluación de ideas comúnmente empleadas (puntuación y clasificación) y ofrece una visión de cómo la eficacia de estas reglas de decisión se ve influida por el número de ideas que hay que evaluar y la cantidad de información disponible. A nivel agregado, la regla de decisión de la puntuación puede permitir mayores niveles de precisión con menos participantes, ya que el proceso de puntuación es más eficaz para evaluar ideas cuando hay varias ideas que evaluar. Este estudio contribuye a la investigación descriptiva al destacar las posibles limitaciones cognitivas asociadas al uso de diferentes reglas de decisión utilizadas en la

evaluación y selección de ideas, un espacio relativamente inexplorado en el desarrollo de nuevos productos y la innovación abierta. Además, el ensayo contribuye a la prescripción al aportar ideas sobre los elementos de diseño de los entornos de innovación abierta (por ejemplo, número de ideas que se presentarán, estimaciones del tamaño de la multitud, etc.).

Los dos últimos capítulos se centran en un ámbito relativamente inexplorado, aunque importante, de la previsión de series temporales caracterizadas por rupturas estructurales. Estos ensayos tienden un puente entre las dos corrientes de pronóstico de juicio de series temporales y la sabiduría de los pronósticos agregados. Estos estudios analizan el rendimiento de diferentes métodos de agregación de pronósticos y descubren que los métodos de agregación comúnmente utilizados, como las medias simples y el recorte simétrico, pueden ser menos eficaces en entornos de previsión caracterizados por rupturas estructurales. Además, los capítulos dos y tres proponen una nueva regla de agregación de recorte asimétrico para superar las limitaciones de los métodos de agregación tradicionales. Estos capítulos contribuyen colectivamente a la investigación descriptiva al demostrar la presencia de sesgos sistemáticos en los pronósticos, causados por discontinuidades de las series temporales como las rupturas estructurales en tareas de pronóstico de juicio, y al destacar cómo la presencia de dichos sesgos puede afectar a las reglas de agregación de previsiones utilizadas habitualmente. Además, los ensayos contribuyen a la investigación prescriptiva proponiendo reglas robustas de conjunto de pronósticos para agregar pronósticos

de juicio de series temporales que se caracterizan por cambios estructurales. Uno de los métodos de conjunto de pronósticos prescritos también se probó con variables de previsión macroeconómica en tiempo real y los resultados en términos de rendimiento de pronósticos resultaron ser robustos.

En conjunto, los ensayos de la tesis han pretendido mejorar nuestra comprensión de los mecanismos y las reglas de decisión utilizados para condensar la sabiduría colectiva en diferentes contextos de aplicación. En conjunto, los estudios realizados en la presente tesis han demostrado el impacto de los sesgos sistemáticos y las limitaciones cognitivas del responsable de la toma de decisiones en diferentes contextos en el proceso de obtención de la sabiduría colectiva. Además, la tesis prescribe directrices y métodos para minimizar el impacto de dichos sesgos y limitaciones al condensar la sabiduría colectiva en diferentes contextos de aplicación. En general, los estudios de esta tesis han pretendido contribuir a la investigación tanto descriptiva como prescriptiva en la agregación de juicios individuales para mejorar la calidad de la decisión.

Bibliography

Aiolfi, Marco, and Allan Timmermann. "Persistence in forecasting performance and conditional combination strategies." *Journal of Econometrics* 135, no. 1-2 (2006): 31-53.

Alwin, D. F., J. A. Krosnick. 1985. The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly*, 49(4): 535-552. doi: 10.1086/268949.

Armstrong, J Scott, ed. *Principles of forecasting: a handbook for researchers and practitioners*. Vol. 30. Springer Science & Business Media, 2001.

Armstrong, J Scott., and Fred Collopy. "Causal forces: Structuring knowledge for time-series extrapolation." *Journal of Forecasting* 12, no. 2 (1993): 103-115.

Armstrong, J. Scott. "Combining forecasts: The end of the beginning or the beginning of the end?." *International Journal of Forecasting* 5, no. 4 (1989): 585-588.

Atanasov, Pavel, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. "Distilling the wisdom of crowds: Prediction markets vs. prediction polls." *Management science* 63, no. 3 (2017): 691-706.

Atiya, Amir F. "Why does forecast combination work so well?." *International Journal of Forecasting* 36, no. 1 (2020): 197-200.

Aue, Alexander, and Lajos Horváth. "Structural breaks in time series." *Journal of Time Series Analysis* 34, no. 1 (2013): 1-16.

Baddeley, A. 1992. Working memory. *Science*, 255(5044): 556-559. doi: 10.1126/science.1736359.

Bai, Jushan, and Pierre Perron. "Computation and analysis of multiple structural change models." *Journal of applied econometrics* 18, no. 1 (2003): 1-22.

Bai, Jushan, and Pierre Perron. "Estimating and testing linear models with multiple structural changes." *Econometrica* (1998): 47-78.

Batchelor, Roy, and Pami Dua. "Conservatism and consensus-seeking among economic forecasters." Journal of Forecasting 11, no. 2 (1992): 169-181.

Bates, John M., and Clive WJ Granger. "The combination of forecasts." *Journal of the Operational Research Society* 20, no. 4 (1969): 451-468.

Bayus, Barry L. "Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community." *Management science* 59, no. 1 (2013): 226-244.

Beach, L.R., 1993. Broadening the definition of decision making: The role of prechoice screening of options. *Psychological Science*, 4(4), pp.215-220.

Becker, Otwin, Johannes Leitner, and Ulrike Leopold-Wildburger. "Expectation formation and regime switches." *Experimental Economics* 12, no. 3 (2009): 350-364.

Berinsky, A. J., G. A. Huber, G. S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3): 351-368.

Bettman, R., E. Johnson, J. Payne. 1991. Consumer decision making. *Handbook of Consumer Behavior,* 44(2): 50-84.

Bettman, R., P. Kakkar. 1977. Effects of information presentation format on consumer information acquisition strategies. *Journal of Consumer Research*, 3(4): 233-240. doi: 10.1086/208672.

Bjelland, O. M., R. C. Wood. 2008. An inside view of IBM's 'Innovation Jam.' *MIT Sloan Management Review,* 50(1): 32.

Bockstedt, J., C. Druehl, and A. Mishra. 2016. Heterogeneous submission behavior and its implications for success in innovation contests with public submissions. *Production and Operations Management* 25(7): 1157-1176.

Bockstedt, Jesse, Cheryl Druehl, and Anant Mishra. "Problem-solving effort and success in innovation contests: The role of national wealth and national culture." *Journal of Operations Management* 36 (2015): 187-200.

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

Bolger, Fergus, and Dilek Önkal-Atay. "The effects of feedback on judgmental interval predictions." *international Journal of forecasting* 20, no. 1 (2004): 29-39.

Boone, L. 2020. OECD Interim Economic Outlook. *Organization for Economic Co-operation and Development.* https://www.oecd.org/economic-outlook/. Accessed on 04-27-2020.

Boone, Laurence, David Haugh, Nigel Pain, and Veronique Salins. "Tackling the fallout from COVID-19." *Economics in the Time of COVID-19* 37 (2020).

Boudreau, Kevin J., Nicola Lacetera, and Karim R. Lakhani. "Incentives and problem uncertainty in innovation contests: An empirical analysis." *Management science* 57, no. 5 (2011): 843-863.

Brown, Scott, and Mark Steyvers. "The dynamics of experimentally induced criterion shifts." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, no. 4 (2005): 587.

Budescu, David V., and Eva Chen. "Identifying expertise to extract the wisdom of crowds." *Management Science* 61, no. 2 (2015): 267-280.

Buskirk, E. V. (2010). Google Struggles to Give Away $10 Million. Wired Magazine. June 28, 2010, http://www.wired.com/business/2010/06/google-struggles-to-give-away-10-million/all/.

Campbell, Sean D., and Steven A. Sharpe. "Anchoring bias in consensus forecasts and its effect on market prices." Journal of Financial and Quantitative Analysis 44, no. 2 (2009): 369-390.

Chang, Tzu-Pu, and Ray Yeutien Chou. "Anchoring effect on macroeconomic forecasts: A heterogeneity approach." Proceedings of MAC-MME 2016 (2016): 180.

Chesbrough, H. W. 2006. *Open Innovation: the New Imperative for Creating and Profiting from Technology*. Harvard Business School Press, Boston, MA.

Clemen, Robert T. "Combining forecasts: A review and annotated bibliography." *International journal of forecasting* 5, no. 4 (1989): 559-583.

Clemen, R. T., Winkler, R. L. “Limits for the Precision and Value of Information from Dependent Sources” Operations Research. 33 (1985). 427-442

Clements, Michael P., and David F. Hendry. "Forecasting economic processes." *International Journal of Forecasting* 14, no. 1 (1998): 111-131.

Cruz-Cunha, M. M. (Ed.). 2012. *Handbook of Research on Serious Games as Educational, Business and Research Tools*. IGI Global, Hershey, PA.

Dahan, E., A. Soukhoroukova, M. Spann. 2010. New product development 2.0: preference markets—how scalable securities markets identify winning product concepts and attributes. *Journal of Product Innovation Management*, 27(7): 937-954. doi: 10.1111/j.1540-5885.2010.00763.x.

Dahan, E., H. Mendelson. 2001. An extreme-value model of concept testing. *Management Science*, 47(1): 102-116. doi: 10.1287/mnsc.47.1.102.10666.

Dahan, E., J. R. Hauser. 2002. The virtual customer. *Journal of Product Innovation Management*, 19(5): 332-353. doi: 10.1111/1540-5885.1950332.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell. "When is a crowd wise?." *Decision* 1, no. 2 (2014): 79.

Dhar, Ravi, and S. Nowlis. 1999. The effect of time pressure on consumer choice deferral. *Journal of Consumer Research* 25(4): 369-384

Diebold, Francis X., and Minchul Shin. "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives." *International Journal of Forecasting* 35, no. 4 (2019): 1679-1691.

Durbach, Ian N., and Gilberto Montibeller. "Behavioural Analytics: Exploring judgments and choices in large data sets." *Journal of the Operational Research Society* 70, no. 2 (2019): 255-268.

Edmundson, R. H. "Decomposition; a strategy for judgemental forecasting." *Journal of Forecasting* 9, no. 4 (1990): 305-314.

Edmundson, Bob, Michael Lawrence, and Marcus O'Connor. "The use of non-time series information in sales forecasting: A case study." *Journal of Forecasting* 7, no. 3 (1988): 201-211.

Einhorn, Hillel J., and Robin M. Hogarth. 1981. Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology* 32(1): 53-88.

Erat, S. 2017. Making the Best Even Better: The Role of Idea Pool Structure. *Production and Operations Management*, 26(10), 1946-1959

Erat, Sanjiv, and Vish Krishnan. "Managing delegated search over design spaces." *Management Science* 58, no. 3 (2012): 606-623.

Faure, C. 2004. Beyond brainstorming: Effects of different group procedures on selection of ideas and satisfaction with the process. *Journal of Creative Behavior*, 38(1): 13–34.

Fildes, Robert, and Paul Goodwin. "Against your better judgment? How organizations can improve their use of management judgment in forecasting." *Interfaces* 37, no. 6 (2007): 570-576.

Fujiwara, Ippei, Hibiki Ichiue, Yoshiyuki Nakazono, and Yosuke Shigemi. "Financial markets forecasts revisited: Are they rational, stubborn or jumpy?." Economics Letters 118, no. 3 (2013): 526-530.

Furlong, Michael J., and Bruce E. Wampold. "Visual analysis of single-subject studies by school psychologists." *Psychology in the Schools* 18, no. 1 (1981): 80-86.

Gaba, Anil, Ilia Tsetlin, and Robert L. Winkler. "Combining interval forecasts." *Decision Analysis* 14, no. 1 (2017): 1-20.

Galton, Francis. "Vox populi (the wisdom of crowds)." *Nature* 75, no. 7 (1907): 450-451.

Gao, P., Kaas, H. W., Mohr, D., and Wee, D. (2016). *Disruptive trends that will transform the auto industry*. McKinsey Company.

Giacomini, Raffaella, and Barbara Rossi. "Forecast comparisons in unstable environments." *Journal of Applied Econometrics* 25, no. 4 (2010): 595-620.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). Simple Heuristics That Make Us Smart. New York: Oxford University Press.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review, 103, 595–596.

Girotra, K., C. Terwiesch, K. T. Ulrich. 2010. Idea generation and the quality of the best idea. *Management Science*, 56(4): 591-605. doi: 10.1287/mnsc.1090.1144.

Gobet, Fernand, and Herbert A. Simon. "Templates in chess memory: A mechanism for recalling several boards." *Cognitive psychology* 31, no. 1 (1996): 1-40.

Goodwin, P. (2009). New evidence on the value of combining forecasts. *Foresight: The International Journal of Applied Forecasting*, (12), 33-35.

Graefe, Andreas, J. Scott Armstrong, Randall J. Jones Jr, and Alfred G. Cuzán. "Combining forecasts: An application to elections." *International Journal of Forecasting* 30, no. 1 (2014): 43-54.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, 63(4), 1110-1130.

Gwet, K. L. "Benchmarking inter-rater reliability coefficients." *Handbook of inter-rater reliability. 3rd edn. Gaithersburg, MD* (2012): 121-128.

Gwet, K. 2002. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing, Gaithersburg, MD.

Hall, S. G. and Mitchell, J. (2007) Combining density forecasts, International Journal of Forecasting, 23, 1–13.

Harvey, Nigel, Matthew Twyman, and Maarten Speekenbrink. "Asymmetric detection of changes in volatility: Implications for risk perception." In Proceedings of the 39th Annual Meeting of the Cognitive Science Society, pp. 2162-2167. Cognitive Science Society, 2017.

Harvey, Nigel. "Judgmental forecasting of univariate time series." *Journal of Behavioral Decision Making* 1, no. 2 (1988): 95-110.

Harzing, A.W., J. Baldueza, W. Barner-Rasmussen, C. Barzantny, A. Canabal, A. Davila, A. Espejo, R. Ferreira, A. Giroud, K. Koester, Y.K Liang. 2009. Rating versus ranking: what is the best way to reduce response and language bias in cross-national research? *International Business Review*, 18(4): 417-432.

Hastie, R. and Dawes, R.M., 2001. *Rational Decision in an Uncertainty World: The Psychology of Judgment and Decision Making*. Sage Publisher, London

Hendry, David F., and Michael P. Clements. "Pooling of forecasts." *The Econometrics Journal* 7, no. 1 (2004): 1-31.

Hibon, Michele, and Theodoros Evgeniou. "To combine or not to combine: selecting among forecasts and their combinations." *International Journal of Forecasting* 21, no. 1 (2005): 15-24.

Hoffrage, U. (2019). Modeling experts with fast-and-frugal heuristics. In The Oxford Handbook of Expertise (pp. 148–172). Oxford University Press.

Hong, Lu, and Scott E. Page. "Some microfoundations of collective wisdom." *Collective wisdom* (2012): 56-71.

Ichiue, Hibiki, and Tomonori Yuyama. "Using survey data to correct the bias in policy expectations extracted from fed funds futures." *Journal of Money, Credit and Banking* 41, no. 8 (2009): 1631-1647.

Jore, A. S., Mitchell, J. and Vahey, S. P. (2010) Combining forecast densities from VARs with uncertain instabilities, Journal of Applied Econometrics, 25, 621–34

Jose, Victor Richmond R., Yael Grushka-Cockayne, and Kenneth C. Lichtendahl Jr. "Trimmed opinion pools and the crowd's calibration problem." *Management Science* 60, no. 2 (2014): 463-475.

Jose, Victor Richmond R., and Robert L. Winkler. "Simple robust averages of forecasts: Some empirical results." *International Journal of Forecasting* 24, no. 1 (2008): 163-169.

Kavadias S. and Chao R.. 2007. Resource Allocation and New Product Development Portfolio Management, chapter 7, 135-163, in Loch C. H., and Kavadias S. (eds.) *Handbook of Research in New Product Development Management*, Elsevier/Butterworth, Oxford UK.

Kerr, Norbert L., and R. Scott Tindale. "Group-based forecasting?: A social psychological analysis." *International journal of forecasting* 27, no. 1 (2011): 14-40.

Kim, Oliver, Steve C. Lim, and Kenneth W. Shaw. "The inefficiency of the mean analyst forecast as a summary forecast of earnings." *Journal of Accounting Research* 39, no. 2 (2001): 329-335.

King, A., K. R. Lakhani. 2013. Using open innovation to identify the best ideas. *MIT Sloan Management Review*, 55(1): 41.

Klein, M., A. Garcia. 2015. High-speed idea filtering with the bag of lemons. *Decision Support System* 78(1): 39–50

Klein, Mark, and Gregorio Convertino. "An embarrassment of riches." *Communications of the ACM* 57, no. 11 (2014): 40-42.

Kornish, L. and J. Hutchison-Krupat. 2017. Research on idea generation and selection: Implications for management of technology *Production and Operations Management* 26(4): 633-651.

Kornish, L., and K. Ulrich. 2014. The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *Journal of Marketing Research* 51(1): 14-26.

Kremer, Mirko, Brent Moritz, and Enno Siemsen. "Demand forecasting behavior: System neglect and change detection." *Management Science* 57, no. 10 (2011): 1827-1843.

Langville, Amy N., and Carl D. Meyer. Who's# 1?: the science of rating and ranking. Princeton University Press, 2012.

Larrick, Richard P., and Jack B. Soll. "Intuitions about combining opinions: Misappreciation of the averaging principle." *Management science* 52, no. 1 (2006): 111-127.

Lawrence, Michael, Paul Goodwin, Marcus O'Connor, and Dilek Önkal. "Judgmental forecasting: A review of progress over the last 25 years." *International Journal of Forecasting*22, no. 3 (2006): 493-518.

Lawrence, Michael, and Spyros Makridakis. "Factors affecting judgmental forecasts and confidence intervals." *Organizational Behavior and Human Decision Processes* 43, no. 2 (1989): 172-187.

Lee, Michael D., Irina Danileiko, and Julie Vi. "Testing the ability of the surprisingly popular method to predict NFL games." *Judgment and Decision Making* 13, no. 4 (2018): 322.

Lee, Yun Shin, and Enno Siemsen. "Task decomposition and newsvendor decision making." *Management Science* 63, no. 10 (2017): 3226-3245.

Lee, Yun Shin, Yong Won Seo, and Enno Siemsen. 2018. Running Behavioral Operations Experiments Using Amazon's Mechanical Turk. Production and Operations Management (2018). 27 (5), 973-989.

Lichtendahl Jr, Kenneth C., Yael Grushka-Cockayne, and Robert L. Winkler. "Is it better to average probabilities or quantiles?." *Management Science* 59, no. 7 (2013): 1594-1611.

Luce, M., J. Bettman, J. Payne. 1997. Choice processing in emotionally difficult decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23(2): 384.

Makridakis, Spyros, and Robert L. Winkler. "Averages of forecasts: Some empirical results." *Management Science* 29, no. 9 (1983): 987-996.

Malone, T., R. Laubacher, C. Dellarocas. 2010. The collective intelligence genome. *MIT Sloan Management Review*, 51(3): 21-31.

Mannes, Albert E., Jack B. Soll, and Richard P. Larrick. "The wisdom of select crowds." *Journal of personality and social psychology* 107, no. 2 (2014): 276.

Massey, Cade, and George Wu. "Detecting regime shifts: The causes of under-and overreaction." *Management Science* 51, no. 6 (2005): 932-947.

Matyas, Thomas A., and Kenneth M. Greenwood. "Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects." *Journal of Applied Behavior Analysis* 23, no. 3 (1990): 341-351.

Maule, A., G. Hockey, and L. Bdzola. 2000. Effects of time-pressure on decision-making under uncertainty: changes in affective state and information processing strategy. *Acta Psychologica* 104(3): 283-301.

Medin, D., B. Ross, A. Markman. 2004. *Cognitive Psychology*, 4th edn. Wiley, Hoboken, NJ.

Miller, G. A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2): 81–97. doi: 10.1037/h0043158.

Montibeller, Gilberto, and Detlof von Winterfeldt. "Individual and group biases in value and uncertainty judgments." In *Elicitation*, pp. 377-392. Springer, Cham, 2018.

Moore, M. 1975. Rating versus ranking in the Rokeach value survey: an Israeli comparison. *European Journal of Social Psychology*, 5(3): 405-408. doi: 10.1002/ejsp.2420050313.

Moritz, Brent, Enno Siemsen, and Mirko Kremer. "Judgmental forecasting: Cognitive reflection and decision speed." *Production and Operations Management* 23, no. 7 (2014): 1146-1160.

Nakazono, Yoshiyuki. "Heterogeneity and anchoring in financial markets." Applied financial economics 22, no. 21 (2012): 1821-1826.

O'Connor, Marcus, William Remus, and Kenneth Griggs. "The asymmetry of judgmental confidence intervals in time series forecasting." *International Journal of Forecasting* 17, no. 4 (2001): 623-633.

O'Connor, Marcus, William Remus, and Ken Griggs. "Going up–going down: How good are people at forecasting trends and changes in trends?." *Journal of Forecasting* 16, no. 3 (1997): 165-176.

O'Connor, Marcus, William Remus, and Ken Griggs. "Judgmental forecasting in times of change." *International Journal of Forecasting* 9, no. 2 (1993): 163-172.

O'Connor, Marcus, and Michael Lawrence. "An examination of the accuracy of judgmental confidence intervals in time series forecasting." *Journal of Forecasting* 8, no. 2 (1989): 141-155.

Ozer, M. 2009. The roles of product lead-users and product experts in new product evaluation. *Research Policy*, 38(8): 1340-1349. doi: 10.1016/j.respol.2009.07.001.

Palley, A. B., & Soll, J. B. (2019). Extracting the Wisdom of Crowds When Information is Shared. *Management Science*, 65(5), 2291-2309.

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

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

Paolacci, G., J. Chandler. 2014. Inside the Turk: understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3): 184-188. doi: 10.1177/0963721414531598.

Paolacci, G., J. Chandler. 2014. Inside the Turk: understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3): 184-188. doi: 10.1177/0963721414531598.

Parks, C.D. and Cowlin, R., 1995. Group discussion as affected by number of alternatives and by a time limit. *Organizational Behavior and Human Decision Processes*, 62(3): 267-275.

Paye, Bradley S., and Allan Timmermann. "Instability of return prediction models." *Journal of Empirical Finance* 13, no. 3 (2006): 274-315.

Pesaran, M. Hashem, Andreas Pick, and Mikhail Pranovich. "Optimal forecasts in the presence of structural breaks." *Journal of Econometrics* 177, no. 2 (2013): 134-152.

Pesaran, M. Hashem, Davide Pettenuzzo, and Allan Timmermann. "Forecasting time series subject to multiple structural breaks." *The Review of Economic Studies* 73, no. 4 (2006): 1057-1084.

Pesaran, M. Hashem, and Allan Timmermann. "How costly is it to ignore breaks when forecasting the direction of a time series?." *International Journal of Forecasting* 20, no. 3 (2004): 411-425.

Poetz, M. K., M. Schreier. 2012. The value of crowdsourcing: can users really compete with professionals in generating new product ideas?. *Journal of Product Innovation Management*, 29(2): 245-256. doi: 10.1111/j.1540-5885.2011.00893.x.

Prelec, Dražen, H. Sebastian Seung, and John McCoy. "A solution to the single-question crowd wisdom problem." *Nature* 541, no. 7638 (2017): 532.

Prelec, D. "A Bayesian Truth Serum for Subjective Data." Science 306 (2004). 462-466.

Putman, V.L. and Paulus, P.B., 2009. Brainstorming, brainstorming rules and decision making. *The Journal of Creative Behavior*, 43(1): 29-40.

Rand, D. G. 2012. The promise of mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299: 172-179. doi: 10.1016/j.jtbi.2011.03.004.

Rand, D. G. 2012. The promise of mechanical Turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299: 172-179. doi: 10.1016/j.jtbi.2011.03.004.

Rietzschel, E., Nijstad, B. and Stroebe, W., 2010. The selection of creative ideas after individual idea generation: Choosing between creativity and impact. *British Journal of Psychology*, 101(1), 47-68.

Rossi, Barbara, and Tatevik Sekhposyan. "Forecast rationality tests in the presence of instabilities, with applications to Federal Reserve and survey forecasts." *Journal of Applied Econometrics* 31, no. 3 (2016): 507-532.

Rossi, Barbara. "Advances in forecasting under instability." In Handbook of economic forecasting, vol. 2, pp. 1203-1324. Elsevier, 2013.

Rossi, Barbara. "Are exchange rates really random walks? Some evidence robust to parameter instability." *Macroeconomic dynamics* 10, no. 1 (2006): 20-38.

Russell, P. A., C. D. Gray. 1994. Ranking or rating? Some data and their implications for the measurement of evaluative response. *British Journal of Psychology*, 85(1): 79-92. doi: 10.1111/j.2044-8295.1994.tb02509.x.

Russo, J., G. Krieser, S. Miyashita. 1975. An effective display of unit price information. *Journal of Marketing*, 39: 11-19. doi: 10.2307/1250110.

Rutchick, Abraham M., Bryan J. Ross, Dustin P. Calvillo, and Catherine C. Mesick. "Does the “surprisingly popular” method yield accurate crowdsourced predictions?." *Cognitive research: principles and implications* 5, no. 1 (2020): 1-10.

Seifert, Matthias, and Allègre L. Hadida. "On the relative importance of linear model and human judge (s) in combined forecasting." *Organizational Behavior and Human Decision Processes* 120, no. 1 (2013): 24-36.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? Acta Psychologia, 81, 75–86.

Simon, H. A. 1974. How big is a chunk. *Science*, 183(4124): 482-488. doi: 10.1126/science.183.4124.482.

Simon, Herbert A. "Invariants of human behavior." Annual review of psychology 41, no. 1 (1990): 1-20.

Speekenbrink, Maarten, Matthew Twyman, and Nigel Harvey. "Change detection under autocorrelation." In Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 34, no. 34. 2012.

Stock, James H., and Mark W. Watson. "Combination forecasts of output growth in a seven-country data set." *Journal of forecasting* 23, no. 6 (2004): 405-430.

Stock, James H., and Mark W. Watson. "Evidence on structural instability in macroeconomic time series relations." Journal of Business & Economic Statistics 14, no. 1 (1996): 11-30.

Surowiecki, James. "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business." *Economies, Societies and Nations* 296 (2004).

Terwiesch, C., K. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*. Harvard Business School Press, Boston, MA.

Terwiesch, C., Y. Xu. 2008. Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9): 1529-1543. doi: 10.1287/mnsc.1080.0884.

Tetlock, Philip E., and Dan Gardner. *Superforecasting: The art and science of prediction*. Random House, 2016.

Thomson, Mary E., Andrew C. Pollock, Dilek Önkal, and M. Sinan Gönül. "Combining forecasts: Performance and coherence." *International Journal of Forecasting* 35, no. 2 (2019): 474-484.

Tian, Jing, and Heather M. Anderson. "Forecast combinations under structural break uncertainty." *International Journal of Forecasting* 30, no. 1 (2014): 161-175.

Todd, P. M. & Gigerenzer, G. (2000). Précis of Simple Heuristics That Make Us Smart. Behavioral and Brain Sciences, 23, 727–780.

Toubia, O., L. Florès. 2007. Adaptive idea screening using consumers. *Marketing Science*, 26(3): 342-360. doi: 10.1287/mksc.1070.0273.

Tversky, Amos, and Daniel Kahneman. "Judgment under uncertainty: Heuristics and biases." science 185, no. 4157 (1974): 1124-1131.

Ulrich, K. T., S. Eppinger. 2015. *Product Design and Development*. 6th Edition. McGraw-Hill Higher Education, New York, NY.

Von Hippel, E. 1986. Lead users: a source of novel product concepts. *Management Science*, 32(7): 791-805. doi: 10.1287/mnsc.32.7.791.

Von Hippel, E. A. 2005. *Democratizing Innovation*. MIT Press, Cambridge, MA.

Wallis, Kenneth F. "Combining forecasts–forty years later." *Applied Financial Economics* 21, no. 1-2 (2011): 33-41.

Webby, Richard, Marcus O'Connor, and Bob Edmundson. "Forecasting support systems for the incorporation of event information: An empirical investigation." *International Journal of Forecasting* 21, no. 3 (2005): 411-423.

Wilson, T. D., J. W. Schooler. 1991. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60 (2), 181-192

Wooten J. and K. Ulrich, 2017. Idea Generation and the Role of Feedback: Evidence from Field Experiments with Innovation Tournaments. *Production & Operations Management*. 26(1), 80-99.

Yaniv, Ilan. "The benefit of additional opinions." *Current directions in psychological science* 13, no. 2 (2004): 75-78.

Yaniv, Ilan. "Weighting and trimming: Heuristics for aggregating judgments under uncertainty." *Organizational behavior and human decision processes* 69, no. 3 (1997): 237-249.

Appendix A-1

Results of the Weak Order Ranking Experiment

In this experiment, 103 newly recruited participants were provided with brief (short) pieces of information, and 95 newly recruited participants were provided with detailed (long) pieces of information. The IP address of each participant in the new experiment was checked to ensure that they differ from those in the previous strong-order ranking experiment. Resembling the previous experiment, the eight ideas were ranked under two treatments, i.e., the detailed (long) and brief (short) information treatments. Next, the percentage of matches was calculated for each participant with the procedure described in Section 1.3.2. Finally, the percentage of matches in the weak order ranking process with the results in the scoring process were compared.

		Information		Average
		Short	Long	
Process	Ranking (weak order)	26%	26%	26%
	Scoring[18]	39%	37%	38%
Average		32%	32%	32%

Table A1-0-1: Percentage of "matches"-top two ideas (weak order ranking)

[18] The data of the weak order ranking experiment with long and short information were collected after the first experiment had been conducted. Thus, the comparison with the scoring group can only be done as a quasi-experimental design (i.e., there was no random assignment between scoring and weak order ranking). However, a series of comparisons showed that there were no significant differences on the demographic characteristics or on the domain awareness between the participants in the first and the second data collection. Therefore, any differences in the accuracy or duration cannot be attributed to differences in these aspects of participants.

ANOVA analysis was conducted, and the results are reported in Table 10. This analysis confirms that the main effect of *process* remains significant ($p<0.01$). The percentage of matches under the scoring process is on average 11.8% higher than that in the weak order ranking process ($p<0.01$). Specifically, the percentage of matches in the scoring process is 10.65% (12.93%) higher than that in the ranking process ($p<0.01$), when the detailed (brief) information is provided. However, neither *information* ($p = 0.77$) nor the interaction term between information and process ($p = 0.68$) has a significant influence on the match probability.

Variable	F-ANOVA	P-value
Process	18.12	<0.01***
Information	0.08	0.772
Process × Information	0.17	0.681

Table A1-0-2: ANOVA analysis (weak order ranking)

Appendix A-2

Robustness check with varying the number of ideas

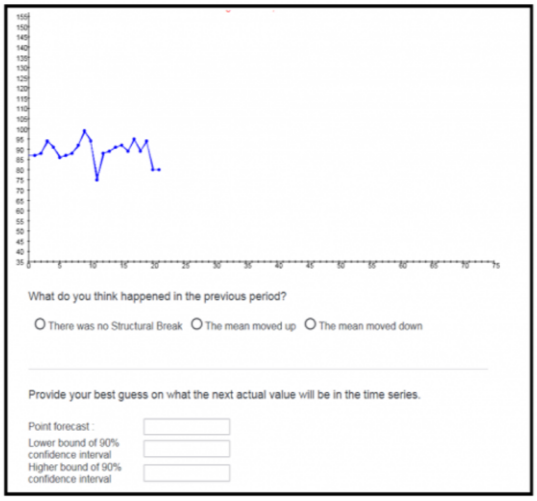
In this section, the aim is to check the robustness of the main result, i.e., the scoring process outperforms the ranking process when eight ideas are evaluated. One implicit assumption in the previous analysis is that there should not be any significant idea-specific characteristics that may systematically influence *all* of the participants' evaluations regarding the order of these ideas. As the dependent variable is the percentage of matches for top two ideas, the potential existence of such unobserved heterogeneities related with the intrinsic orders among ideas may affect the results. To exclude this possibility, a simulation program was used to repeatedly test the model by randomly picking different combinations of ideas from the existing data and check the sign as well as the significance level of the key variable, *process*. For example, if we pick 7 ideas out of 8 existing ideas, there will be eight different combinations. Then, regressions were conducted on each combination of ideas using the existing data associated with the combination and examined how likely the results would hold among these eight combinations. After checking the 7 ideas case, the simulation repeated the aforementioned procedure while examining 6 ideas (28 combinations), 5 ideas (56 combinations) and 4 ideas (70 possible combinations). In this process, all ideas have equal chances of being included or excluded, which enables us to systematically exclude the idea-specific heterogeneities and examine the robustness of the main results. In all the regressions, the dependent variable remains the percentage of matches for the top two ideas.

	Coefficient of the *process*	p<0.05	p<0.1	Not significant
7 ideas (8 possible idea combinations)	Total	7	1	0
	Positive	7	1	0
	Negative	0	0	0
6 ideas (28 possible combinations)	Total	18	3	7
	Positive	18	3	7
	Negative	0	0	0
5 ideas (56 possible combinations)	Total	28	4	24
	Positive	28	4	20
	Negative	0	0	4
4 ideas (70 possible combinations)	Total	25	9	36
	Positive	25	9	28
	Negative	0	0	8

Table A2-0-1: The coefficients of process and associated significance levels.

Appendix B-1

Study material for the experimental study (Chapter 2)



In each game, we are interested in asking you to provide the following forecasting judgments:

(1) Specify what you think happened in the previous period, i.e., no structural break, a structural break where the mean went up, or a structural break where the mean went down.

(2) Estimate a point forecast:
Provide your best guess on what the next actual value will be in the time series.

(3) Estimate the boundaries of a 90% confidence interval: We are interested in the range of values that you believe the next actual value will fall into with a likelihood of 90%. In other words, you should expect 90% of the actual values to lie somewhere between the upper and lower bounds that you specify. There will be a 5% chance that the actual value will be higher than the upper bound and 5% chance that it will be lower than the lower bound.

Example: Imagine your forecast the next point to be at 100, and you are 90% sure that the value will be in the range of 110 and 90. In this case, you will enter the value for the Upper Bound of 110, and for the Lower Bound of 90.

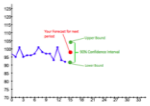
Here is a graphical representation of what a 90% confidence interval would look like. 

Figure B1: Experimental task screenshot on Qualtrics

Appendix B-2

Description of variables used in real-world data analysis (Sec. 2.6)

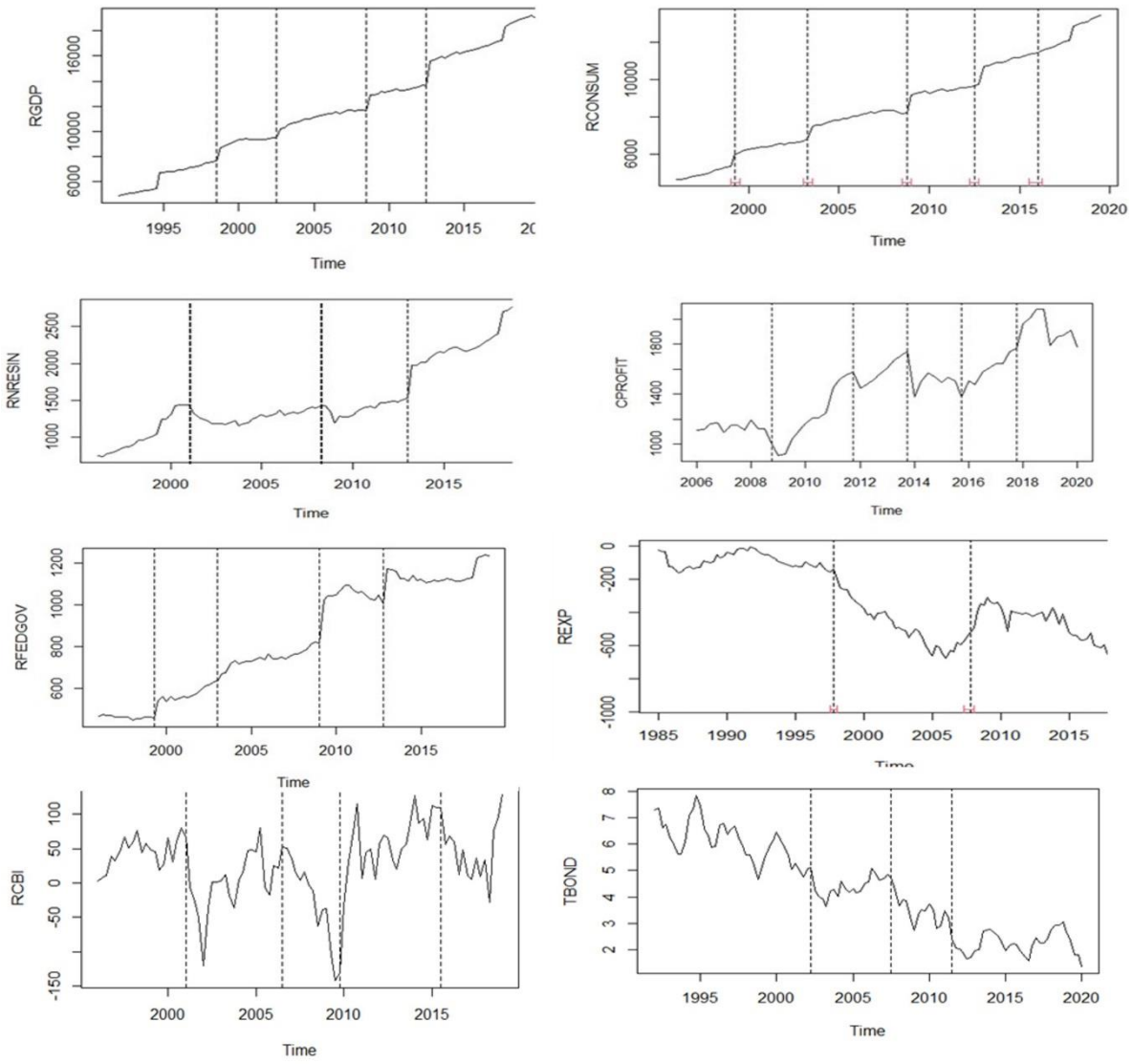
1. Real GDP – We use historical quarterly forecasts (1992-2020) of real GDP. The actual quarterly real GDP data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
2. Real Personal Consumption Expenditures - We use historical quarterly forecasts (1996-2019) of the annual level of chain-weighted real personal consumption expenditures. The actual quarterly real personal consumption expenditures data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
3. Real Federal Government Consumption and Gross Investment – We use historical forecasts (1996-2019) for the quarterly real federal government consumption and gross investment and use the actual RFEDGOV data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
4. Real Non-residential Fixed Investment – We use historical forecasts (1996-2019) from the SPF for the quarterly levels of real non-residential fixed investment (also known as the business fixed investment) and actual quarterly business fixed investment data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
5. Nominal Corporate Profits - We use historical quarterly forecasts (2006-2020) of the annual level of nominal corporate profits after tax. The actual quarterly

nominal corporate profit data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.

6. Real Net Export - We use historical forecasts (1985-2019) for the quarterly level of real net-exports and actual quarterly net exports data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
7. Real Change in Private Inventories – We use historical forecasts (1996-2019) of the quarterly real change in private inventories from the SPF and actual change in private inventories data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
8. T-Bill Rate – We use historical quarterly average forecasts (1990-2015) of the three-month Treasury bill rate. The actual three-month Treasury bill rates for this period is obtained from the daily treasury yield curve rates provided by the US department of Treasury.
9. 10-year Treasury Bond Rate - We use historical quarterly forecasts (1992-2020) of the average 10-year Treasury bond rate. The actual T-Bond rate data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
10. Unemployment – We use historical quarterly forecasts (1968-2019) of average unemployment rate. The actual quarterly average unemployment rate is obtained by averaging the actual monthly unemployment rate data obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.

11. Real State Local Government Consumption and Gross Investment – We use historical forecasts (1981-2014) for the quarterly real state and local government consumption and gross investment. The actual quarterly RLGOV data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
12. Nonfarm Payroll Employment – We use historical forecasts (2004-2019) from the SPF for the quarterly average level of nonfarm payroll employment in thousands of jobs and the actual data for this variable is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
13. Nominal GDP – We use historical forecasts (1992-2019) from the SPF for quarterly nominal GDP in billion dollars and actual quarterly nominal GDP data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
14. Real Residential Fixed Investment – We use historical forecasts (1996-2019) from the SPF for the quarterly level of real residential fixed investment and actual quarterly real residential fixed investment data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.
15. Real personal consumption expenditures – We use historical forecasts (1996-2019) of the quarterly levels of real personal consumption expenditures from the SPF and the actual total quarterly real personal consumption expenditures data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.

16. GDP Price Index – We use historical forecasts (1992-2019) of the quarterly GDP price index from the SPF and the actual quarterly GDP price index data is obtained from the real time data research center of the Federal Reserve Bank of Philadelphia.



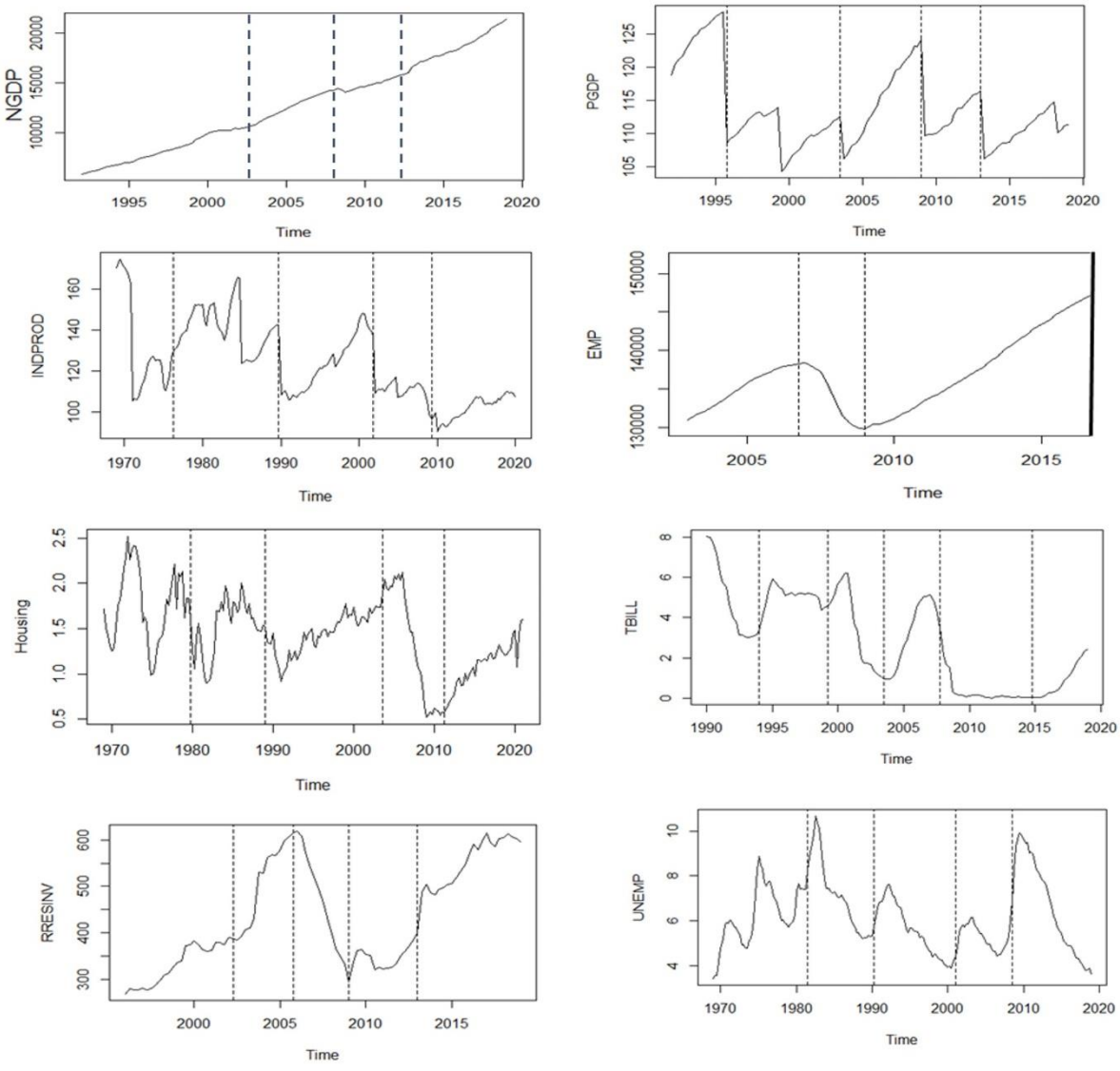


Figure B2: Actual Time Series Plots for the SPF Variables

Appendix C

Break detection accuracy (shift magnitude – 20 & 15 units)

Mean Shift Magnitude - 20 units		Actual Event		
		No Shift	**Upward Shift**	**Downward Shift**
Predicted Event	**No Shift**	28.90%	19.51%	17.30%
	Upward Shift	37.05%	52.91%	30.91%
	Downward Shift	34.05%	27.58%	51.79%
	Total	100.00%	100.00%	100.00%

Mean Shift Magnitude - 15 units		Actual Event		
		No Shift	**Upward Shift**	**Downward Shift**
Predicted Event	**No Shift**	23.46%	16.63%	18.21%
	Upward Shift	40.70%	51.27%	37.59%
	Downward Shift	35.84%	32.10%	44.20%
	Total	100.00%	100.00%	100.00%

Table C-0-1: Break detection accuracy results (shift magnitude 20 and 15 units)

MAPE comparison aggregate level (shift magnitude – 20 & 15 Units)

Mean Shift Magnitude - 20 units

Forecast Aggregation Approach	Pre-Break Segment	Post-Break Segment	Overall
Simple Averaging	0.0590 (0.0014)	0.1143 (0.0024)	0.1075 (0.0021)
Symmetric trimming	0.0511 (0.0012)	0.0952 (0.0020)	0.0904 (0.0017)
Left Trim	0.0939 (0.0020)	0.1541 (0.0051)	0.1468 (0.0044)
Right Trim	0.0797 (0.0024)	0.1344 (0.0024)	0.1275 (0.0023)
Break Judgment Ensemble	0.0580 (0.0013)	0.1010 (0.0020)	0.0963 (0.0017)
Empirical Performance Ensemble	0.0556 (0.0024)	0.0744 (0.0009)	0.0742 (0.0008)

Mean Shift Magnitude - 15 units

Forecast Aggregation Approach	Pre-Break Segment	Post-Break Segment	Overall
Simple Averaging	0.0565 (0.0011)	0.0859 (0.0009)	0.0821 (0.0008)
Symmetric trimming	0.0473 (0.0009)	0.0757 (0.0008)	0.0716 (0.0007)
Left Trim	0.0759 (0.0014)	0.0948 (0.0018)	0.0908 (0.0016)
Right Trim	0.0738 (0.0018)	0.1102 (0.0016)	0.1063 (0.0015)
Break Judgment Ensemble	0.0550 (0.0011)	0.0802 (0.0009)	0.0765 (0.0007)
Empirical Performance Ensemble	0.0610 (0.0028)	0.0640 (0.0006)	0.0633 (0.0005)

Table C-0-2: Mean Values of MAPE for different aggregation approaches (shift magnitude 20 and 15 units)

MAPE comparisons across different series (shift magnitudes 15 & 20 units)

Mean Shift Magnitude - 20 units

Forecast Aggregation Approach	FUUU	FUUD	FUDD	FUDU	FDDD	FDDU	FDUU	FDUD
Simple Averaging	0.1261 (0.0047)	0.0920 (0.0027)	0.0877 (0.0031)	0.0701 (0.0016)	0.1901 (0.0086)	0.1482 (0.0081)	0.0743 (0.0019)	0.0716 (0.0022)
Symmetric trimming	0.0822 (0.0042)	0.0790 (0.0026)	0.0664 (0.0023)	0.0649 (0.0013)	0.1290 (0.0052)	0.1077 (0.0040))	0.0648 (0.0017)	0.0641 (0.0017)
Left Trim	0.0461 (0.0013)	0.0680 (0.0026)	0.1433 (0.0078)	0.0836 (0.0023))	0.3441 (0.0164)	0.2699 (0.0168)	0.1057 (0.0037)	0.1142 (0.0048)
Right Trim	0.2464 (0.0077)	0.1750 (0.0061)	0.1111 (0.0055)	0.1049 (0.0038)	0.0990 (0.0012)	0.0916 (0.0016)	0.1044 (0.0030)	0.1275 (0.0023)
Break Judgment Ensemble	0.0934 (0.0034)	0.0800 (0.0024)	0.0829 (0.0032)	0.0683 (0.0013)	0.1602 (0.0062)	0.1330 (0.0062)	0.0707 (0.0017)	0.0706 (0.0020)
Empirical Performance Ensemble	0.0522 (0.0009)	0.0557 (0.0019)	0.0640 (0.0015)	0.0634 (0.0009)	0.1142 (0.0026)	0.0977 0.0016)	0.0666 (0.0014)	0.0720 (0.0009)

Mean Shift Magnitude - 15 units

Forecast Aggregation Approach	FUUU	FUUD	FUDD	FUDU	FDDD	FDDU	FDUU	FDUD
Simple Averaging	0.0861 (0.0030)	0.0900 (0.0026)	0.0644 (0.0015)	0.0764 (0.0019)	0.0873 (0.0024)	0.0855 (0.0013)	0.0907 (0.0013)	0.0762 (0.0025)
Symmetric trimming	0.0629 (0.0023)	0.0658 (0.0025)	0.0668 (0.0012)	0.0646 (0.0020)	0.0768 (0.0022)	0.0783 (0.0011)	0.0783 (0.0011)	0.0677 (0.0023)
Left Trim	0.0431 (0.0007)	0.0558 (0.0010)	0.0643 (0.0018)	0.0667 (0.0025)	0.1494 (0.0043)	0.1365 (0.0019)	0.1091 (0.0019)	0.0879 (0.0044)
Right Trim	0.1400 (0.0043)	0.1554 (0.0043)	0.0757 (0.0028)	0.0882 (0.0026)	0.0660 (0.0014)	0.0713 (0.0011)	0.0713 (0.0011)	0.0686 (0.0017)
Break Judgment Ensemble	0.0730 (0.0026)	0.0813 (0.0027)	0.0596 (0.0011)	0.0696 (0.0019)	0.0.079 (0.0022)	0.0807 (0.0013)	0.0807 (0.0013)	0.0703 (0.0024)
Empirical Performance Ensemble	0.0487 (0.0011)	0.0569 (0.0012)	0.0592 (0.0009)	0.0604 (0.0019)	0.0656 (0.0018)	0.0716 0.0014)	0.0717 (0.0014)	0.0648 (0.0020)

Table C-0-3:Mean Values of MAPE across series (shift magnitude 20 and 15 units)