



University of Peloponnese

Department of Economics

Boosting organizational decision making through
Structured Super-forecasting

A risk management based approach

Ilias D. Katsagounos

2019



UNIVERSITY OF PELOPONNESE

SCHOOL OF ECONOMY, MANAGEMENT & INFORMATICS

DEPARTMENT OF ECONOMICS



**BOOSTING ORGANIZATIONAL DECISION MAKING THROUGH STRUCTURED
SUPER-FORECASTING**

A RISK MANAGEMENT BASED APPROACH

by

ILIAS KATSAGOUNOS

PhD Dissertation

TRIPOLI, 2019



UNIVERSITY OF PELOPONNESE

SCHOOL OF ECONOMY, MANAGEMENT & INFORMATICS

DEPARTMENT OF ECONOMICS

**BOOSTING ORGANIZATIONAL DECISION MAKING THROUGH STRUCTURED
SUPER-FORECASTING**

A RISK MANAGEMENT BASED APPROACH

by

ILIAS KATSAGOUNOS

PhD Dissertation

Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

TRIPOLI, 2019

Scientific Committee:

Professor Dimitrios Thomakos, Supervisor

Professor Konstantinos Nikolopoulos, Member

Professor Konstantinos Masselos, Member

Approved by

First Reader (Supervisor)	Dr. Dimitrios Thomakos, Professor, Department of Economics, University of Tripoli
Second Reader	Dr. Konstantinos Nikolopoulos, Professor, Business School, Bangor University
Third Reader	Dr. Konstantinos Masselos, Professor, Department of Economics, University of Tripoli
Fourth Reader	Dr. Georgios Fotopoulos, Professor, Department of Economics, University of Tripoli
Fifth Reader	Dr. Constantina Kottaridi, Associate Professor, Department of Economics, University of Piraeus
Sixth Reader	Dr. Thomas Alexopoulos, Assistant Professor, Department of Economics, University of Tripoli
Seventh Reader	Dr. Krina Griva, Assistant Professor, Department of Economics, University of Tripoli

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Ilias D. Katsagounos

2019

Acknowledgements

For some reason it feels like this section is the most difficult one out of the entire thesis. It is so hard to find the correct words to sincerely thank so many people that stepped in and supported me throughout this journey. I believe that the most correct way would be to do it in chronological order:

The reason I entered the PhD world was Dr Dr H.C. Yapalis Aristides. I still cherish our long discussions while he was challenging me to embark on this journey. The triggering point was a book by Michael Dertouzos, titled: *'The Unfinished Revolution: How to Make Technology Work for Us - Instead of the Other Way Around'*. We were both astonished by Dertouzos's humancentric approach in computer systems, that considered it would have been the ideal stepping stone for further research. In this research though, we address the humancentric approach from a totally different aspect, than the one described in Dertouzos's book: Instead of dealing with how a computer system interacts with a person, we deal with how a person interacts with computers (well, sort of). We stop being the raw data feeders of 'thinking algorithms' and instead we feed the algorithms with well thought-of information that comes as the logical outcome of human mind processing. Dear Aristides, I once again sincerely thank you for your parental affection and guidance.

Well, guess who comes 2nd in chronological order; that wouldn't have been anybody else than my beloved wife, Iro. It was to her that I run back and asked if it would have been feasible (given we already had two kids at that time) to proceed with my PhD application. Her answer was breathtaking, although she knew it would have been overwhelmingly difficult for the entire family. I love her and thank her for all her support and patience the past 3,5 years.

Now the ball goes into my principal supervisor's court, Professor Dimitrios Thomakos, or should I say 'Dr.T'? I'm grateful to him in so many ways. He was a supervisor, a teacher, and above all a dear friend! He was always there, whenever I needed him, to motivate, guide me and even get me back on track whenever necessary. A hard-working man with profound knowledge in econometrics and unparalleled skills in R. He dedicated numerous days of his personal time, even during weekends, to teach me R programming and support me in my data analysis. It was so funny whenever I was describing to fellow PhD researchers from other Universities, my academic relationship with Professor Thomakos; it was difficult for them to grasp that something like that could actually be true. Dear Professor, thanks for everything; I couldn't have imagined having a better supervisor than you.

I'm also grateful to Professor Konstantinos Nikolopoulos for his strategic guidance and insights. A renowned expert in the field of

forecasting and with immense research in the field of analogies, he guided me in the structuring of the experimental procedure and the subsequent analysis of the collected data. Moreover, I feel in debt to him for constantly pushing me to achieve higher goals and not settling with the minimum viable solution. Dr.N thanks for everything; I owe you a lot!

A very special gratitude goes out to all the members of my scientific committee: Professors Konstantinos Masselos and Georgios Fotopoulos, Associate Professor Constantina Kottaridi, Assistant Professors Thomas Alexopoulos and Krina Griva for their insightful comments and the time they devoted in this research.

I am also grateful to our university staff and dear friends, Ms Dimitra Psychogiou, for her continuous support from the University Secretariat, and Mr Dimitris Dimitrakopoulos, our librarian and IT guru, because he was there whenever we needed him, for whatever reason!

Finally, a special thanks to my cordial friend and colleague, Lazaros Rizopoulos PhD, for our long discussions and exchanges on our researches and (mostly) beyond!

Thank you all!

Ilias Katsagounos

Dedications

To my wife Iro for her insurmountable patience and support!

To my three little ones that were fed-up seeing their father sitting in an office for 3,5 years...



Contents

Approved by	vii
Declaration	ix
Acknowledgements	xi
Dedications	xv
Contents	xvii
List of Figures	xxiii
List of Tables	xxvii
Nomenclature	xxix
Abstract (GR)	xxxii
Abstract (EN)	xxxiv
Chapter 1 Introduction	1
1.1 <i>Judgemental forecasting background</i>	1
1.2 <i>The Human mind: re-acknowledging its importance</i>	3
1.3 <i>Interpretations and Applications</i>	8
1.4 <i>Research Overview</i>	11
1.5 <i>Research question and thesis</i>	18
Chapter 2 Examining the status quo in forecasting	20
2.1 <i>Judgemental Forecasting</i>	20
2.2 <i>Forecasting by Analogies</i>	23
2.3 <i>Forecasting and Biases</i>	26
2.4 <i>Forecasting and Training</i>	28

2.5 Forecasting Tournaments	31
Chapter 3 Methodology	36
3.1 General	36
3.2 Methodology	37
3.2.1 Procedure	38
3.2.1.1 Initial demographics	39
3.2.1.2 Factorial design (research team formation)	40
3.2.1.3 Training	41
3.2.1.4 Question setting	43
3.2.1.5 Question scoring	44
3.2.1.6 Post Research “Superforecaster” evaluation	44
3.2.2 Ethical, moral and privacy boundaries	45
3.2.2.1 Ethical and moral issues	45
3.2.2.2 Confidentiality, privacy and relevant legal issues	45
3.2.3 Constraints & assumptions taken under consideration prior to the research	46
3.2.3.1 Constraints	46
3.2.3.2 Assumptions	46
3.2.4 Incentives	47
3.2.5 Benefits from research	47
3.2.5.1 Benefits for participating forecasters	47
3.2.5.2 Direct and Indirect Benefits for the Participating Organizations	48
Chapter 4 Managing decisions under constraints	50
4.1 <i>Structured Superforecasting”: Judgmental forecasting, Supercharged</i>	50
4.1.1 General	50
4.1.2 Abstract	51

4.1.3	Introduction	52
4.1.4	Literature Review	54
4.1.5	Hypotheses	54
4.1.6	Project Design	54
4.1.6.1	Project Overview	56
4.1.6.2	Subjects and research design	57
4.1.6.3	Questions, scoring and feedback	58
4.1.6.4	Incentives	64
4.1.6.5	Training Design	65
4.1.7	Results-Analysis and Hypotheses Testing	67
4.1.7.1	Potential existence and early identification of Superforecasters	69
4.1.7.2	The contribution of training in forecasting performance	74
4.1.7.3	Stochastic Dominance Tests	87
4.1.8	General Discussion	89
4.2	<i>Early vs late forecasting: Do forecasting tournaments help us identify a time related performance of forecasters?</i>	92
4.2.1	General	92
4.2.2	Hypothesis	94
4.2.3	Project design	94
4.2.4	Results-Analysis and Hypotheses Testing	94
4.2.5	General Discussion	97
4.3	<i>"PESCO - PM² - ESDC": Could E-Learning Bring Closer Together EU's Success Stories?</i>	98
4.3.1	General	98
4.3.2	Abstract	99
4.3.3	Introduction	100
4.3.3.1	Historical Background	100
4.3.3.2	Entry into the PESCO era	102

4.3.4	Is there a link between PESCO and a structured project management approach?	106
4.3.4.1	The Council's provisions for project governance	107
4.3.4.2	The road towards a better project management approach. The case of PM2.	110
4.3.5	Is there room for the European Security & Defence College (ESDC) between PESCO & PM ² ?	113
4.3.5.1	Win-Win-Win! A multiple-gain approach	115
4.3.6	Training Method	117
4.3.6.1	Portfolio Manager	118
4.3.6.2	Programme Managers	119
4.3.6.3	Project Managers	121
4.3.6.4	Project team members and external cooperating entities	122
4.3.6.5	The common denominator	125
4.3.7	Conclusion	128
4.4	<i>On the M4.0 forecasting competition: can you tell a 4.0 earthquake from a 3.0?</i>	129
4.4.1	General	129
4.4.2	Abstract	131
4.4.3	First cut is the deepest	132
4.4.4	Thinner, Lighter, Faster	132
4.4.5	A new competition	133
4.4.6	Reality matters and more can be done	133
4.4.7	Sins of commission	135
4.4.8	The winner takes it all	135
4.4.9	Omelets and eggs	136
4.4.10	Time is of the Essence	137
4.4.11	The one to beat	137

Chapter 5 Discussion		141
5.1 <i>Summary</i>		141
5.2 <i>The way ahead</i>		146
References		149
Appendix A	Per Team Performance Analysis	A-1
Appendix B	Descriptive Statistics per Question	B-6
Appendix C	R code for data processing	C-13
C.1	Data collection per question	C-13
C.2	Data standardization	C-18
C.3	Stochastic Dominance calculations	C-22
C.3.1	Source code for 'SD thom' function	C-35
C.3.2	Source code for 'SD per question 3SD' function	C-39
C.4	Comparative performance of forecasters	C-40
Appendix D	PESCO Projects Stratification	D-43
D.1	Short description of the above PESCO projects	D-44
Appendix E	Demographics Survey	E-49
Appendix F	Forecasting question example	F-56

List of Figures

Figure 1: Methodology Tree for Forecasting by J. Armstrong & K. Green.....	2
Figure 2: Dilbert and forecasting	10
Figure 3: Research methodology flowchart	38
Figure 4: Experiment's web interface	39
Figure 5: Final experimental structure flowchart	56
Figure 6: Superofrecaster destribution per (%) bin	70
Figure 7: Namber of forecasters per (%) bin for Avg scores.....	75
Figure 8:AVG scores per (%) bin	75
Figure 9:Namber of forecasters per (%) bin for Net scores.....	76
Figure 10: Net scores per (%) bin	76
Figure 11:Namber of forecasters per (%) bin for standardized over the mean average brier scores.....	77
Figure 12: Standardized over the mean average brier scores per (%) bin.....	77
Figure 13: Namber of forecasters per (%) bin for Standardized over the Mean Net Brier Points	78
Figure 14: Standardized over the Mean Net Brier Points per (%) bins.....	78
Figure 15: Improvement of Team 'B' over Team 'A'.....	80
Figure 16: Average Brier Score per team boxplot	81
Figure 17: Net Brier scores per team Boxplots.....	82
Figure 18: Standardized over the mean Average Brier Scores per Teams Boxplots.....	83

Figure 19: Standardized over the mean Net Brier Scores per Teams Boxplots	.84
Figure 20: Average Brier Scores per Question and Teams Boxplots85
Figure 21: Net Brier Scores per Question and Teams Boxplots85
Figure 22: Standardized over the mean Average Brier Scores per Question and Teams Boxplots86
Figure 23: Standardized over the mean Average Net Scores per Question and Teams Boxplots86
Figure 24: ECDF plots for 1 st order Stochastic Dominance Test89
Figure 25: Standardized over the Mean Average Brier Scores per Teams Boxplots96
Figure 26: ECDF plots for 1st order Stochastic Dominance Test97
Figure 27: CARD formation approach105
Figure 28: EDF breakdown105
Figure 29: Project Management levels109
Figure 30: ESDC contribution115
Figure 31: Smartphones in Billions126
Figure 32: Active mobile broadband subscriptions (in millions) * estimation	.127
Figure 33: Standardized (Median IQR) Average Brier Scores per Teams boxplotsA-1
Figure 34: Standardized (MedianIQR)Net Brier Scores per Teams BoxplotsA-2
Figure 35: Standardized (MedianMAD) Average Brier Scores per Teams BoxplotsA-3
Figure 36 : Standardized (MediaMAD)Net Brier Scores per Teams Boxplots	..A-4
Figure 37: PESCo Project StratificationD-43

List of Tables

Table 1: Source (Kahneman, 2013). The examples are presented in order of complexity.....	5
Table 2: Forecasting tournament training programme	43
Table 3: Demographic Characteristics of forecasters	58
Table 4: Average Brier Score Calculation	61
Table 5: Net Brier Points calculation	62
Table 6: Preparatory training modules	65
Table 7: Superforecaster count per standardization type	70
Table 8: Statistical analysis of forecaster's diversification factors	73
Table 9: Descriptive statistics for average Brier scores per team	81
Table 10: Descriptive statistics for Net Brier scores per team	82
Table 11: Descriptive statistics for Standardized over the mean Average Brier Scores per Teams	83
Table 12: Descriptive statistics for Standardized over the mean Net Brier Scores per Teams.....	84
Table 13: 1st, 2nd and 3rd order Stochastic Dominance tests	88
Table 14: Cross tema descriptive statistics	95
Table 15: Proposed Stratification	118
Table 16: SME determining factors according to the EC	143
Table 17: Descriptive statistics for Standardized (Median IQR) Average Brier Scores per Teams	A-1

Table 18: Descriptive Statistics for Standardized (MedianIQR)Net Brier Scores per Teams.....	A-2
Table 19: Descriptive Statistics for Standardized (MedianMAD) Average Brier Scores per Teams	A-3
Table 20: Descriptive Statistics for Standardized (MediaMAD)Net Brier Scores per Teams.....	A-4
Table 21: Descriptive Statistics for 1st Question	B-6
Table 22: Descriptive Statistics for 2nd Question	B-6
Table 23: Descriptive Statistics for 3rd Question	B-7
Table 24: Descriptive Statistics for 4th Question	B-7
Table 25: Descriptive Statistics for 5th Question	B-8
Table 26: Descriptive Statistics for 6th Question	B-8
Table 27: Descriptive Statistics for 7th Question	B-9
Table 28: Descriptive Statistics for 8th Question	B-9
Table 29: Descriptive Statistics for 9th Question	B-10
Table 30: Descriptive Statistics for 10th Question	B-10
Table 31: Descriptive Statistics for 11th Question	B-11
Table 32: Descriptive Statistics for 12th Question	B-11
Table 33: Descriptive Statistics for 13th Question	B-12
Table 34: Descriptive Statistics for 14th Question	B-12

Nomenclature

Acronym	Description
ACE	Aggregative Contingent Estimation
AI	Artificial Intelligence
ARCS	Attention, Relevance, Confidence, Satisfaction (J. Keller's model)
ARIMA	Auto-Regressive Integrated Moving Average
CARD	Coordinated Annual Review on Defence
CMC	computer-mediated communication
CRT	Cognitive Reflection Test
CSDP	Common Security and Defence College
DNI	Director of National Intelligence
ECDF	Empirical Cumulative Distribution Function
EDA	European Defence Agency
EDC	European Defence Community
EDF	European Defence Fund
EEAS	European External Action Service

ESDC	European Security and Defence College
EUMS	European Union Military Staff
F2F	Face to face training
GJP	Good Judgement Program
GUI	Graphical User Interface
H-R	Human Resources
HFC	Hybrid Forecasting Competition
IARPA	International Advanced Research Program Activity
IC	Intelligence Community
IIF	International Institute of Forecasting
ITU	UN International Telecommunications Union
JF	Judgmental Forecasting
KPI	Key Performance Indicator
LMS	Learning Management System
MAPA	Multiple Aggregation Prediction Algorithm
ML	Machine Learning
NBP	Net Brier points

NIP	national implementation plans
PESCO	Permanent Structure Cooperation (projects)
PM ²	Project Management Methodology
RAM	Random Access Memory
SA	Structured Analogies
SD	Standard Deviation
SD	Stochastic Dominance
SF	Superforecasting
SKU	Stock Keeping Unit
SME	Small Medium Enterprise
SSA	Semi-Structured Analogies
TEU	Treaty of the European Union
TFEU	Treaty of the Functioning of the European Union
UNFCCC	United Nations Framework Convention on Climate Change

Abstract (GR)

Η επιτυχία του Good Judgment Project στην αναγνώριση και αξιοποίηση των 'superforecasters' οδηγεί φυσικά στο ερώτημα πώς μπορεί κανείς να εφαρμόσει αυτή την προσέγγιση σε μικρότερη κλίμακα με περιορισμένους πόρους και λιγότερους συμμετέχοντες. Τα μικρά επιχειρησιακά περιβάλλοντα και οι δομές λήψης αποφάσεων τύπου MME αποτελούν πρωταρχικά παραδείγματα όπου μπορεί να χρησιμοποιηθεί μια τροποποιημένη προσέγγιση του superforecasting.

Σε αυτή την έρευνα επικεντρωνόμαστε σε μια 'υβριδική προσέγγιση' της δια κρίσεως πρόβλεψης για ειδικά γεγονότα όπου συνδυάζουμε την εκπαίδευση των μελλοντικών - δυνητικών 'superforecasters' με μια τροποποιημένη έκδοση των δομημένων αναλογιών. Ονομάζουμε την προκύπτουσα προσέγγιση δομημένο superforecasting και καταδεικνύουμε την αποτελεσματικότητά της σε δείγματα συμμετεχόντων από τον ευρύτερο δημόσιο τομέα και την ακαδημαϊκή κοινότητα.

Συγκεκριμένα, μέσω ενός πειραματικού σχεδιασμού που περιλαμβάνει μία εκπαιδευμένη ομάδα και μία ομάδα αναφοράς, εφαρμόζουμε την παραπάνω μεθοδολογία και συγκρίνουμε τις επιδόσεις. Η ανάλυση των αποτελεσμάτων χρησιμοποιεί πέρα από τις τυποποιημένες έννοιες μέτρησης της κριτικής

πρόβλεψης, τη μεθοδολογία της στοχαστικής δεσπόζουσας θέσης (SD) για την αξιολόγηση της απόδοσης των συμμετεχόντων - για πρώτη φορά εξ όσων γνωρίζουμε σε αυτό το σκέλος της βιβλιογραφίας.

Η χρήση της ιδέας SD είναι σημαντική για δύο λόγους: πρώτον, επιτρέπει μια πλήρη / καλύτερη εικόνα της υπεραπόδοσης σε σύγκριση με τα πιο παραδοσιακά στατιστικά στοιχεία και δεύτερον, παρέχει οπτικές απεικονίσεις των διαφορετικών επιδόσεων για κάθε τρόπο διαμόρφωσης του δείγματός μας.

Είναι ιδιαίτερα σημαντική η διαπίστωσή μας ότι οι συμμετέχοντες οι οποίοι έχουν εκπαιδευτεί στις δομημένες αναλογίες ξεπερνούν το δείγμα ελέγχου σχεδόν σε όλες τις τεθείσες ερωτήσεις, ενώ παράλληλα, μέσω της παρούσας πειραματικής διαδικασίας, επιτυγχάνεται η έγκαιρη αναγνώριση (κατόπιν 6 ερωτήσεων) των συστηματικά αποτελεσματικών 'forecasters'.

Τα οφέλη από την παραπάνω προσέγγιση διαφαίνονται ως ιδιαίτερα σημαντικά και ως εκ τούτου κρίνεται σκόπιμη η επέκταση της παρούσας έρευνας και σε διαφορετικά δείγματα.

Λέξεις-κλειδιά: Προβλέψεις, Superforecasting, Δομημένες Αναλογίες, Προβλέψεις Κρίσεων, Λήψη Αποφάσεων, Συγκριτική Αξιολόγηση, E-Learning

Abstract (EN)

The success of the Good Judgment Project in harnessing the power of superforecasting naturally leads to the question as to how one can implement that approach on a smaller scale with more limited resources and fewer participants. Small(er) corporate environments and SME-type decision structures are prime examples where a modified superforecasting approach can be used.

In this research we focus on a hybrid approach of judgmental forecasting on special events where we combine training of superforecasters-to-be via the concept of a modified version of structured analogies, a staple of judgmental forecasting in the literature. We call the resulting approach structured superforecasting and illustrate its efficacy over samples of participants from the wider public sector and the academic community.

In particular, with a proper experimental design that includes a training and a control group, we apply the above methodology and compare performances. Our analysis of the results utilizes, beyond standard measurement concepts of judgmental forecasting, the methodology of stochastic dominance (SD) to evaluate the performance of participants -- for the first time to the best of our knowledge in this strand of the literature.

The use of the SD concept is important for two reasons: first, it allows a complete/better view of outperformance compared to more traditional statistics and, second, it provides compelling visuals for the results across individual questions and across any sample split we wish.

We find that, across most questions employed, analogies trained participants outperform the control group. Moreover, we also succeeded in the early identification of the consistently top performing forecasters. This is an important and practical result which we validate for the first time.

The implications of extending this research in other environments and different samples is obvious: expending effort and resources in training on analogies can super-charge super forecasting and thus tapping into the wisdom of the crowds does not need larger crowds but, importantly, smarter (by training) ones.

Key words: Forecasting, Superforecasting, Structured Analogies, Judgmental Forecasting, Decision Making, Benchmarking, E-Learning

Chapter 1 Introduction

1.1 Judgemental forecasting background

The past three decades we have observed a phenomenal pivoting towards Judgmental Forecasting (JF) which actually demonstrates the importance of “critical thinking” in providing accurate forecasts. The “Superforecasting” concept as communicated to the wider public through the book ‘Superforecasting, the art and science of prediction’ (Philip Tetlock & Gardner, 2015) clearly substantiated the above observation while, at the same time, opened the door for further research on that particular domain.

JF cannot be considered as panacea when it comes to providing accurate forecasts. The forecaster should be in a position to analyse the situation at hand, including the available information and its nature, and decide upon the optimal method (or even a combination of methods) to be used. One amongst the principal researchers in the field of JF is Scott Armstrong¹. His methodological approach in mapping the forecasting field and defining the key principles that govern the selection procedure of the

¹ https://en.wikipedia.org/wiki/J._Scott_Armstrong

appropriate forecasting method have really contributed in the upscaling of JF (Armstrong, 2001). The below diagram brings together some of the key forecasting methods as mapped together by S. Armstrong and K. Green ².

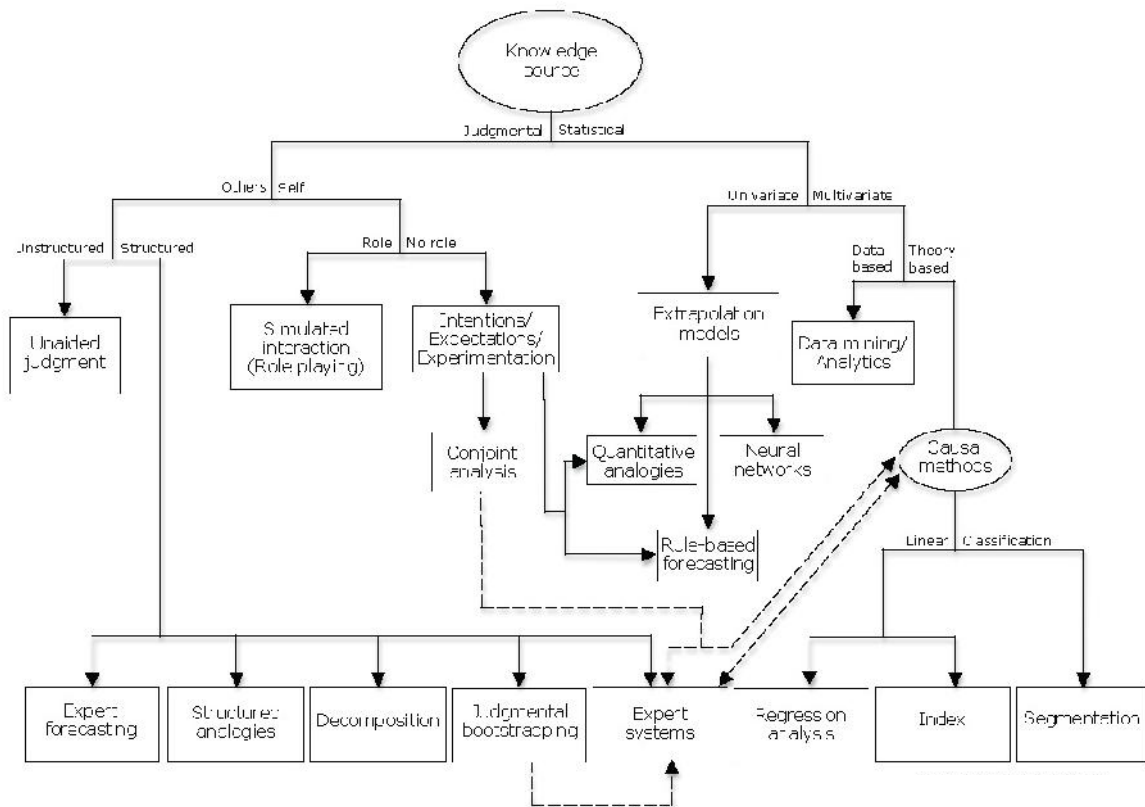


Figure 1: Methodology Tree for Forecasting by J. Armstrong & K. Green

It is evident that judgmental approaches, although they sound rather abstract and unsophisticated, on the contrary they are well patterned, structured (in

² <http://forecastingprinciples.com/index.php/methodology-tree>

most cases) and with an added value wherever pure quantitative information is absent or insufficient.

Excluding the very recent researches where the judgemental approach is systematically measured (in terms of performance) within a forecasting tournament framework, in the past it was primarily being used as an adjustment tool for statistical methods. Lawrence et al (2006), have performed a very thorough mapping of the progress performed within a period of 25 years³ within their review.

1.2 The Human mind: re-acknowledging its importance

The human mind is important for many reasons. The principal reasoning behind its existence is to support our hyper-demanding bodies with their overly complexed functions and demands. Our hand would move, only in the case where our mind would give the respective signal. Equivalently, we would only be able to articulate some arguments under the prerequisite that the mind mobilizes the mouth and vocal muscles, in order to encode the mental reasoning into audible sounds.

But the human mind has evolved dramatically throughout the ages, becoming able to memorize, analyse and synthesize information of much

³ Approximately from 1980 to 2005

more complex nature than just survivability related. Thinking, Fast and Slow (Kahneman, 2013), a best-selling book by Nobel laureate Daniel Kahneman, who was the 2012 winner of the National Academies Communication Award⁴ for best creative work that helped the wider public (i.e. outside the pure scientific community) to understand the complex topics of behavioural science. In this book Kahneman demystifies our mind's functionalities behind reasoning. He actually fights against the unsound ideas of the Social scientists back in the 70s. Their perception of the Social scientists back in the 70s, whose perception was that people are generally rational and emotions are the elements that diverge the human thinking from the rational path. In his seminal paper titled: 'Judgment under uncertainty: heuristics and biases' (Tversky & Kahneman, 1974), he managed to trace down systematic cognition errors that were of higher importance than emotional corruption. He coined the terms of 'System 1' and 'System 2' that remained undisputed since then.

System 1 is a kind of a preconfigured part of our minds that resides on the back part of our brains. It's the place from where fast and intuitive actions-thoughts spring. On the contrary, System 2 is the well-known field of conscious thought that resides on the front part of our brains. Both systems 1 and 2 come

⁴ <http://www.nationalacademies.org>

with advantages and disadvantages. In the first chapter of his book, Kahneman discriminates them by describing the set of actions that each can perform:

System 1:	System 2:
Fast, automatic, frequent, emotional, stereotypic, unconscious.	Slow, effortful, infrequent, logical, calculating, conscious.
determine that an object is at a greater distance than another	brace yourself before the start of a sprint
localize the source of a specific sound	direct your attention towards the clowns at the circus
complete the phrase "war and ..."	direct your attention towards someone at a loud party
display disgust when seeing a gruesome image	look out for the woman with the grey hair
solve $2+2=?$	dig into your memory to recognize a sound
read text on a billboard	sustain a higher than normal walking rate
drive a car on an empty road	determine the appropriateness of a particular behaviour in a social setting
come up with a good chess move (if you're a chess master)	count the number of A's in a certain text
understand simple sentences	give someone your phone number
connect the description 'quiet and structured person with an eye for details' to a specific job	park into a tight parking space

Table 1: Source (Kahneman, 2013). The examples are presented in order of complexity

Despite the fact that System 2 triggers deliberate effortful actions to derive to a logically accepted solution, that does not mean that it is not error – prone. Correspondingly, System 1 can occasionally be extremely efficient and correct, despite the fact of working fast and intuitively. But System 1 and

2 do not work independently. System 2 can only be triggered by an initial answer delivered through System 1. It only then starts questioning it and analysing it. But yet again sometimes, although it is necessary to activate System 2 in order to proceed to a logical action, System 2 seems to step aside and give space to System 1. Everybody dwelling in the world of biases is aware of the ball and bat question:

'A bat and ball together cost \$1.10. The bat costs a dollar more than the ball. How much does the ball cost?'

Sometimes even when we know that the correct answer is 5 cents, our minds keep on telling us that it is 10 cents. That's the apparent intervention of System 1 that provides us with a fast and effortless solution, that sounds so true and logical that no signal is being sent to System 2 in order to activate it and challenge the answer. Our mental logic follows the primitive psycho-logic: *'if it feels true, then it must be true'*. Kahneman has clearly stated that, *"System 1 is designed to jump to conclusions from little evidence"*.

The bat and ball is just one of the cognitive reflection tests (CRT), that Shane Frederick mentions in his ground breaking paper: *'Cognitive Reflection and Decision Making'* (Frederick, 2005). CRT measures a person's tendency to override an incorrect intuitive reaction and engross in additional reflection in order to find a correct answer. CRT has a moderate positive correlation with

measures of intelligence (i.e. IQ tests), and a high positive correlation with several measures of mental heuristics. Some other well-known CRT tests are the following (Frederick, 2005):

'If it takes 5 machines 5 minutes to make 5 widgets, how long would it take

100 machines to make 100 widgets? _____ minutes'

'In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take

for the patch to cover half of the lake? _____ days'

Frederick and Kahneman have worked and researched together, and that is apparent in the way they approach cognitive thinking. The questions that they use to test the various subjects focus primarily on numerical reasoning. Mathematics is a 2nd nature skill that requires arduous System 2 training. Consequently, we should keep that in mind when we analyse and judge the way our subjects behave.

Apart from the north – south (System 2 and System 1 respectively) split of our brains, we should also be aware of the east - west one (right and left hemisphere) as well. Each part of our brain serves different purposes (Goleman, 2011). The right part has stronger connections with the emotional centres, i.e. the amygdala and the subcortical regions but when it comes to creativity, the entire brain is activated. The left part is more independent consisted of perfectly stacked vertical columns that allow the differentiation of separate mental functions, while limiting their integration. Nonetheless, when it comes to

creative thinking the entire brain is activated through a large web of connections. Problem solving can definitively be approached as a creativity trigger, especially when an out of the box approach is the most desired one. In that case Wallas's 4 stages of creativity(Sadler-Smith, 2015; Wallas, 1926) actually do kick in: preparation, incubation, illumination, and verification

- In the first stage, our brain is gathering information.
- In the second stage, we let our mind wander and stretch our ideas.
- In the third stage, we make connections between ideas.
- In the fourth stage, creative ideas are polished by critical thinking in order to persuasively reach their audience.

1.3 Interpretations and Applications

Forecasting has been out there for thousands of years, practiced by magicians, witches, oracles and recently by business people, pundits and scientists. Its existence throughout the ages should thus be founded on something solid, otherwise its lifespan would have been way shorter. Whether justified scientifically or not, there is something intrinsic in it that makes it 'tick'. It is not by chance that forecasting has come about to become a rather separate scientific field attracting the interest of both corporate and scientific worlds. Philip Tetlock and his team made a breakthrough with their recent

findings under the Good Judgement Program (Philip Eyrikson Tetlock & Gardner, 2015). The scientific turbulence brought about, triggered even further research and since then an active community has emerged dealing and researching on the field of 'superforecasting'. That is a tournament based approach that helps decision makers identify gifted forecasters and exploit balanced and quantitative forecasts to derive to more justified and accurate decisions.

It appears to be a growing demand on this field coming from numerous discrete fields/stakeholders, like:

- Policy makers
- The intelligence community
- The wider defence sector
- The corporate world
- Academics and scholars
- And lately even SMEs

Forecasting and more specifically, superforecasting, have developed their own dynamics and there is an apparent high mobility in this area, with the aim to get it even further. Nevertheless, there will continue to be conflicting arguments in terms of how far can 'forecasting' go. A pundit will always be a pundit, and a pure disruptor in forecasting's 'brand name'. Additionally, intentionally twisted forecasts will continue being the case in many situations,

where the self-fulfilling prophecy has to be served, by all means. Scott Adams, managed to capture this concept in one of his famous sketches⁵ :

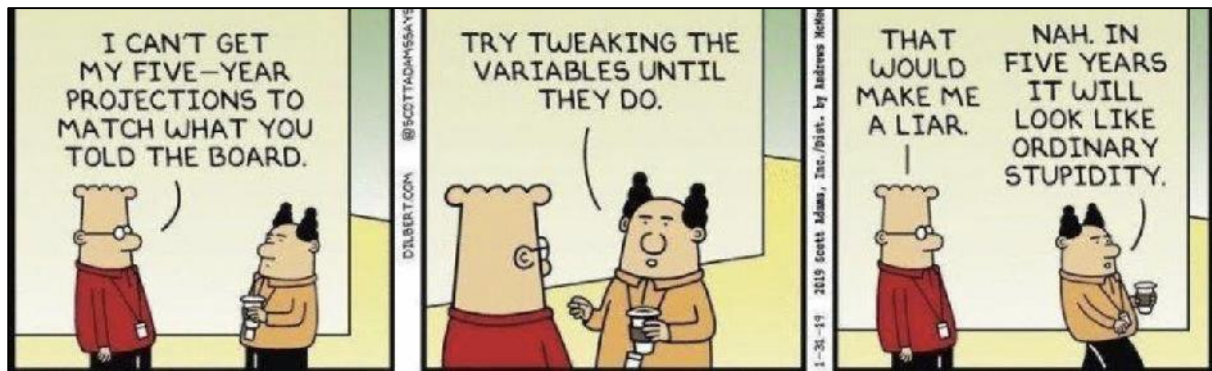


Figure 2: Dilbert and forecasting

So now we do realize why Warren Buffett⁶ once said: *'Forecasts may tell you a great deal about the forecaster; they tell you nothing about the future.'* This is actually an extreme approach, but one cannot deny that there is some validity in it. The funny thing is that even Buffett himself verified his quote when he started making claims that everybody should stay away from Bitcoin considering it 'just a joke'! In that vein, it is apparent that not all forecasts are either justified or supported enough.

Additionally, the miss-perception of the wider audience of what a good forecast is, will continue being a huge disruptor. Take for example the weather forecasts. Most of us have a twisted view of the accuracy of weather forecasts.

⁵ <https://dilbert.com/>

⁶ https://en.wikipedia.org/wiki/Warren_Buffett

We totally ignore the supporting probabilities of each weather forecast and thus judge them as deterministic events. Furthermore, the effect of hindsight bias (Kahneman, 2013; Tversky & Kahneman, 1974) is so immense that we have become the most fierce judges of the respective forecasts. The truth is that weather forecasting is, if not the most, one of the most accurate forecasting practices. The immense amount of traced historical data serve as calibrating factors that have significantly helped the positive evolution of the specific scientific field.

To bring everything together, it seems that forecasting evolves within the constraining field of bounded rationality (Simon, 1979). Bounded stipulates that rationality is constraint when one makes a decision by the controllability of the decision problem, their cognitive limitations and the available time. In that sense each decision maker would act like a satisficer, rather than somebody trying to identify and act upon the optimal solution.

The scientific community is in our case the key contributing factor in the evolutive chain of forecasting, helping identify all these elements that wither its validity, but above all its reputation.

1.4 Research Overview

The principle differentiation of this particular research procedure will be its focus in forecasting events with minimum historical past. This particular characteristic deprives the potential forecaster from using a wide variety of

tools, like extrapolation, time series analysis etc. thus making forecast accuracy an ambivalent outcome.

The aforementioned difficulty will, in no way become an impediment to our effort for providing justifiable and accurate forecasts. Even ancient literature, and in particular Plutarch's «On the "E" at Delphi» (Charles William King, 1908) provides sufficient justification for triggering a research towards the improvement of our forecasts:

«Nothing comes into being without a cause, nothing is known beforehand without a reason. Things which come into being follow things which have been, things which are to be follow things which now are coming into being, all bound in one continuous chain of evolution. Therefore, he who knows how to link causes together into one, and combine them into a natural process, can also declare beforehand things».

Hundreds of years later, and particularly in 1814, the French mathematician and astronomer Pierre-Simon Laplace (Laplace & Dale, 1995) made an equivalent statement:

«We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to

analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes. »

That was Laplace's 'demon': an omniscient entity, entirely knowledgeable of the present, thus fully capable of predicting the future.

Both statements, the one from Plutarch and the one from Laplace, highlight the feasibility to predict the future, but each inserts a different hindering parameter. Plutarch focuses on the identification of causal relationship between events, and Laplace on the unfeasibility of one to be fully aware of all the events - facts pertaining to the forecasting question. Nevertheless, none of the two approaches deems forecasting as an 'exercise in futility'.

This facts-based study takes under consideration both the finding of the 'giants of the past' as well as recent scientific discoveries and aims to identify a new path on the way to efficient and effective forecasting procedures. It thus evaluates the performance of an adapted version of "Forecasting by Analogies" (Armstrong, 2001), which blends the principles of Structured Analogies (Green & Armstrong, 2007) and semi-structured analogies (K. Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015; N. D. Savio & Nikolopoulos, 2013; N. Savio & Nikolopoulos, 2010) with the "Superforecasting" approach (Dhami, Mandel, Mellers, & Tetlock, 2015; Philip Eyrikson Tetlock &

Gardner, 2015). In particular, participants after undergoing a short (example based) e-training on the proposed approach, provide their forecasts, in the form of probability estimations for a number of special events.

We try to prove through the implementation of the aforementioned approach that forecasters put easier aside the “chaotic” feeling of “how do I give my estimation?” and thus deliver more accurate forecasts.

All the above is achieved within a constrained environment deprived of the luxury of infinite resources, adequate information and unlimited time. We accomplish to verify the superforecasting concept and identify consistent and accurate forecasters within a relevantly short period by following a tailored approach of structured analogies, that allows for flexibility to the forecaster. At the same time, it paves the way towards a sane forecasting method while trying to diminish bias influence.

A special set of rules that we have imposed to our experiment was the following:

- All participants had to answer almost all forecasting questions
- All participants had to provide an early forecast, within the 1st ten days each new question was provided. Subsequently they could update their forecasts at their discretion according to their personal judgement and the relevant information flow.

It is apparent that the above approach clearly deviates from the one applied to the forecasting tournaments run under IARPA's project (Office of the Director of National Intelligence, 2017) where each participant had the 'luxury' to reply only to his preferable questions and at the time of his/her discretion. The reasoning behind our approach was that in the corporate world, companies have only a relevantly limited pool of resources and thus by allowing an excessive amount of flexibility they would risk receiving just a few forecasts, if none at all. The above constraint is even more apparent in SMEs (Small-Medium Enterprises) where the pool of resources is significantly downsized.

The present research also allowed us to prove that there existed a positive correlation between overall accurate forecasts and early accurate forecasts. In particular, we derived to almost identical conclusions when we compared the participants' performance throughout the entire lifecycle of the experiment with their performance during the early stages of each question. The most accurate forecasters at the early stages of each question, were the ones that ended up within the superforecasters pool. Although logical on the first reading, the present finding is of rather great importance. Superforecasters were not extremely conservative in their early forecasts, staying around 50% just to play safe and protect their scores. In most of the cases their scores were very close to the extremities (0% or 100%), and yet proven accurate in the long run, despite the 'time chasm'.

An additional innovation in our experimental approach was the training provided to our 'benchmark team'. We did not try to go for the easy comparisons, but rather go for the extra mile. The participants were split into two teams, one of which served as the performance benchmark for the other. Both teams received a common foundational training on probabilistic reasoning and de-biasing. Best practices already identified and verified during previous research were considered as facts (and left intact) and further modifications were made in order to derive to alternative "profitable" approaches. In other words, we implemented Occam's' Razor (Simon, 1979), in terms of accepting the simplest theory that works and elaborating further on, by utilizing particular methods in suitable context (Gigerenzer, 1996). Our approach rather minimized the 'hooray' effect after analysing our results, but yet again it provided adequate information to prove that even by competing with a higher standard benchmark, the method actually works.

The training approach that we followed for our participants was that of micro-learning. We were using small, snackable and easy to 'digest' training content, in the form of video tutorials, that were always available whenever a participant would like to revisit them. There exists vast research on the time related attenuation of the training outcomes. An one-off training would have made apparent the diminishing returns on forecasting performance. The attenuation effects would have been even more ostensible in the case where

the pre-experiment training would have been long and complex. The domain specific 8-9 minutes training content that we were providing had no 'intimidation' impact on the forecasters, that showed a tendency to revisit them, again and again, in order to refresh their knowledge on the selected topic.

Beyond the use of a trained sample for comparison purposes, in our research we also highlight the importance of using more advance quantitative performance measures. The combinational power of judgemental forecasting with quantitative methods has been under scientific investigation for many years. There exist numerous studies analysing the contribution of judgemental forecasting for quantified data, either in the form of a standalone forecast, or as an adjustment forecast. Even the Good Judgement Project has moved towards that direction and now has placed under close investigation the combinational power of judgemental forecasting, as provided by the tournament forecasters and quantified forecasts, produced by complex algorithms. A contributing factor towards the combinational use of judgemental and pure quantitative forecasts, is the blowing effect of big-data, Artificial Intelligence (AI) and Machine Learning (ML).

Over the years, the IIF community has seen many forecasting studies that proposed new methods that could only outperform Naïve, a moving average or just ETS, although it has been obvious for the last two decades that there is a series of very accurate methods, which are computationally cheap and free

in R and Python packages. We acknowledge that through our proposal, if it eventually gets accepted by the scientific community, the life of many researchers will become much more difficult. Our aim is not that, but to help scientific research produce more robust results that highlight the comparative performance of innovative approaches to the best benchmark, and not just the 'lowest performance possible'.

1.5 Research question and thesis

The present research seeks to identify viable answers and solutions to a wide set of questions. The utmost aim is to describe a set of rules and procedures that will foster a healthy forecasting environment for SMEs, in order to allow them to have a rational and cost effective approach in deriving to safe forecasts, adequately justified and ready to feed the decision making process.

Therefore, our key questions are:

- How could we identify the gifted ones (superforecasters) within a rationally small pool of resources (forecasters)?
- How could we define the framing characteristics of these super-forecasters in order to streamline the mapping process for future forecasts?

- How feasible is it for an SME to anticipate to identify superforecasters within its limited capacity?
- How should an SME frame its rules and procedures in order to be able to make adequate use of its resources in the field of forecasting and decision making?
- How does time influence the identification process?
- What is the contribution of training in forecasting performance?
- Does the use of analogies facilitate forecasting performance, to the extent that it could be considered the primary methodological approach?
- What form of training is the most efficient and effective?
- What is the contribution of micro learning in retaining already acquired knowledge?
- How should one engage with dual mode forecasting: Judgemental + probabilistic?

Chapter 2 Examining the status quo in forecasting

2.1 Judgemental Forecasting

The research on the topic of judgmental forecasting is quite immense (M. Lawrence, Goodwin, O'Connor, & Önkal, 2006). Furthermore, there exists several studies that compare the relative performance of judgmental forecasting to the statistical one, with outcomes varying, in a case by case state (Carbone & Gorr, 1985; M. J. Lawrence, Edmundson, & O'Connor, 1985; O'Connor, Remus, & Griggs, 1993; Sanders, 1992).

In a corporate environment, it has been proven that expert management judgement (expressed as a forecast) has significant importance for the company's decision making process (Fildes & Goodwin, 2007), either as a form of adjustments to statistical forecasts (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009) or as a standalone process in conditions where either there is no historical data, or it is chosen that historical data and/or statistical forecasts should be ignored (Franses & Legerstee, 2010).

To sum it up, it can be considered that the judgmental approaches are very helpful but their relative effectiveness is tangled to a number of limitations, with the most salient being the forecaster's inherent biases (S. G. Makridakis, Wheelwright, & Hyndman, 1998). Makridakis in his 2010 MIT SLOAN article (S. G. Makridakis, Hogarth, & Gaba, 2010) encapsulates the extent of forecasting uncertainty within a single phrase '*Human beings are often extremely surprised by the extent of their forecasting mistakes. If statistical models were capable of emotion, they would be surprised by the size of their errors too*'.

The more recent findings on Judgmental Forecasting, came from a wide research sponsored by IARPA (Intelligence Advanced Research Programs Activity), in the form of a series of geopolitical forecasting tournaments, which managed to shed light on the strategies being used for making intuitive probability judgments. Particularly, throughout its four-year duration, the "Good Judgment Project" (Office of the Director of National Intelligence, 2017; Philip Eyrickson Tetlock & Gardner, 2015) team from Wharton University, managed to prevail in all four consecutive years, and its approach is considered a "lighthouse" in the recent scientific literature.

The key findings of the above research present as reinforcing explanations of the "superforecasting" performance four principal elements:

(a) cognitive abilities and styles,

(b) task- specific skills,

(c) motivation and commitments, and

(d) enriched environments (B. Mellers et al., 2015).

Furthermore, it designates the following key (psychological) drivers of accuracy:

(a) recruitment and retention of better forecasters,

(b) cognitive de-biasing training,

(c) more engaging environments in the form of teamwork and prediction markets and

(d) better statistical methods (P. E. Tetlock, Mellers, Rohrbaugh, & Chen, 2014).

At this point we should not put aside a prior study conducted by P.E. Tetlock, where he focused on judgmental shortcomings. He clearly portrays “experts”, or pundits in some cases, as no better in making long term predictions than most people. Furthermore, he pin-points the lack of accountability as a critical contributor to the propagation of bad forecasts (Gardner, 2011; Philip E. Tetlock, 2005).

2.2 Forecasting by Analogies

The analogical ability is intrinsic in human cognition and could be considered as the cornerstone in human evolution. In particular, the analogical reasoning, as described in cognitive psychology, is a kind of reasoning that takes place between specific and discrete cases, where, what is known about one case is used to infer new information about the other. The ability to perceive and use purely relational similarity could be defined as the major contributor for the remarkable human mental powers (Gentner & Goldin-Meadow, 2003; Penn, Holyoak, & Povinelli, 2008). This is further certified by studies that were conducted on the history of science. Those studies, clearly show that analogy was a means of discovery for scientists like Faraday (Tweney, 1991) and Kepler (Gentner, 2002). The principal tool for analogical reasoning is "Analogical Mapping" (Gentner, 1983), which presupposes aligning the two situations (known and unknown) in terms of finding the correspondences between the two and projecting inferences from the one serving as the base, to the other serving as the target.

Gentner (1983) described the *Structure mapping theory* which aims to define the psychological processes that carry out the analogical mapping. A key concept of the above theory, that is taken under consideration for the present research as well, is that the comparison process involves finding a maximal alignment between base and target cases that reveal common relational structure.

By “extrapolating” human mental processes like the above, we can make logical inferences about the effectiveness of analogical reasoning in providing accurate forecasts. The concept is not far-fetched given recent literature verifies it. In particular, text books, dating back in the 1930s, describe analogies as a tool for economic and business forecasting (Green & Armstrong, 2007). Furthermore, Khong Y.F., in his book “*Analogies at war: Korea, Munich Dien Ben Phu and the Vietnam Decisions of 1965*”, (Khong, 1992) clearly describes the decision process during the Vietnam War, as one founded on the utilization of analogies. Deriving from the aforementioned, Green and Armstrong (2007) formulated the structured analogies (SA) approach which tries to blend together all the positive aspects of analogies and rule-out potential inconsistencies, namely bias.

The SA procedure, involves the following five steps, in chronological order:

(a) Administrator

- a. Description of target situation, in cooperation with Subject Matter Experts (SME).
- b. Selection of experts under the criteria of possessed domain knowledge and relevant experience. A lower threshold of five

forecasters is considered crucial for the effectiveness of the method (Armstrong, 2001).

(b) Experts

a. Identify and describe analogies, without considering the extend of the similarity to the target. They should also match analogy to target outcomes.

b. Rate analogy similarity to target situation, by using a predefined scale.

(c) Administrator

Derive to forecast using a predefined mechanical rule

A complementary approach to the above, is the one proposed by Savio and Nikolopoulos (2010,2013), which suggests the utilization of a semi-structured methodology that is different from the original SA (Green & Armstrong, 2007) in the below sectors:

(a) Active involvement of administrator in deriving forecasts from provided, rated analogies is suspended.

(b) Exploitation of expert knowledge through the provision of point forecasts within a preselected 90% prediction intervals

Finally, another study by (Litsa, Petropoulos, & Nikolopoulos, 2012), implemented a slight variation of the original SA approach, were instead of experts, they used semi-experts (forecasters that were trained in forecasting

techniques but if they did not possess domain knowledge on the topic, they were called to provide their forecasts for) for forecasting the success of a policy implemented by a European Government. The results revealed relative positive outcomes for the performance of the modified SAs.

2.3 Forecasting and Biases

Given the human mind principally (~95%) works on intuitive mode (Lakoff & Johnson, 1999) we have a natural tendency to use heuristics in order to save time and effort (Tversky & Kahneman, 1974). Those systematic errors, are called biases (Kahneman, 2013) and the number of them is quite immense (Montibeller & von Winterfeldt, 2015). Biases are therefore inherent in human judgment and not necessarily with a strictly negative meaning. They are meant to serve specific needs that have helped the human species to survive throughout the ages (Kahneman, 2013).

Nevertheless, these biases occasionally hinder the decision making process, given they primarily rely on automated mental mechanisms. Those biases are principally "Type 1" errors (outcomes of intuitive processes), that drive the decision making process. Although Type 1 prevails in the bias world, we should not rule out Type 2 errors (those deriving from slow systematic

thinking), that might be the outcome of erroneous strategy selection or imperfect decision rules (Arkes, 1991).

The key to counteracting biases, is de-biasing. In a very recent paper (Chang, Chen, & Mellers, 2016) that summarizes the impact of various training practices on judgmental accuracy in geopolitical forecasting tournaments, de-biasing techniques, are organized into four major categories:

- a) didactic,
- b) process based,
- c) feedback based and finally
- d) format based.

The effectiveness of each of the above techniques is not the same, yet their combination can lead to enhanced levels of performance, in terms of mitigating or even out-ruling biases.

In a recent paper by Liu et. al. (Liu, Vlaev, Fang, Denrell, & Chater, 2017) a complementary approach to de-biasing is presented. Given de-biasing primarily focuses on the 'restructuring' of our slow (but yet effective) system 2 thinking (Kahneman, 2013), the so called 'MindSpace Approach'⁷ (Dolan et al., 2012) aims at triggering our fast System "1" thinking towards getting faster and more reliable decisions. The present research did not make use of the

⁷ https://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE-Practical-guide-final-Web_1.pdf

specific approach given for experimental purposes we aimed at limiting external influence on the participants.

2.4 Forecasting and Training

The GJP (Office of the Director of National Intelligence, 2017) and all subsequent research, did not only highlight new ways of identifying superforecasters and using their competences to feed the decision making process, but also went back and performed a profound research on what was already there, spanning from training, to biases, to scoring, to aggregating and many more. Chang et al. (2016) provided us with a clear outline of this overview both in terms of topics and methods of delivery. They clustered training into a set of modules - principles that were defined by the gradually elaborated acronym 'CHAMPS KNOW⁸' (4th year elaborated version).

In general, the key training areas identified and squeezed within an hour's distant learning module were the following:

- Identifying and tackling biases

⁸ CHAMPS is the acronym for: (1) **C**omparison Classes, (2) **H**unt for the right information, (3) **A**djust and update forecasts when appropriate, (4) **M**athematical and statistical models, (5) **P**ost mortem analysis, (6) **S**elect the right level of effort to devote to each question
KNOW is the acronym for: (1) **K**now the power players and their preferences, (2) **N**orms and protocols of institutions, (3) **O**ther perspectives should inform forecasts, (4) **W**ildcards, accidents and black swans

- Probabilistic reasoning, including
 - The use of base rates
 - The basic principles on belief updating, including the Bayesian approach
 - The value of taking under consideration different estimates
- How Brier Scoring works (difference between calibration and resolution)

The above approach was highly effective and during the 4th year of the tournament, the trained forecasters had better brier scores by 7%. Beyond the pure score improvement, trained participants were exhibiting improved calibration and resolution by reducing overconfidence.

When one addresses great audiences aiming to recruit and train forecasters in order to participate to a large scale tournament, cannot afford the luxury of residential training. The f2f training although (traditionally considered) more effective when it comes to knowledge retention, cannot be considered as a cost effective, not to mention 'viable' solution. Asynchronous e-learning can thus be considered as the optimal solution in order to tackle large audiences and yet manage to pass the key messages.

Some of the principal ground rules to be considered when setting up an asynchronous e-training module are as follows:

- Motivate participants (Law, Lee, & Yu, 2010): We need to substitute the missing interaction with the instructor with other actions that promote learner engagement. A best practice is to communicate tangible

course goals up-front that highlight the usability of the course content. Another effective approach is to use realistic scenario training, with the scenarios being similar to the situations that the trainee is going to face in his or her day-to-day job. Furthermore, having a clear view of Keller's ARCS model (Keller, 1987) of instructional design helps us understand the major influences on motivation to learn. (A concise summary of the model can be found on the Learning Theories website ("ARCS MODEL OF MOTIVATIONAL DESIGN THEORIES (KELLER)," n.d.).)

- User-friendly interface: The graphical user interface (GUI) is of tremendous significance in an e-learning course (Ahmad, Basir, & Hassanein, 2004; Zhang & Nunamaker, 2003). We should aim for the best first impression and active engagement by using clear navigation schemes and well-structured content.
- Keep participants interested by incorporating variety in the learning activities: The list of potential tools is endless and could include interactive simulations, case studies, quizzes and games.
- Content chunking (Clark & Mayer, n.d.; Mayer & Moreno, 2003; Mödritscher, 2006): The human brain, which is capable of storing a quadrillion bytes of data and performing extremely complex operations, slows down to the speed of a snail when asked to recall 10 numbers or

repeat just a few simple words. This has to do with the actual working memory of a human brain (similar to the RAM in our PCs). In e-learning, content chunking is the process of presenting content in the form of crisp sentences and bulleted or numbered lists. Instructional designers break down long strings of information into bite-sized absorbable pieces, helping learners to stay focused.

- Include effective assessment strategies (Roberts, 2006; Wang, 2007): Constructive feedback as an outcome of a well-structured assessment has multiple positive effects over the training experience. Apart from helping learners identify their weak points, it improves training effectiveness by boosting memory retention. The effort of retrieving information (no matter the outcome) makes it easier to retrieve when needed (Lahey, 2014; Roediger & Karpicke, 2006).

2.5 Forecasting Tournaments

Evolving from forecasting competitions to forecasting tournaments. Some would say that there is no difference between the two, but just a use of a different noun. In fact, these two are totally different!

Although one could claim that forecasting competitions date thousands of years before, where oracles were competing on their accuracy (Raphals, 2013), the actual breakthrough with forecasting competitions came about in 1998, with the famous M-competition, by Makrydakis and Hybon. We have

recently reached the 4th version⁹ of the famous competitions (M-4) and the scientific community is still startled by the never ending findings. The key characteristic of these competitions is that they are performed over existing time-series (Hyndman, 2019). This is indeed the key difference with the forecasting tournaments, that are actually open to any type of forecasts, spanning from pure judgemental, to hybrid, or even purely data driven.

Forecasting tournaments were actually introduced into the scientific community by the Intelligence Advanced Research Projects Activity (IARPA) that created the Aggregative Contingent Estimation (ACE) Programme (IARPA, 2010) on June 30, 2010. The objective of the specific programme the cost of which was tens of millions of dollars, (as stated in the respective call: IARPA-BAA-10-05) was to: *'develop and test methods for generating accurate and timely probabilistic forecasts, leading indicators and early warning of events, by aggregating the judgements of many widely – dispersed analysts'*.

Forecasting tournaments challenge the participants to give accurate probabilistic estimates for a wide range of events while at the same time offer a constant cross-check on the accuracy of each provided forecast (Philip Eyrikson Tetlock & Gardner, 2015). Tournaments form a special social

⁹ <https://www.mcompetitions.unic.ac.cy/>

environment where three key debiasing factors are being placed to work (Barbara Mellers, Tetlock, & Arkes, 2019):

- Forecasters are accountable for the accuracy of their views.
- Forecasters are forced to acknowledge opposing opinions and limitations in their views-knowledge and thus reduce their overconfidence.
- Forecasters are advised to engage in perspective thinking.

The ACE programme, included five competing research programmes, all aiming to generate the most accurate probabilistic estimates. Each of the participating entities (consortiums) formed its own tournament and was fully flexible to develop and test its own methods for the provision of the forecasts.

The Good Judgment Project (GJP) was eventually the one to win the IARPA/ACE tournament. The statistics were shocking both for the academic and the intelligent community. Its forecasts were on the right side of 50/50 on 86.2% of all daily forecasts, outperforming the simple average of the control group by 60% and other teams by 40%. Upon analysis of the outcomes of the performed factorial design, further conclusions were drawn, identifying some key psychological drivers of accuracy (P. E. Tetlock et al., 2014):

- Recruitment and retention of better forecasters
- Cognitive-debiasing training
- Engaging work environments, in the form of collaborative teamwork and prediction markets

- Better statistical methods of distilling the wisdom of the crowd — and sorting out the madness.
- Creating superforecaster teams including the top 2% performers of each tournament

The results collected by the GJP paved the way for another tournament that kicked in almost a year ago, and brings together human predictions and artificial intelligence. The new tournaments are organized by the new IARPA project that comes by the name: Hybrid Forecasting Competition (HFC) (IARPA, n.d.).

According to IARPA:

'The HFC program is developing and testing hybrid geopolitical forecasting systems. These systems integrate human and machine forecasting components to create maximally accurate, flexible, and scalable forecasting capabilities.'

Hybrid approaches hold promise for combining the strengths of these two approaches while mitigating their individual weaknesses.'

It is obvious that, forecasting tournaments are here to stay. They proved to be an indispensable tool, both for the academic and the intelligence community (IC), that brings to the light the actual weaknesses of our

forecasting approaches, by benchmarking them against reality and enforcing accountability in the form of score penalties.

Chapter 3 Methodology

3.1 General

In order to provide solid answers to our research questions, as stated in [Section 1.5](#) we have stratified our research in four discreet sections, as follows:

1. The primary topic of the present research is included in our paper titled *'Looking for the needle in the haystack: Evidence of the superforecasting hypothesis when time and samples are limited'*. The present paper has been submitted to the European Journal of Operational Research and is under review.
2. Subsequent analysis on the early performance of forecasters is included in our working paper titled: *'Early vs late forecasting: Do forecasting tournaments help us identify a time related performance of forecasters?'* In the present paper we analyze the effectiveness of our method (as described in our previous paper) on the performance of the forecasters at the early stages (1st ten days) of each forecasting question.
3. Our paper titled: *""PESCO - PM2 - ESDC"" could e-learning bring closer together EU's success stories?'* presented at the 14th edition of the

International Conference, e-Learning and Software for Education, makes a comparative presentation - analysis of the educational approaches that can be followed in order to address the needs of various types of stakeholders. Its principal focus is on e-learning and in particular, micro-learning, the training approach implemented as well in our superforecasting experiment (For more information see analysis in [Section 4.1.6](#)).

4. On our short and sharp paper, titled '*On the M4.0 forecasting competition: can you tell a 4.0 earthquake from a 3.0?*' that was accepted for publication at the International Journal of Forecasting, we present a solid argumentation on the benchmarking approaches that should be followed when comparing the performance of new forecasting approaches. We try to raise the bar of performance by using as benchmarks already established and validated best practices, and avoid the typical comparison to the typical lowest threshold.

The methodological approach followed in order to achieve the aforementioned is described in the following chapter.

3.2 Methodology

The experiment was conducted in the context of the wider public and private sectors, plus the Academia. Given the special nature of the

participants, mutually agreed confidentiality rules applied. These rules were further defined upon negotiation.

3.2.1 Procedure

The general procedure that was followed, is described in the below flowchart:

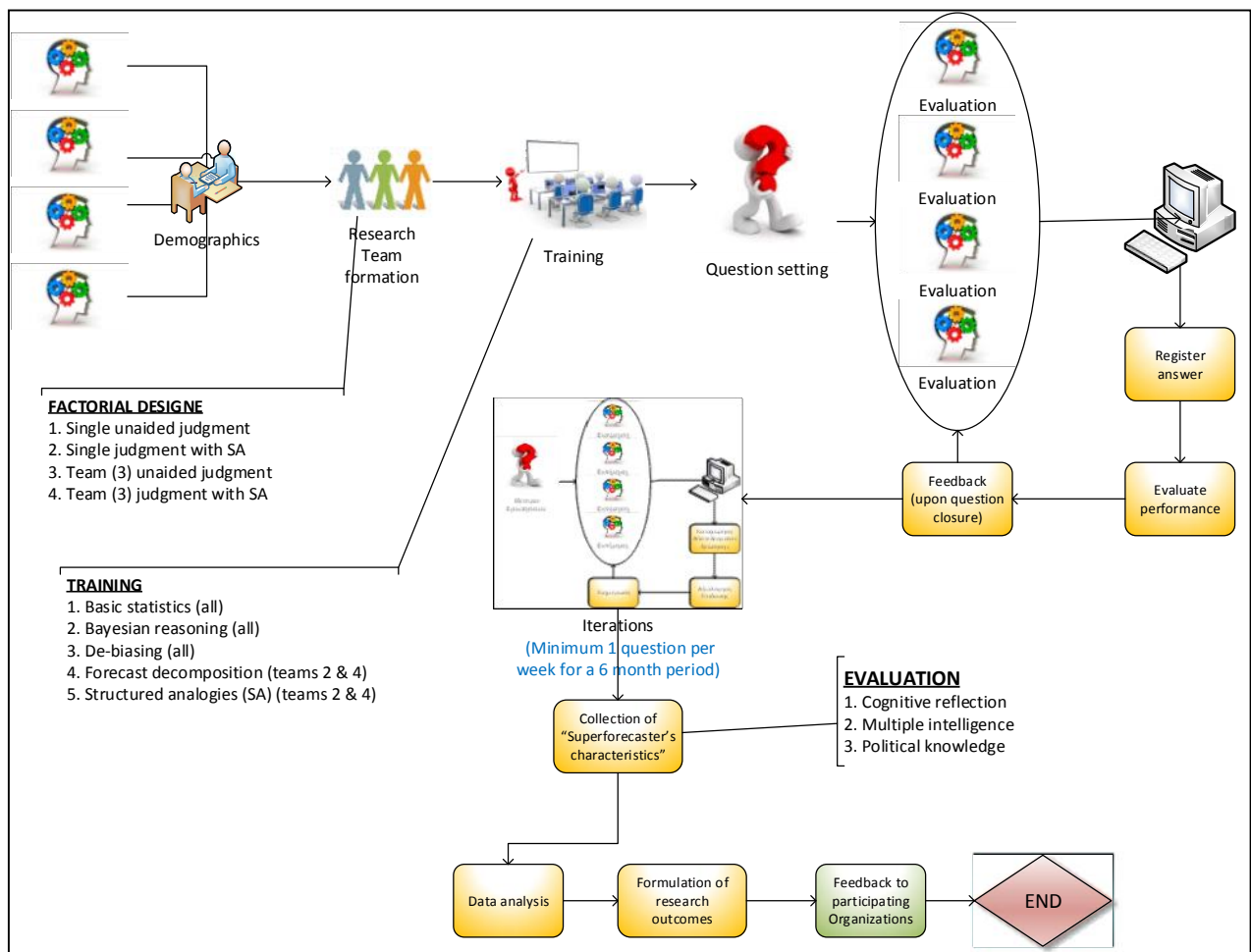


Figure 3: Research methodology flowchart

Participants were engaged with the tournament through a carefully tailored web interface that allowed them to interact only with elements referring to their assigned status.

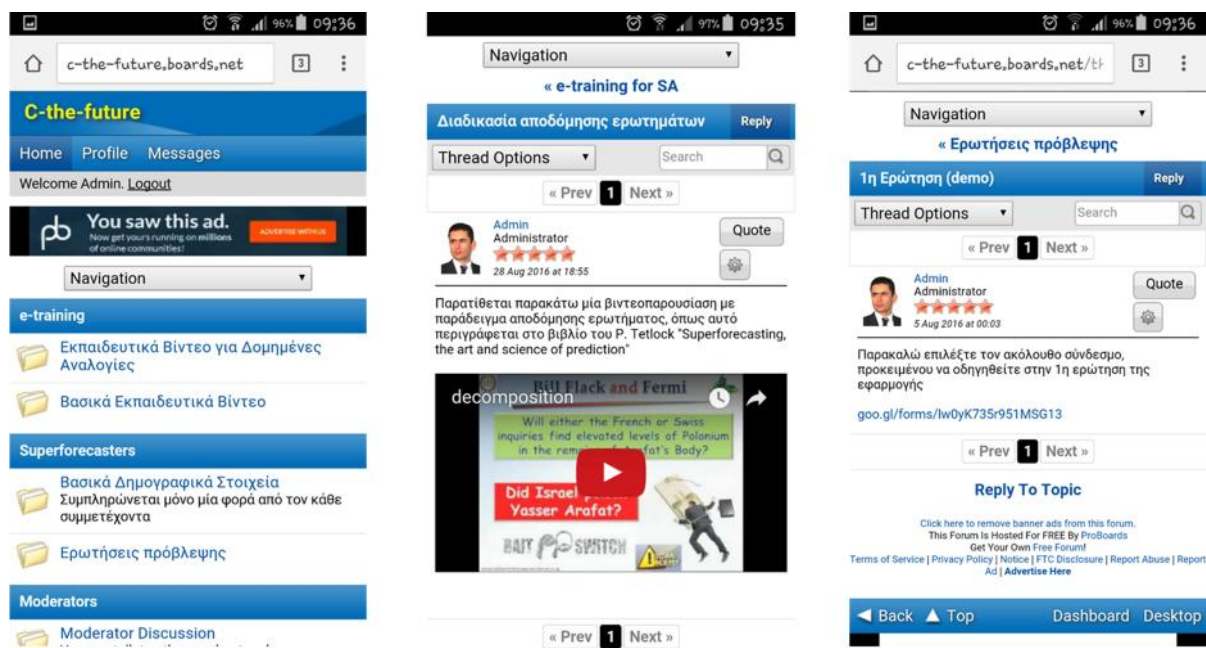


Figure 4: Experiment's web interface¹⁰

Below follows an analysis of the basic steps depicted in the above flowchart

3.2.1.1 Initial demographics

In order to be able to make actual comparisons between forecasters, we needed to register as much as possible of the participants' background status.

¹⁰ <http://c-the-future.boards.net/>

As a prerequisite for entering into the forecasting tournament, an initial survey was delivered that was imperative for all to complete prior entering into the forecasting tournament. Given all participants were Greeks the survey was delivered in Greek and is included in Appendix E as reference.

3.2.1.2 Factorial design (research team formation)

All participants to the experiment were divided into 4 discrete teams, as follows:

- Single unaided judgment, where each forecaster will provide his judgment (forecast) on his own, within his own capability framework
- Single judgment with the use of a modified version of Structured Analogies (SA). Here, each participant will be called to provide and justify his forecast by using the aforementioned method (SA).
- Team unaided judgment, where forecasters will work and cooperate anonymously.
- Team judgment with the use of the above mentioned, modified version of Structured Analogies.

To further analyze forecasting in the team framework:

- All teams were containing three individuals

- Anonymity was achieved through the use of coded email accounts (see Appendix “E”), thus allowing participants to work in a “Delphi method” alike environment.
- Each participant in the team was prompted to provide his particular forecast after having exchanged options with the rest of his team members (there was no “per team” forecast).

3.2.1.3 Training

Building on Meller’s findings, during the Good Judgment Project, several e-training sessions were created and were available to forecasters, through a web interface¹¹.






These e-trainings included the following, short modules:

- Basic knowledge in statistics and probabilities
- Bayesian reasoning
- Forecast decomposition
- Cognitive de-biasing
- Structured analogies

All training followed the concept of micro learning and were condensed into short 7-10 min video tutorials that were constantly available for revisiting.

¹¹ A hyperlink was provided from within the app that redirected forecasters to attend specific asynchronous e-training modules.

All video tutorials are available to all, through the below links:

Clip preview	Description	URL
	<p>Basic video on de-biasing</p>	<p>https://youtu.be/vA1hu_hdnXI</p>
	<p>Advanced de-biasing</p>	<p>https://youtu.be/wBKrlBckdCY</p>
	<p>Basic statistics</p>	<p>https://youtu.be/7UvZYFKPdAM</p>
	<p>Bayesian reasoning</p>	<p>https://youtu.be/6l29iT8dxVY</p>
	<p>Structured analogies</p>	<p>https://youtu.be/wndSekKzs8A</p>


	<p style="text-align: center;">Forecast decomposition</p>	<p style="text-align: center;">https://youtu.be/ngdFU9QyT_4</p>
-----------------------------------------------------------------------------------	---------------------------------------------------------------	------------------------------------------------------------------------------------------------------------

Table 2: Forecasting tournament training programme

3.2.1.4 Question setting

The effectiveness of the research, was primarily based on the selection of the proper questions. Therefore, all questions fulfilled the following criteria:

- They were not conveying lateral messages thus influencing the forecasters.
- They were articulated using the principles of “clean language”.
- The answer to each question was measurable and identifiable.
- The time horizon of each question was not longer than 6 months, in order to be able to effectively measure forecaster’s effectiveness within the experiment timeframe.
- The questions were provided by predefined appointees from the participating organizations (under the guidance of the researcher) in order to cover the organization’s special information needs.
- They did rely on confidential information.
- The answer to the questions was not falling under disclosure constraints.

- The statistical (and not the content data¹²) data from each question would be fully available to the researcher for scientific analysis and publication.

In Appendix F we provide an example of how the questions were presented to the participants (in Greek).

3.2.1.5 Question scoring

All provided answers to the questions being set were evaluated with the use of Brier Scoring and alike variants.

3.2.1.6 Post Research “Superforecaster” evaluation

Aspiring to have as many participants as possible, we avoided practices that tend to succumb the will for participation. In particular, forecasters did not undergo any psychological or mentality tests, prior to entering the tournament. After several informal interviews that were conducted with potential participants, it was made clear that such practices will definitely form a

¹² The particular questions being set and the corresponding justifications provided by the participating forecasters, will not be publicized. Only the corresponding statistical data will be available to the researcher for analysis and publication.

constraining factor, since they tend to expose personal traits and characteristics.

3.2.2 Ethical, moral and privacy boundaries

3.2.2.1 Ethical and moral issues

“There are a number of ethical principles that should be taken into account when performing research. At the core, these ethical principles stress the need to (a) do good (known as beneficence) and (b) do no harm (known as non-maleficence). In practice, these ethical principles mean that as a researcher, I need to: (a) obtain informed consent from potential research participants; (b) minimize the risk of harm to participants; (c) protect their anonymity and confidentiality; (d) avoid using deceptive practices; and (e) give participants the right to withdraw from your research (Lærd Dissertation 2016)”.

3.2.2.2 Confidentiality, privacy and relevant legal issues

Confidentiality agreements to protect sensitive data must be explicitly defined and established between the researcher and the participating organizations.

Both sides (researcher and participating organizations) should work together towards limiting confidentiality issues to the minimum acceptable level, thus promoting mutual interests.

3.2.3 Constraints & assumptions taken under consideration prior to the research

3.2.3.1 Constraints

- All research will be done using the web based interface “C-the-future”, developed by the researcher.
- The choice to develop a new similar application for the participating organizations does not fall under the scope of this particular research, and will be examined as a different project in due time.
- Forecasting questions will be provided by the participating organizations, in the Greek language upon cooperation with the researcher.
- All publications relevant to the present research, given they comply with the aforementioned restrictions, fall under the direct authority of the researcher and his supervising Professors.

3.2.3.2 Assumptions

- Approval will be granted by the participating organizations to conduct the experiment, under the aforementioned constraints.

- A significant number of participants will conclude the research cycle.

3.2.4 Incentives

As per §3.2.5 (benefits from research).

Furthermore, all participants that conclude the research cycle (answer to all questions being set), are entitled to the below benefits, regardless of their scoring:

- Certification from the University of Peloponnese/Department of Economics, stating that they successfully attended the corresponding training and completed a scientific tournament in judgmental forecasting,
- Participation free of charge, in a Project Management Professional (PMP) certification preparation course, provided by the researcher¹³ and hosted by the University of Peloponnese/Department of Economics.

3.2.5 Benefits from research

3.2.5.1 Benefits for participating forecasters

- Assess their skills in providing accurate forecasts

¹³ The researcher is a certified Project in Risk Management Professional (PMP & RMP), by the Project Management Institute (PMI) of the United States. For further information please check his LinkedIn profile at: <https://gr.linkedin.com/in/katsagounos>

- The gain of new knowledge,
 - Through their participation in the research procedure
 - By attending, free of charge relevant training courses (as per §4.4)
- Improve their reasoning skills as a natural outcome of the above-mentioned.

3.2.5.2 Direct and Indirect Benefits for the Participating Organizations

- Improvement of H-R management procedures
 - Enhancement of recruiting methodologies
 - Enhancement of appraisal procedures
 - Better allocation of resources
 - Contribution in creating well “calibrated” job descriptions
- Improvement in forecasting abilities:
 - Enhancement of planning procedures
 - Enhancement of risk management procedures
 - Effective and efficient use of resources (primarily monetary)
- Insight for future development.

Chapter 4 Managing decisions under constraints

4.1 Structured Superforecasting”: Judgmental forecasting, Supercharged

4.1.1 General

This research paper is the first within a series of research activities where we try to bridge the findings of the Good Judgement Project, with an already well established forecasting methodology, that of analogies. The reasoning behind our methodological approach, lays at the assumption, that the injection of a structured or semi-structured approach within the superforecasting framework will help forecasters provide more accurate probabilistic estimates, by using a tool that provides them with a framework of actions, without being very limiting in its implementation. Our target group is primarily SMEs where each resource is considered a valuable asset. The specific entities due to their limited size, primarily in terms of 'headcount', should pose a flexible enough tool, that will help them identify their skilled personnel at the earliest possible stage, and subsequently make use of their

forecasting capabilities, in order to feed the decision making process with more accurate information.

By analyzing our findings, we derived to the conclusion that within approximately 6 forecasting questions, we can have a somewhat clear view of the over performing and consistent forecasters and thus invest on them either in terms of using their early forecasts to feed the decision making process, or by using them as 'sources of knowledge', and trying to decrypt and replicate their way of thinking.

4.1.2 Abstract

The success of the Good Judgmental Project in harnessing the power of superforecasting naturally leads to the question of how one can implement the approach on a smaller scale with more limited resources in less time and with fewer participants. Small(er) corporate environments and SME-type decision structures are prime examples of contexts in which the modified superforecasting approach can be applied. In this paper, we will present a hybrid approach to judgmental forecasting in relation to special events in which we combine training of superforecasters-to-be with the concept of a modified version of structured analogies (s-SA), a staple of judgmental forecasting in the literature. We call the resulting approach 'structured superforecasting' and go on to illustrate its efficacy using samples of participants from both the wider public sector and the academic community. Specifically, we apply the above methodology and compare

performances employing a proper experimental design including training and control groups. Significantly, we found evidence for the superforecasting hypothesis both when working with smaller samples--a few hundred experts, and when the superforecaster selection process needs to be completed much faster—in less than a year.

Key words: Forecasting, Superforecasting, Structured Analogies, Judgmental Forecasting, Decision Making

4.1.3 Introduction

The past three decades have witnessed a phenomenal pivoting towards Judgmental Forecasting (JF), which demonstrates the importance of “critical thinking” in providing accurate forecasts. The “Superforecasting” concept (Philip Eyrickson Tetlock & Gardner, 2015) has clearly substantiated the above observation, while at the same time opening the door for further research into that particular domain.

The principal differentiation of this particular research procedure will be its focus on forecasting events with a minimal historical past. This particular characteristic deprives the potential forecaster of a wide variety of tools including extrapolation and time series analysis, making forecast accuracy an ambivalent outcome.

This difficulty will in no way impede our efforts to provide justifiable and accurate forecasts. Even ancient literature, and in particular Plutarch's "On the 'E' at Delphi" (Charles William King, 1908), provides sufficient justification for initiating research aimed at improving our forecasts:

'Nothing comes into being without a cause, nothing is known beforehand without a reason. Things which come into being follow things which have been, things which are to be follow things which now are coming into being, all bound in one continuous chain of evolution. Therefore, he who knows how to link causes together into one, and combine them into a natural process, can also declare beforehand things.'

This study evaluates the performance of an adapted version of "Forecasting by Analogies" (Armstrong, 2001) which blends the principles of Structured Analogies (Green & Armstrong, 2007) and semi-structured analogies (K. Nikolopoulos et al., 2015; N. D. Savio & Nikolopoulos, 2013; N. Savio & Nikolopoulos, 2010) with the "Superforecasting" approach (Dharami et al., 2015; Philip Tetlock & Gardner, 2015). Specifically, participants receive a short (example-based) e-training in the proposed approach before going on to provide their forecasts for a number of special events in the form of probability estimations.

By implementing this approach, we will try to prove that forecasters can set aside more easily the "chaotic" feeling of "How do I give my estimation?" and thus deliver more accurate forecasts.

4.1.4 Literature Review

An extensive literature review on this paper has been provided in Chapters 2 to 6 of the present Thesis.

4.1.5 Hypotheses

In the light of the above literature review, we draft our research questions as follows:

R1: Do we find evidence supporting the superforecasting hypothesis when there are constraints on sample size (the number of forecasters) and time (the duration of the experiment)?

R2: Does the extra training provided in structured forecasting approaches help create more superforecasters?

4.1.6 Project Design

The research was built on the foundations of the Good Judgment Project (P. E. Tetlock et al., 2014; Philip Eyrickson Tetlock & Gardner, 2015; Ungar et al., 2012). Best practices already identified and verified therein (e.g. the effectiveness of statistical reasoning and de-biasing in providing more accurate forecasts) were considered as facts (and left intact) and further

modifications were made in order to derive to alternative “profitable” approaches. In other words, we implemented Occam's' Razor (Simon, 1979), in terms of accepting the simplest theory that works and elaborating further on, by utilizing particular methods in suitable context (Gigerenzer, 1996). One principal differentiation between the present experimental procedure and the GJP is the fact that participants were incentivized to answer almost all questions. The approach was based on the grounds that:

- In a corporate environment, ruling out resources is a luxury we do not have, given that the aim is to maximize the exploitation of the limited resources available in the most efficient and effective way.
- Scarcity in responses would have produced a lot of ‘missing values’ thus creating analysis issues (Merkle, Steyvers, Mellers, & Tetlock, 2016).

4.1.6.1 Project Overview

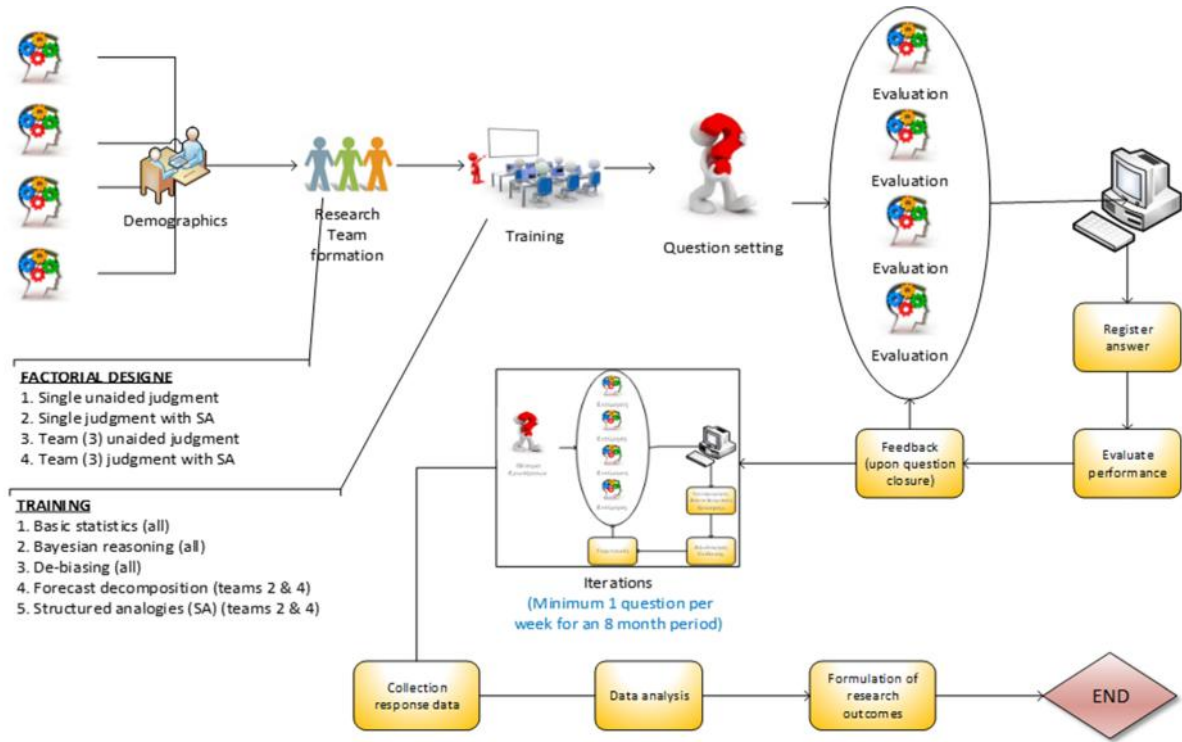


Figure 5: Final experimental structure flowchart

The project took the form of a forecasting tournament, and was conducted between November 2016 and July 2017. Participants were asked to submit their subjective probability estimates for a variety of time-bound questions using a custom-designed web interface. It should be noted that the term “subjective probability” is also known as “belief probability” or “personalist probability” (Hacking, 2001). It expresses a personal belief concerning the likelihood of an outcome and primarily relates to single events rather than repeated ones.

The project was divided into two phases. During the first phase, all participants were split into two groups (control and trained) and were asked to provide their forecasts on their own, without interaction. The second phase included team work, and the participants of each of the aforementioned groups formed teams of three and were asked to provide their individual forecasts after having anonymously collaborated with their team mates. This paper analyses the outcomes of the first phase only.

4.1.6.2 Subjects and research design

Subjects were recruited primarily from the wider public sector and academia. The recruitment process took the form of an informative, face to face presentation in which the project layout was clearly described and several examples provided of what potential participants were up against. Some key demographic characteristics of the sample are provided below:

Characteristic	Value	Comment
Number of participants	314 registered 195 engaged	64% retention rate
Gender	63.6% males 36.4% females	Out of 195
Sample stratification	67.7% academia 32.3% market	The market refers to both public and private sector

Number of respondents per question	100-130	The number of respondents per question was variable, demonstrating a pick during the first 3 weeks (~160), and then regulating around 100-130
Number of questions set	14	Open for 2-6 months
Total number of responses	~2,100	Forecasts registered in the system

Table 3: Demographic Characteristics of forecasters

All participants, upon admission to the project had to create a new email address in the form of w123456@provider.com (a letter followed by 6 numbers of their choice) which would be their user name. Its purpose was to protect anonymity within the project environment, and thus limit the potential for participants to influence one another. All participants then had to take a thorough demographics quiz, before they were randomly allocated to one of the two project teams: Team A (the Control Group) and Team B (the Trained Group).

4.1.6.3 Questions, scoring and feedback

All questions used for the purposes of the present project, were carefully selected by an appointed team (selection board) comprised of the three

writers of the present paper and cautiously selected subject matter experts, depending on the topic of each question. In practical terms, the selection board had each question formulated in a very clear layout, ruling out any information that would potentially guide the forecasters towards a specific direction.

Furthermore a “Clairvoyance test” was issued for each question, which prevented potential “ex postfacto” disputes for the actual outcome of the question at hand (B. Mellers et al., 2015). Apart from the question text, several other supplementary clarifications were given to participants, primarily through the use of “neutral” resources (UN, EU, NATO, governmental entities etc.).

Below we present an example of one of the questions given (translated from Greek to English):

QUESTION:

Will the United States of America submit by May 30th 2017 an official request to withdraw from the United Nations Framework Convention on Climate Change (UNFCCC)?

CLARIFICATIONS:

According to Article 25 of the aforementioned Framework Convention, “...withdrawal shall take effect upon expiry of one year from the date of receipt by the Depositary of the notification of withdrawal...”.

Verification of withdrawal request submission will be performed through the UN's respective official site:

<https://treaties.un.org/Pages/CNs.aspx?cnTab=tab1&clang=en>.

Supplementary information concerning UNFCCC can be retrieved from the Hellenic Ministry's for Environment and Energy official website at:

<http://www.ypeka.gr/Default.aspx?tabid=226&locale=el-GR&language=en-US>

<http://www.ypeka.gr/LinkClick.aspx?fileticket=%2bZ9LThYNEvI%3d&tabid=442&language=el-GR>

Questions provided had a time horizon approximately from two to six months and all participants were instructed to provide an initial forecast within the first ten days and then update it at their own discretion, given the information flow.

The participants (forecasters), where requested to provide their estimates in the form of a point forecast, spanning from 0% to 100%, in increments of '1' (Philip Eyrikson Tetlock & Gardner, 2015). The traditional Brier Score was used in order measure performance accuracy (Brier, 1950). In simple terms the Brier Score measures the squared deviations between probabilistic

point forecasts (as expressed by the participants) and actual outcomes (as verified and coded by the selection board).

A principal characteristic of Brier scoring is that extremely erroneous forecasts are heavily penalized. The actual outcomes have a binary expression, either of “0” if the event under question did not occur, or “1” if the event took place. All participants were receiving Brier-based scores and their corresponding ranking, as follows:

Average Brier Score for each question, where a forecaster gets a score for each day a question is active, starting from the time that he actually places his first point forecast (Horowitz et al., 2016).

	Day 1	Day 2	Day 3	Day 20	Day 21	Day 41	Day 42
Actual Forecasts	-	60%			80%			99%	
Calculated Forecasts	-	60%	60%	60%	80%	80%	80%	99%	99%
Brier Scores (event occurred)	-	0.32	0.32	0.32	0.08	0.08	0.08	0.0002	0.0002

Table 4: Average Brier Score Calculation

In the above example, a forecaster who was participating to a question that was active for 42 days, placed three consecutive forecasts, on the 2nd, 20th and 41st days. In order to take into account, the moment each forecast is being placed, each forecast is being “carried along” until the moment a new one is placed. This assumption is logical given the most probable reason

a forecaster is not updating his forecast is because he actually still believes in its validity. So the score for the above question would be:

$$(20*0.32+22*0.08+2*0.0002)/41=(6.4+1.76+0.0004)/41\approx 0.199.$$

Average Net Brier points¹⁴ (NBP) for each question, while taking into account the performance of other forecasters at the time each forecast was being placed. Particularly, we average all the Brier scores per day (benchmark Brier), for all the active participants in the specific question, and we then subtract them from the forecaster's daily score (individual scores):

	Day 1	Day 2	Day 3	Day 20	Day 21	Day 41	Day 42
Actual Forecasts	-	60%			80%			99%	
Calculated Forecasts	-	60%	60%	60%	80%	80%	80%	99%	99%
Individual Scores	-	0.32	0.32	0.32	0.08	0.08	0.08	0.0002	0.0002
Benchmark Scores	1.02	0.99	0.98	0.97*	0.71	0.40	0.32*	0.11	0.001
Net Brier	0**	-0.67	-0.66	-0.65	-0.63	-0.32	-0.24	-0.1098	-0.0008

Table 5: Net Brier Points calculation

* For the sake of calculations we assume that the benchmark scores from Day 4 to Day 19 and from Day 21 to Day 40, remain the same.

** Forecasters do not receive any penalty for the days that they do not provide any forecast, but get the Benchmark Score

¹⁴ The specific Scoring method was retrieved from the Open Good Judgment tournament at: <http://training.goodjudgment.com/keepingscore/index.html>.

Given the above, NBP, ideally take values in $[-2,2]$, and being in the range $[-2,0]$, indicates that the forecaster performs better than average, with the lowest NBP indicating better performance. So, in the present example, NBP, would be calculated¹⁵ as the average of the Net Briers, in a 42 days period (NBP ≈ -0.45).

The above mentioned scores were accessible to each forecaster through the respective interface in the application being used, where they could also see their ranking comparing to all the other participants that have provided an answer to the specific question.

This approach is fully aligned with the so called 'outcome accountability' as described by (Chang, Atanasov, Patil, Mellers, & Tetlock, 2017) stating that, in the framework of a geopolitical tournament, "accountable forecasters perform better than their non-accountable counterparts, in terms of forecasting accuracy".

Additionally, it should be highlighted that the freedom provided to forecasters to provide their estimates in a pure and unconstrained numerical form (not rounded or in the form of 'bins'), is verified by a more recent analysis of the GJP results (Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2018).

The abovementioned scoring approach can be substantially impacted by extremely unforeseen events (black swans) that lead forecasters to

¹⁵ NBP= $[0+(-0.67)+(-0.66)+(-0.65)*18+(-0.63)+(-0.32)+(-0.24)*20+(-0.1098)+(-0.0008)]/42 \approx -0.45$

extremely erroneous estimations (Schoemaker & Tetlock, 2016) and thus higher brier scores (bad). Fortunately, the questions set during the present experimental procedure were not toppled by any extreme events. In any case a thorough audit of the verbal justifications provided by the forecasters served as an indispensable tool for identifying errors related to skills or chance.

4.1.6.4 Incentives

In the initial presentation given to all participants, during the recruiting process, the following incentives were communicated to them:

A “Certification for the Successful Participation in a Research Program in the Field of Judgmental Forecasting”, issued by the University of Peloponnese. The opportunity to attend, free of any a cost, a hyper-intensive preparatory course towards obtaining the Project Management Professional (PMI/PMP®) certification¹⁶, of nominal value of 650.00€.

The smaller number of questions in our research implies, when compared with other studies, that we will need a higher acceptance threshold to maintain internal consistency and comparability with those studies. Moreover,

¹⁶ Information on the specific certification can be retrieved from: www.pmi.org/certifications/types/project-management-pmp. Practically they were given the chance to obtain one of the most important industry-recognized certifications for project managers.

the presence of a lower threshold in such studies does not necessarily constrain the number of answers which may have been higher and closer to our threshold. To qualify for the incentives, participants had to make a forecast for at list 11 out of the 14 questions of the first phase. Additionally, all participants were able to track their performance in terms of personal brier scores and ranking, relative to all the other participants.

4.1.6.5 Training Design

The training design followed the principles of the Good Judgment Project (Chang et al., 2016). In practical terms, all participants, were randomly distributed into two groups:

- a) Team "A": No Analogies,
- b) Team "B": With Analogies.

The training modules per group, are presented in the below table:

TRAINING MODULE	DURATION	TEAM "A"	TEAM "B"
The world of biases	8'	√	√
De-biasing techniques	12'	√	√
Basic statistics & probabilistic reasoning	11'	√	√
Practical Bayesian thinking	9'	√	√
Techniques for forecast decomposition	12'		√
Structures analogies and their applications	7'		√
Total training duration per Team		40'	59'

Table 6: Preparatory training modules

The construction of the training modules was an arduous effort, and had to go through several revisions. Initially a small survey was performed, to a

sample of the participants (in the form of unstructured interviews), in order to derive to the optimal duration of the presentation modules with the objective to obtain maximum engagement. The above survey verified the concept of “micro-learning” (Hug, 2007; Katsagounos & Rehl, 2018) and an approximate duration of 10’ per module was selected as the rule of thumb.

Following the above duration “restrictions”, a severe content compression had to be performed in order to achieve optimal training material communication, both in terms of content completeness and comprehension. Post training interviews were performed, that verified the effectiveness of the above approach in terms of content engagement, as an outcome of content comprehension and applicability.

It is a fact that a solemnly didactic approach is not always the most fruitful, principally due to one-way communication (Chang et al., 2016; Graber, 2003). Participants were therefore receiving active feedback in the form of a Brier Score (Brier, 1950) for each question, along with their respective comparative performance ranking, relatively to the other participants. The above information, along with a continuous prompt to go through the training material before engaging to a question, was aiming to help them re-calibrate their forecasting methods and adapt towards achieving a better performance

(According to Mellers⁵⁶ training effects were identified to last to training periods (8-10-month duration per period).

A recent practical example of the effectiveness of similar to the aforementioned training within a corporate environment, can be retrieved from a recent HBP article by D. Hernandez(Hernandez, 2017), where he clearly describes the forecasting performance improvement in 'TWITCH' company (subsidiary of Amazon).

4.1.7 Results-Analysis and Hypotheses Testing

In general, the participant's engagement with the whole endeavor could be characterized as balanced and relatively anticipated. In particular, the average forecast update, per participant and per question was 1.42 [in the GJP the corresponding frequency was 1.49 (Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2015)]. Furthermore, given the experiment procedures required justification for all questions, the corresponding character count is relatively high, averaging at 363 characters per person and per question [The corresponding restriction in the GJP was that there should be at list one 50 word (~200 characters) comment throughout the year (B. Mellers et al., 2015)].

In the figure below, we chart the evolution of the aforementioned characteristics over the course of the experiment. The correlation coefficient for the above characteristics was estimated at 0.06, indicating that they are almost absolutely uncorrelated. Conversely, a closer look at the below graph

clearly indicates that they are correlated, with the correlation coefficient drastically changing after the 6th question. Particularly, the coefficient for the first 6 questions is -0.531, and for the rest is 0.317.

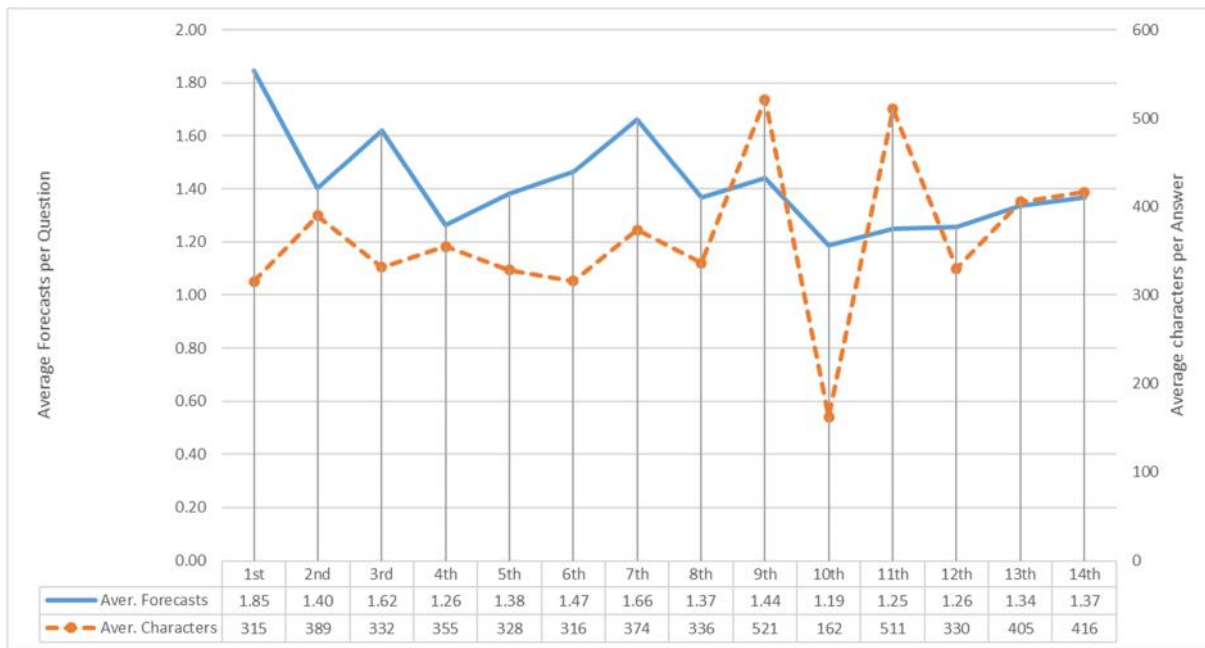


Figure 6: Time performance of forecasters

The above finding may be considered a 'maturity indicator' which reveals the critical point at which forecasters adapt to the experiment environment and react to the external signals (e.g. information flow) in a manner closer to that which was anticipated. Further research into the topic should be considered indispensable in order to verify the finding in other contexts as well.

4.1.7.1 Potential existence and early identification of Superforecasters

In order to test the validity of our 1st hypothesis, we set as a cut-off point the 6th question, which is the median of our acceptance threshold (11 out of 14 questions to be answered) and we identify¹⁷, at that point, the forecasters that belong to the top 2% (check validity of Superforecasting theory in the present experimental outline), 5%, 10% and 25% (top quartile) and see who among those remain within the specific pool, or we face randomness in performance (time tracking of performance).

In the below graph we present with the blue bars the total number of forecasters (trained and untrained) that fulfil the selection procedure (us described in the above paragraph), and with the red bars the respective percentage of the total number of forecasters in our experiment (~200).

¹⁷ The identification was performed in the following way:

- First we calculated the average brier scores per forecaster for the first 6 questions
- Then we ordered them and created the respective 'benchmark bins' per percentage (2%, 5%, 10% and 25%)
- We then identified the respective forecasters, per bin and per question.
- Finally, we detected those that were belonging in both bins for a minimum of 11 out of the 14 questions.

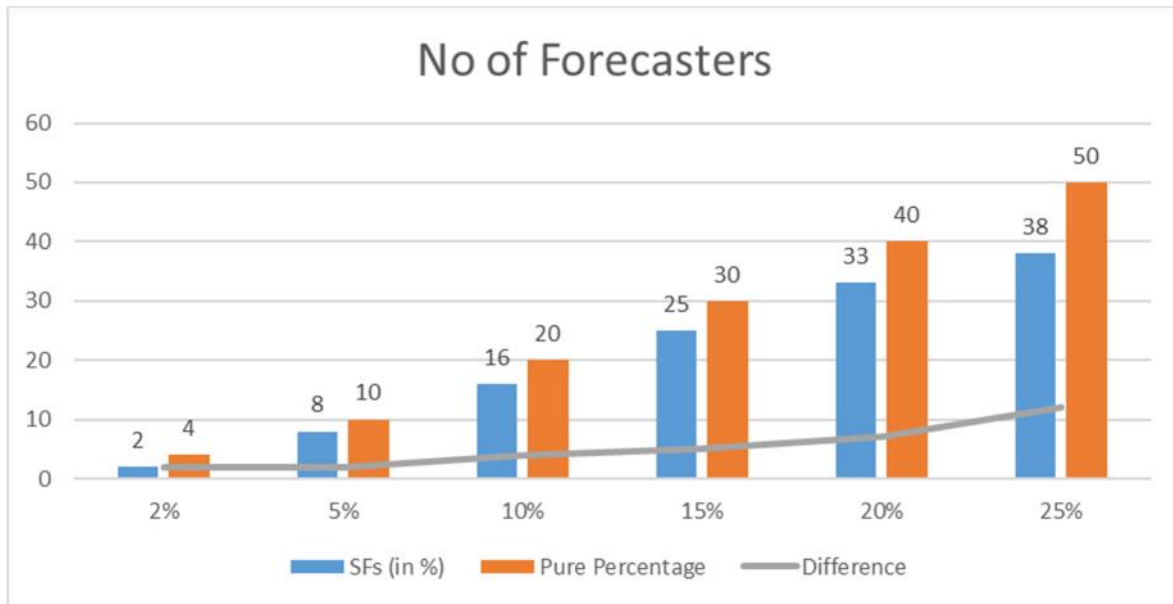


Figure 7: Superforecaster distribution per (%) bin

The graph is created for the standardized over the mean average Brier Scores (sAvg_mean). The SF count for the other standardizations is presented below:

	Avg	Net	sAvg_mean	sNET_mean
2%	2	2	2	2
5%	8	8	8	8
10%	17	17	16	17
15%	23	232	25	24
20%	31	29	33	32
25%	38	37	38	38

Table 7: Superforecaster count per standardization type

Excluding the 2% bin, where we face pure randomness (50%), in all other cases the difference appears to have settled around the 80% mark (Mean=80.4, std=0.025). In other words, the approach provides us with 80%

confidence that the selected pool of forecasters will be consistent in their performance and thus reliable enough for their estimations to be taken into account in the decision-making procedure. We can therefore firmly state that we found evidence to support the superforecasting hypothesis

Over and beyond the above indications, we wanted to identify supplementary characteristics that would serve as early warning signals when it comes to superforecaster identification. Nevertheless, the constraints under which the present experiment was conducted (time and participants) resulted in an insufficient amount of data for profound analysis and evaluation. To maximize the sample data and derive trustworthy conclusions, we conducted additional analysis of the forecasters in in the 25% bin, irrespective of the team to which they belonged, by comparing their demographic characteristics with the rest of the participants. The key profile characteristics that differentiate them from the remaining participants are presented below.

They were identified through the use of Pearson's Chi Square independence test:

Gender	The test identified significant difference in performance ($\chi^2(1)=4.808, p<0.05$) between males (24.2% in the 25% bin) and females (11.3% in the 25% bin).
Working experience	The test identified significant difference in performance ($\chi^2(6)=19.04, p<0.05$) for the various levels of experience. Particularly, from the various experience levels, the greatest contributing percentage (40% in the 25% bin)

	comes from those with 16-20 years of experience, whereas the lowest (7.8% in the 25% bin) from those with no experience (principally academia).
Knowledge of the English language	<p>The test identified significant difference in performance ($X^2(2)=7.31$, $p<0.05$) for the three levels of English language knowledge. Particularly, the greatest contributing percentage (25% in the 25% bin) comes from those with expert knowledge.</p> <p>Those with the intermediate knowledge contributed with 12.3%, whereas those with basic knowledge had no contribution at all.</p>
General frequency of information	<p>The test identified significant difference in performance ($X^2(4)=9.98$, $p<0.05$) for the various levels of information frequency. Particularly, the greatest contributing percentage (31.8% in the 25% bin) comes from those that get informed on a daily basis. The contribution was declining in an almost linear fashion: 20.9% (weekly), 14.3% (monthly), 3.6% (more scarce), 0% (never).</p>
Type of information sources	<p>The participants were requested to denote their principle sources of information. The choices provided where: (1) paper-based periodical publications, (2) internet-based periodical publications (including official websites), (3) independent websites (blogs, personal webpages etc.), (4) social media, (5) other.</p> <p>The test identified significant difference in performance ($X^2(1)=5.32$, $p<0.05$) between those having selected</p>

	<p>choice (2) (24.2% in the 25% bin) and those having not (11.3% in the 25% bin).</p> <p>Similar were the results for the 3rd source (independent websites) as well ($\chi^2(1)=9.84$, $p<0.05$), with the corresponding percentages: 26,4% (yes), 8.1% (no).</p> <p>The other sources (1, 4 & 5) did not provide significant contribution to forecasting accuracy.</p>
<p>Language of information sources</p>	<p>The participants were requested to denote the language of their sources of information (could be more than one). The choices provided where: (1) Gr, (2) En, (3) Fr, (4) De, (5) Ru, (6) Ar, (7) Other.</p> <p>The test identified significant difference in performance ($\chi^2(1)=6.87$, $p<0.05$) only between those having selected choice (2) (24.4% in the 25% bin) and those having not (8.3% in the 25% bin).</p>

Table 8: Statistical analysis of forecaster's diversification factors

We believe that the gender related finding requires some more clarification. Similar findings were identified by Shane Frederick, the father of the Cognitive Reflection Tests (Frederick, 2005), where men were receiving consistently higher scores than women. Although Frederick's tests were not appraising pure forecasting skills, but rather pure cognitive abilities, it has also been verified by B. Meller et al. (2015) that there exists a positive correlation between the two. The difference in performance during Frederick's experiments was not attributed either to biases nor to lack of attention. It was the superior skills in mathematical reasoning that were actually helping men

perform systematically better. Although our experiment's questions were not principally based on mathematics, the approach that was proposed to them in order to help them derive to a more accurate forecast, required some relevant skills, and principally probabilistic reasoning¹⁸.

4.1.7.2 The contribution of training in forecasting performance

Our second question challenges the impact of specialized training. In order to test the validity of our findings, we standardized all scores (both Average Briers and NBPs) as a function of their deviation from the mean for each question (Chang et al., 2016), using the formula:

$$\text{Stand_Value} = (x - \text{mean}) / \text{SD}$$

Below, we provide a visual analysis of the comparative performance of the two teams, Team A (Control Group) and Team B (Trained Group), for both the simple and standardized metrics:

- Analysis for Average Brier Scores (Avg)

¹⁸ They had to define base rates, aggregate probabilities, update their forecast following the Bayesian way etc.

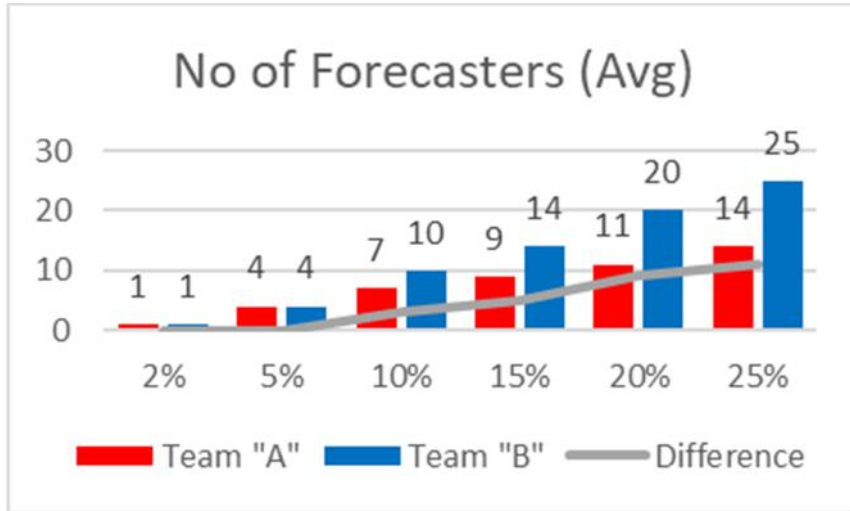


Figure 8: Number of forecasters per (%) bin for Avg scores

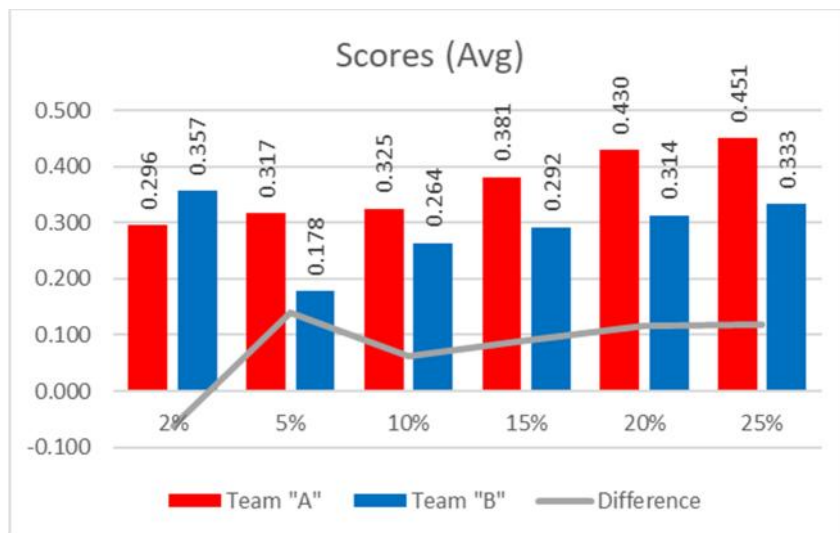


Figure 9:AVG scores per (%) bin

- Net Brier Points (Net)

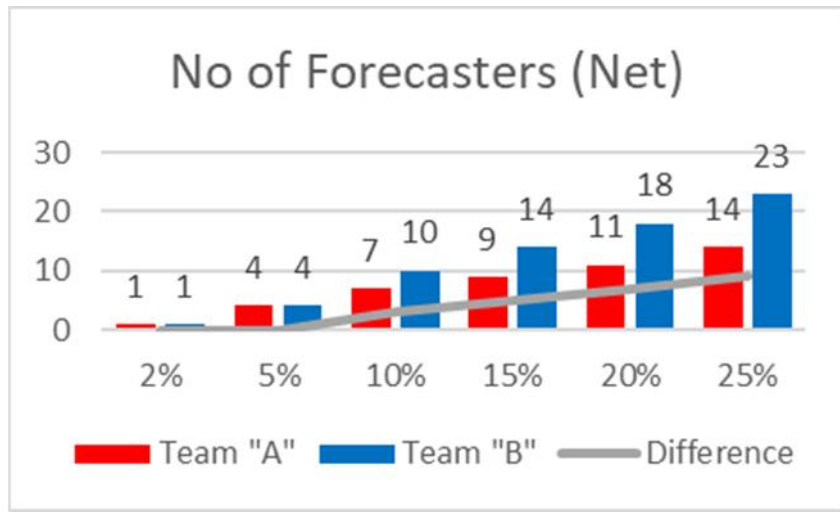


Figure 10: Number of forecasters per (%) bin for Net scores

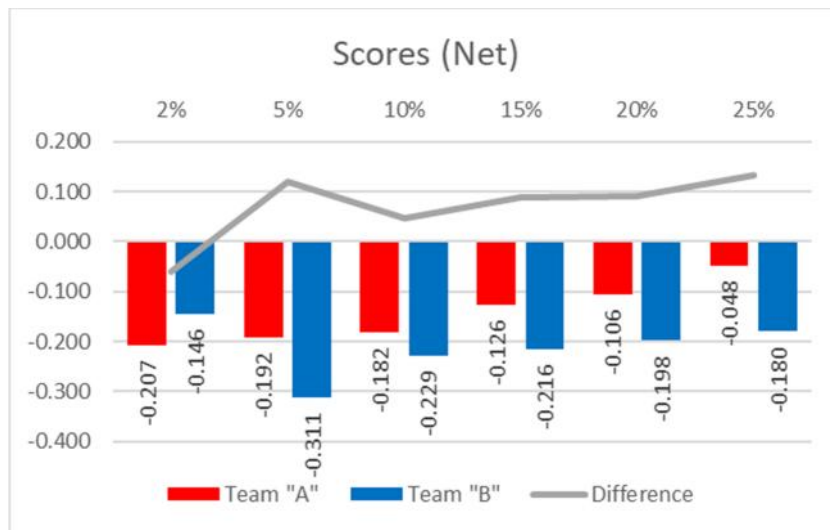


Figure 11: Net scores per (%) bin

- Standardized over the Mean Average Brier Scores (sAvg_mean)

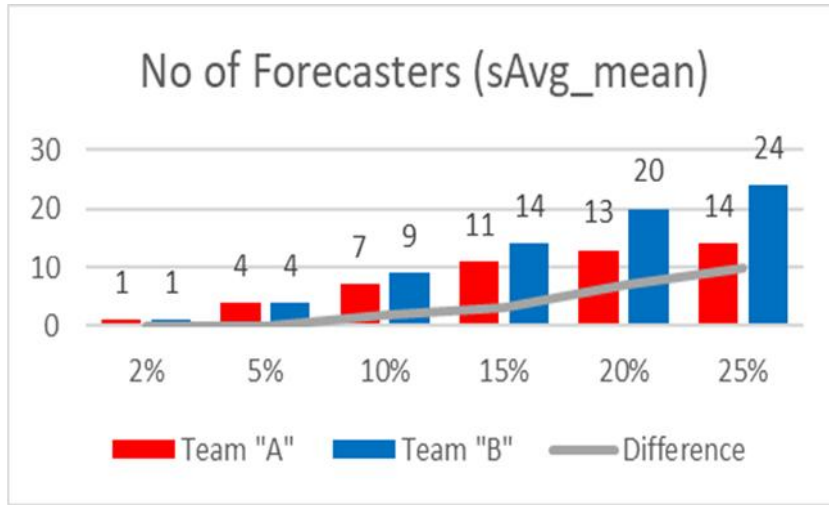


Figure 12: Number of forecasters per (%) bin for standardized over the mean average brier scores

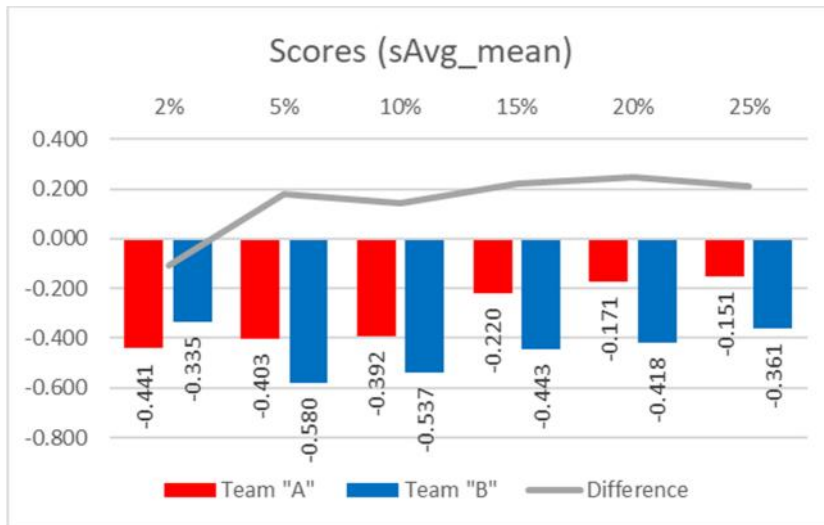


Figure 13: Standardized over the mean average brier scores per (%) bin

- Standardized over the Mean Net Brier Points (sNET_mean)

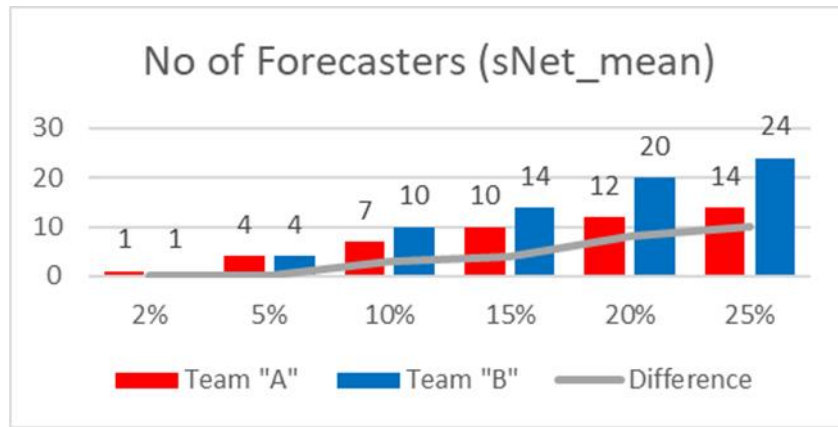


Figure 14: Number of forecasters per (%) bin for Standardized over the Mean Net Brier Points

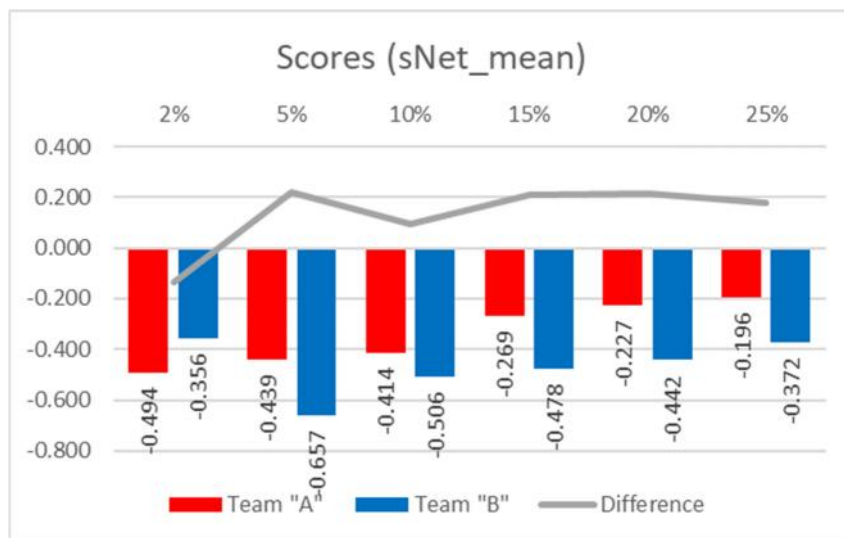


Figure 15: Standardized over the Mean Net Brier Points per (%) bins

The 1st graph in each of the above pairs provides the information concerning the number of forecasters per team & per percentage, denote that the trained group (Team "B") has more systematic "performers", and the

difference between the two groups increases as we increase the respective percentage. The fact that the aforementioned difference changes only above the 5% threshold, should be considered logical and anticipated, given the duration of the experiment and the number of participants. In particular, more systematic research is required, through the use of greater samples, which will identify the critical threshold, where the 2% principle (as per the GJP) starts to apply.

The analysis of the “scores per percentage” graphs on the right, should only be performed while keeping under consideration the number of participants per bin (percentage). In particular, the 2% bin, contains only one forecaster per team thus the provided statistics can't be considered as representative for deriving to solid conclusions. From that point onwards, the number of forecasters per bin, gradually increases and the provided statistics are more substantial.

The conclusion from the score graphs is that the consistent trained forecasters (Team “B”), outperform the respective untrained ones (Team “A”) in terms of absolute scores. Indicatively, the evolution of the percentage of improvement is projected below (for the standardized scores):

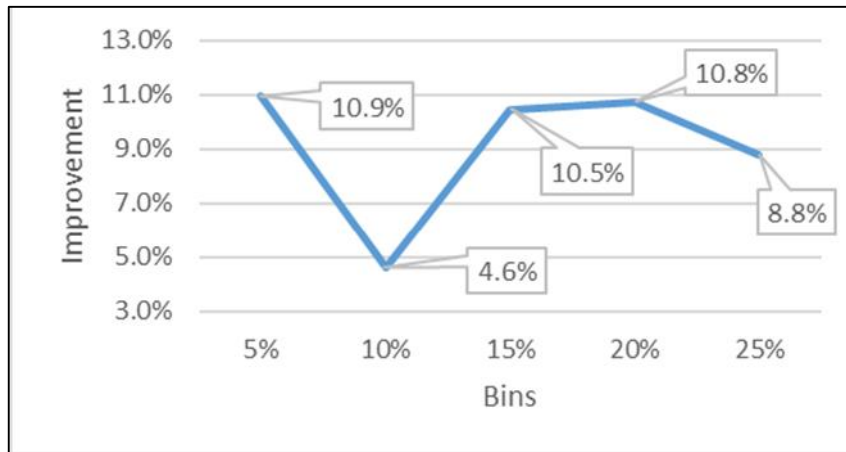


Figure 16: Improvement of Team 'B' over Team 'A'

We could say that the above finding confirms the respective one from the GJP in terms of training contribution to forecasting performance.

It should be noted that, on retrospective, once having identified the forecasters included in the above percentages (bins), an ID per ID cross check indicated that they have all answered all the posed questions (14/14). This is, in a way, in pure contradiction with the 'CHAMPS KNOW' principle (Chang et al., 2017, 2016), and in particular the 'S' one, prompting participants to answer only questions with reasonable pay-off.

Additional confirmation on the non-validity of the 2nd Hypothesis is provided through the below descriptive statistics and respective box-plots. This allows us to identify the superiority of the Structured approach, both visually and numerically:

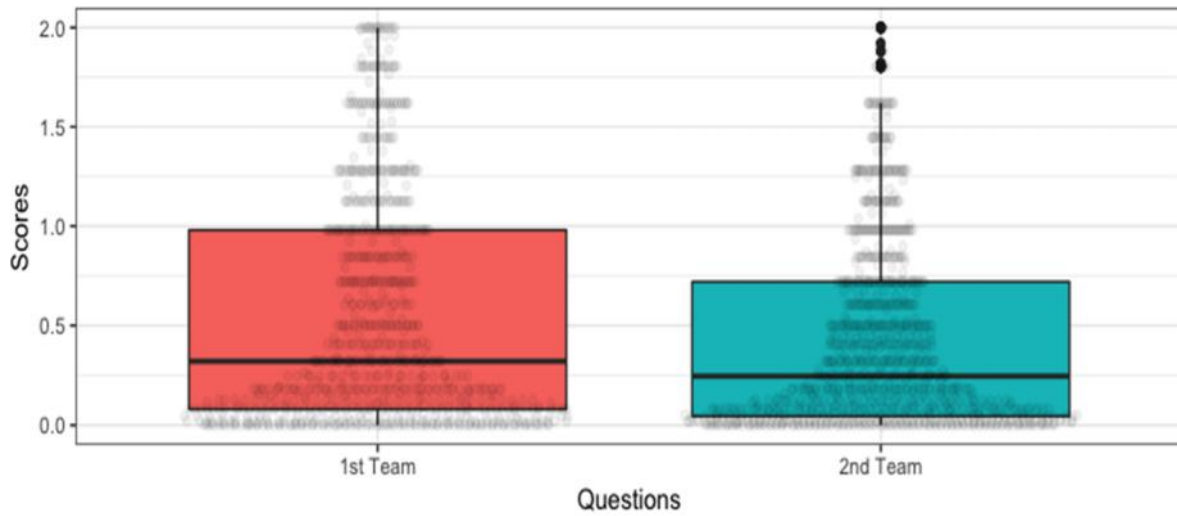


Figure 17: Average Brier Score per team boxplot

Perf measure	Team "A"	Team "B"
min	0	0
max	2	2
median	0.32	0.245
mean	0.57486	0.43979
SE.mean	0.02042	0.01505
CI.mean.0.95	0.04008	0.02954
var	0.34779	0.22512
std.dev	0.58974	0.47447
coef.var	1.02588	1.07885

Table 9: Descriptive statistics for average Brier scores per team

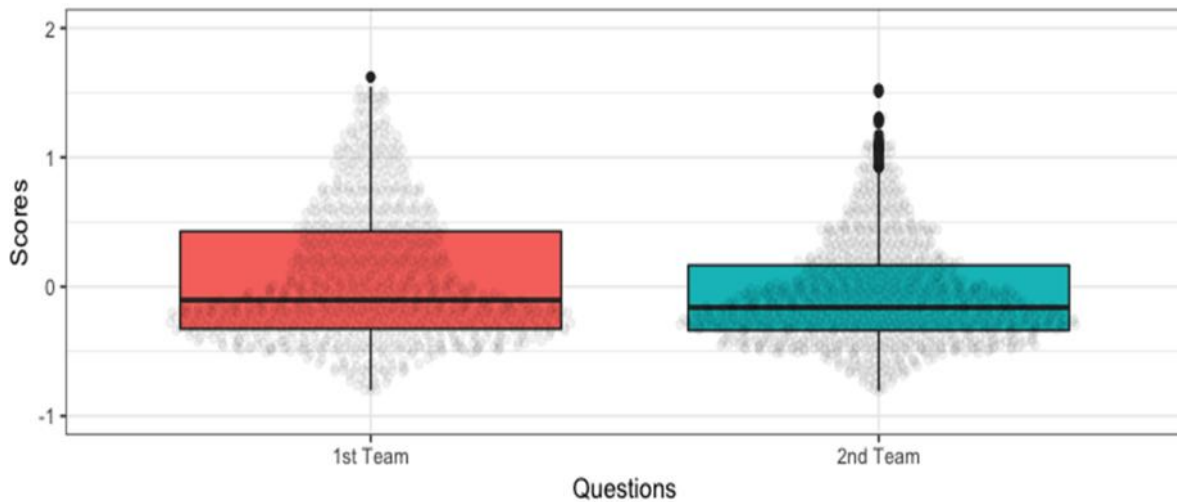


Figure 18: Net Brier scores per team Boxplots

Perf Measure	Team "A"	Team "B"
min	-0.80137	-0.80637
max	1.62154	1.52596
median	-0.10366	-0.16024
mean	0.06936	-0.05859
SE.mean	0.01797	0.01286
CI.mean.0.95	0.03527	0.02525
var	0.26940	0.16441
std.dev	0.51904	0.40548
coef.var	7.48297	-6.92044

Table 10: Descriptive statistics for Net Brier scores per team

The superiority of the 2nd Team is apparent and under the standardized values, showing thus the robustness of the results.

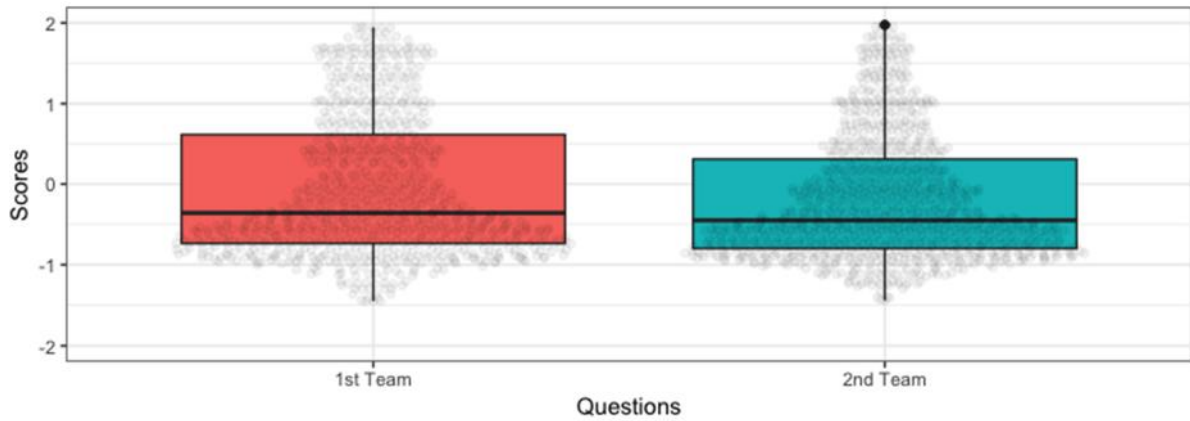


Figure 19: Standardized over the mean Average Brier Scores per Teams Boxplots

Perf Measure	Team "A"	Team "B"
min	-1.45066	-1.44102
max	4.62076	4.62076
median	-0.25153	-0.43725
mean	0.14344	-0.12019
SE.mean	0.03826	0.02790
CI.mean.0.95	0.07511	0.05475
var	1.22131	0.77320
std.dev	1.10512	0.87932
coef.var	7.70441	-7.31561

Table 11: Descriptive statistics for Standardized over the mean Average Brier Scores per Teams

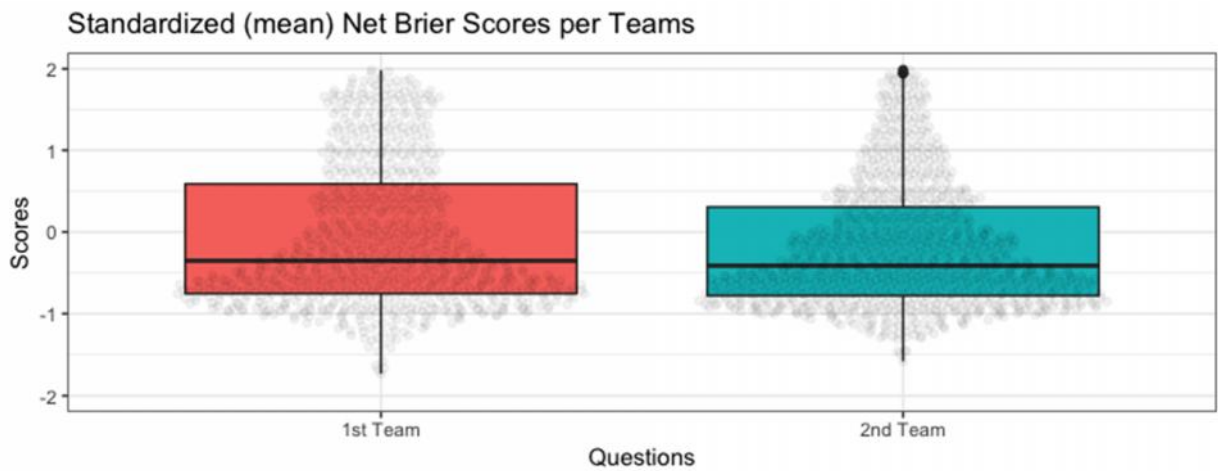


Figure 20: Standardized over the mean Net Brier Scores per Teams Boxplots

Perf Measure	Team "A"	Team "B"
min	-1.73115	-1.57997
max	4.84424	4.55872
median	-0.25138	-0.37948
mean	0.14376	-0.12101
SE.mean	0.03879	0.02738
CI.mean.0.95	0.07614	0.05374
var	1.25503	0.74472
std.dev	1.120282	0.86297
coef.var	7.79250	-7.13125

Table 12: Descriptive statistics for Standardized over the mean Net Brier Scores per Teams

Aiming to further test the robustness of our results, we have performed the following supplementary standardization methods as well, and in all of them the supremacy of Team's "B" performance is being verified:

- Over Median (IQR): $\text{Stand_Value} = (x - \text{median}) / \text{IQR}$
- Over Median (MAD): $\text{Stand_Value} = (x - \text{median}) / \text{MAD}$

Relevant consistency was also apparent on a question by question basis (with minor exceptions in 1st, 9th and 11th questions), as presented in the following boxplots (see **Appendix “A”** for detailed statistics per question.)

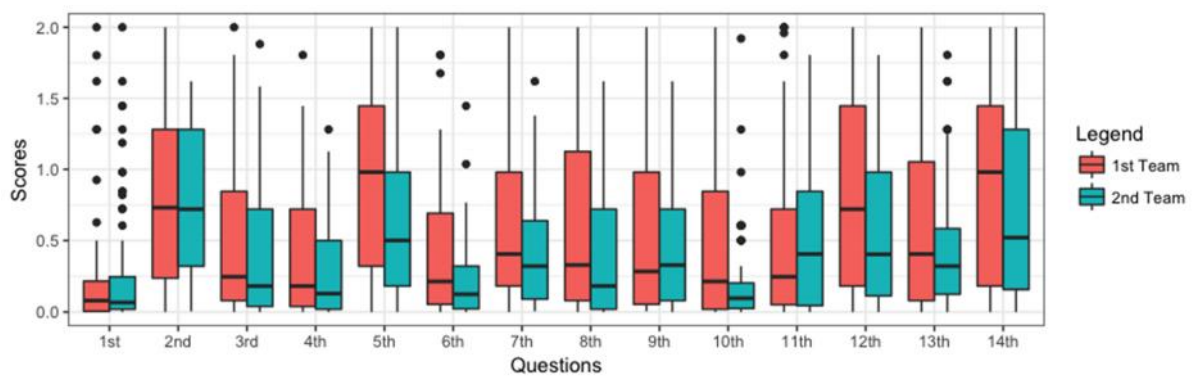


Figure 21: Average Brier Scores per Question and Teams Boxplots

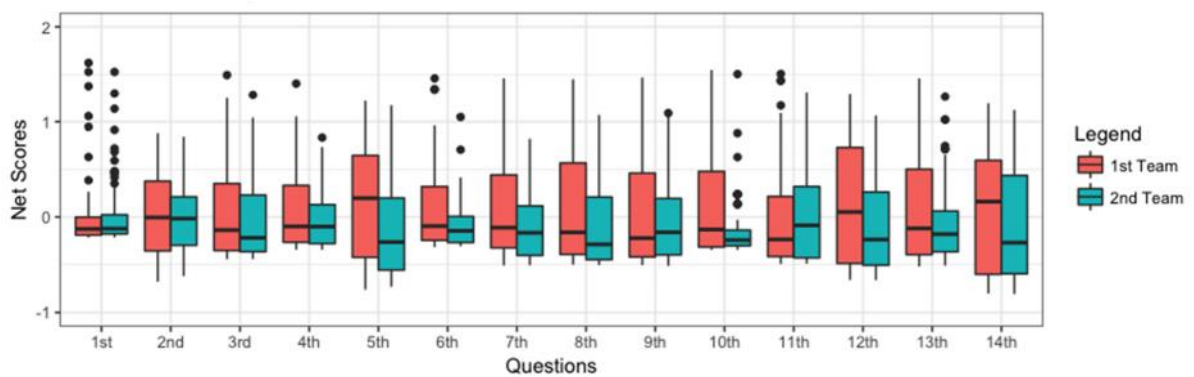


Figure 22: Net Brier Scores per Question and Teams Boxplots

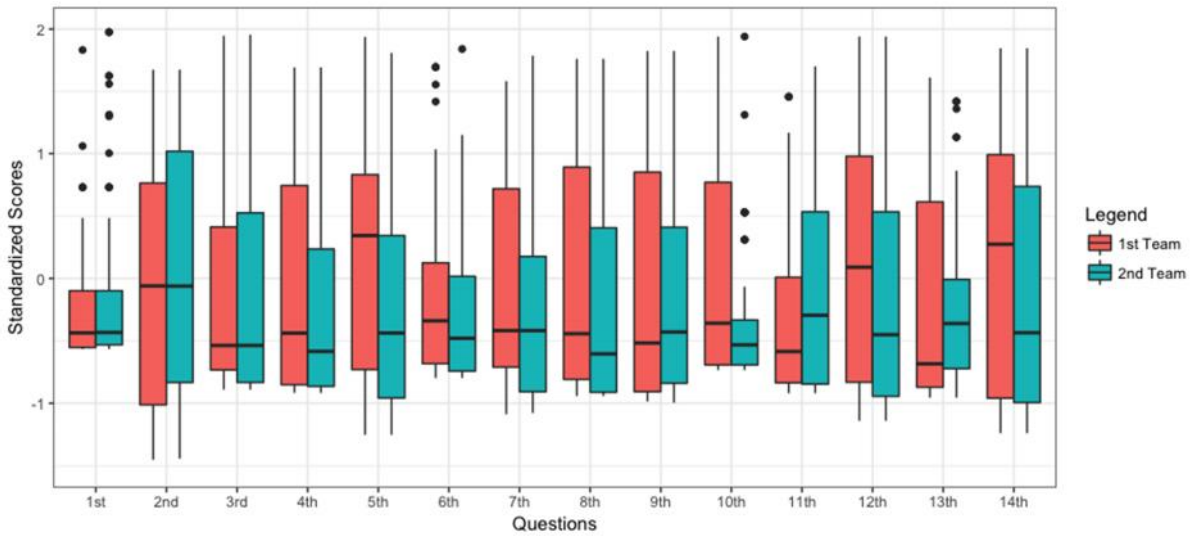


Figure 23: Standardized over the mean Average Brier Scores per Question and Teams

Boxplots

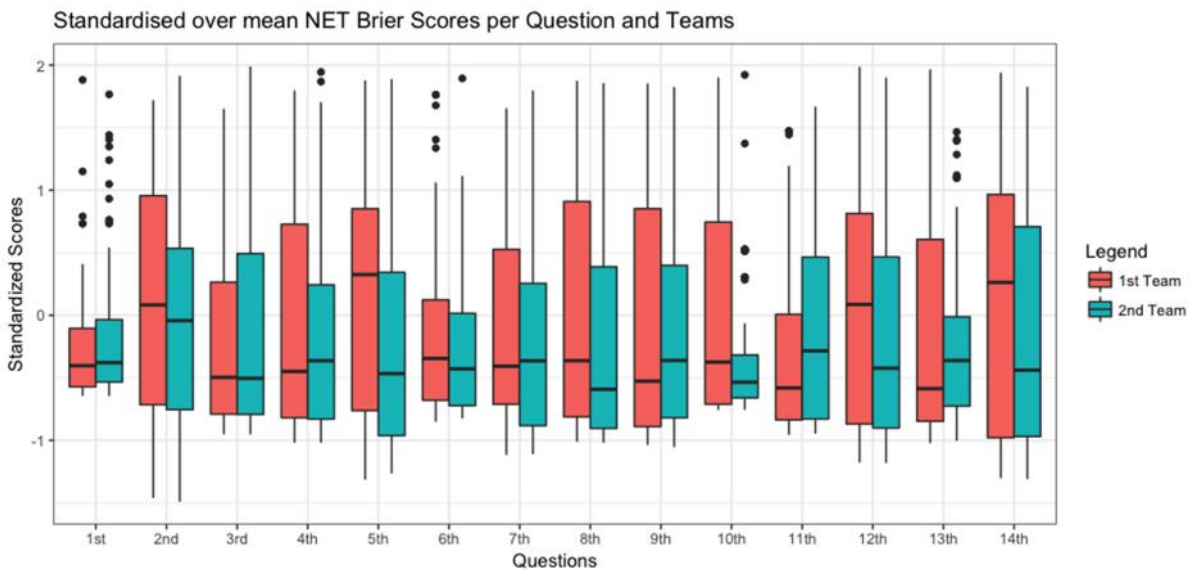


Figure 24: Standardized over the mean Average Net Scores per Question and Teams

Boxplots

Depending on the standardization method, there appears to be a small volatility in the interpretation of the results, particularly for the 2nd and 3rd

questions. This clearly depicts that a single metric can, under circumstances, be misleading. The present, “multidimensional” analysis, eradicates potential errors deriving from single-metric focus (e.g. one type of standardization).

4.1.7.3 Stochastic Dominance Tests

Our analysis of the results utilizes as well, beyond standard measurement concepts of judgmental forecasting, the methodology of stochastic dominance (SD) to evaluate the performance of participants -- for the first time to the best of our knowledge in this strand of the literature. The use of the SD concept is important for two reasons: first, it allows a complete/better view of outperformance compared to more traditional statistics and, second, it provides compelling visuals for the produced results.

SD (Hadar & Russell, 1969) performs a partial ordering between random variables (prospects) for a broad class of decision makers thus revealing the superior prospect. In order to reveal the outperformance of analogies trained teams over the control group, we will use 1st, 2nd and 3rd order SD (Hanoch & Levy, 1969; Whitmore, 1970). The only difference in our approach is that we do not seek to “maximize the profits”, in terms of actual values, but to minimize them, given the lower the Brier Scores, the better the outcomes.

In order to produce and analyze the SD statistics, we perform to our samples (teams per question) two consecutive Bootstrap re-samplings as follows:

- Block Bootstraps (100 iterations per Team) in order to create new samples with equal numbers of entries.
- Bootstraps for the estimation of P-values, for each of the above produced blocks (20 iterations per block).

Given the p-values for all the comparative statistics presented in the table below, Team B outperforms Team A (the 0.05 threshold for the p-values stands for all levels of SD, underscoring the pure outperformance of Team B).

	1SD	2SD	3SD
Avg	0.000073	0.000576	0.001128
sAvg.mean	0.000306	0.000000	0.000000
sAvg.median.IQR	0.000024	0.000257	0.000012
sAvg.medianMAD	0.000711	0.009276	0.000085
Net	0.000294	0.000000	0.000000
sNet.mean	0.000036	0.000000	0.000000
sNet.medianIQR	0.000073	0.001374	0.000306
sNet.medianMAD	0.000159	0.008871	0.000024

Table 13: 1st, 2nd and 3rd order Stochastic Dominance tests

The below graph depicts the ECDFs deriving from the above analysis. Our interpretation is inverse to the original one, given the lower the achieved scores, the greatest the "gains". E.g., the original interpretation would have been "*the red cumulative (first-order) stochastically dominates the blue*

cumulative. When this is true, anyone who prefers larger prizes to smaller ones, will prefer the red cumulative." Thus, aiming for the lowest scores, we can clearly argue that the 2nd Team's performance, stochastically dominates the 1st Team's performance.

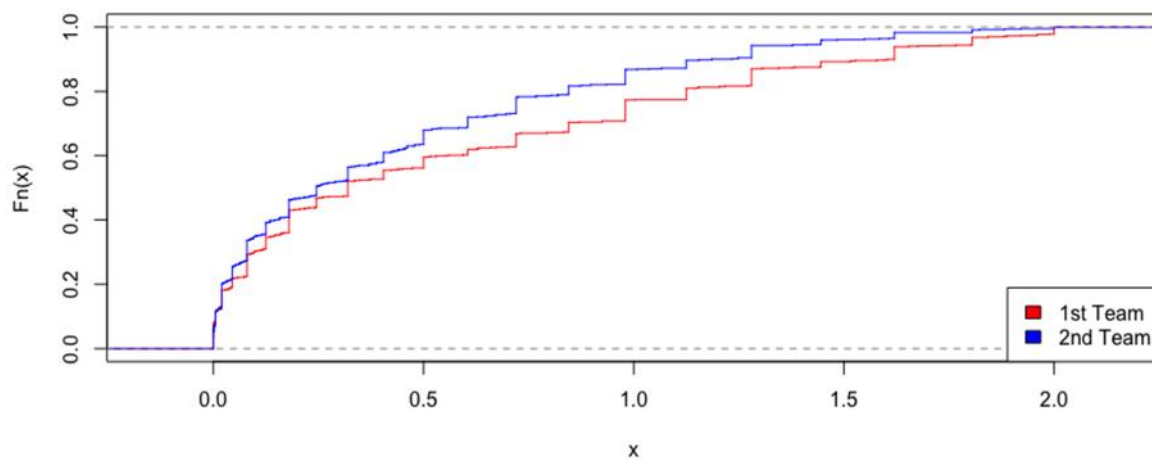


Figure 25: ECDF plots for 1st order Stochastic Dominance Test

4.1.8 General Discussion

The present research aimed at replicating the conditions in a corporate environment where limited resources and time would have been the driving constraints. Furthermore, we avoided the extremely strict experiment environment, given that we knew that similar conditions would not be able to be replicated within the framework of a company or an organization, due to the intrinsic inability to contain and restrict employs. The only restrictive measures that we took, where the following:

- a) Fully anonymized participation in order to avoid interactions and influence.
- b) Password protected training sessions, to make sure that only the designated participants had access to the relevant training modules.

The principal difficulty that we faced throughout the experimental procedure was to keep the drop-out rate at the lowest possible level. Given the limited visibility of our experiment (particularly when comparing it with the GJP/IARPA) we've had to counterbalance their will to discontinue their engagement with relatively high incentives (of nominal value, approximately 650.00€). We thus assume that, in order to have similar outcomes within a corporate environment, a typical reward scheme should be established.

Additionally, in terms of experimental structure, we avoided researching on topics already covered and resolved by the GJP. Particularly, we provided common trainings to both groups, covering the topics of de-biasing and probabilistic reasoning. This approach, although had a negative effect in terms of comparability of the results to those of similar other researches, it gave us the chance to clearly focus on the contribution of the structured approach in the provision of forecasts.

As already shown earlier, we do believe that there exists a turning point in the performance of the participants (irrespective to the group they

pertained to). Thus, there appears to exist a learning curve which in our case seemed to expand until the 6th question (approximately 1,5 month). This probably indicates that the forecasters started interpreting the various 'information signals' more consistently, and their elaborations were following the pace of the forecast updates. Nevertheless, we do believe that there exists ground for further research on this topic in order to identify those key performance indicators (KPIs) that could serve as 'early warning signals' for the forecasting maturity of the participants.

Another critical finding of the present research is that the 2% principle (in terms of forecaster identification) does not seem to apply in so small samples (~200). We've nevertheless witnessed some indications of constant positive performance around the 5% threshold and above. We thus believe that there should be further investigations – research in terms of identifying the critical percentage that is applicable for the various sample sizes.

As mentioned above, in the presented research we've avoided the replication and cross testing of findings already established and in good standing. So instead of testing upon dispositional, situational and behavioural variables (Barbara Mellers et al., 2015) we've focused on some key demographic characteristics and retrieved some rather interesting results. Given the analysis is clearly provided in par. 4.2, we will only focus on one finding, and that is the English language skills. It appears to be a huge impact of the language being used when retrieving information. The majority of

information in the WWW is provided in the English language thus forecasters are, in a way, enforced to adapt. Consequently, English language skills can be considered a key asset when it comes to information collection. We believe that further research should be conducted on the topic in order to identify the contribution of language skills to the various types of forecasting questions (in terms of context).

Finally, we should highlight for one more time the fact that all our key performers had provided answers for all questions thus defied the risk of a potential loss, without significant negative impact on their ranking.

The study highlights the importance of using tournaments for the identification of top forecasters, but in our case, while defying sample size. At the same time, it isolates and identifies the effectiveness of one of the oldest and verified forecasting approaches, that of analogies.

4.2 Early vs late forecasting: Do forecasting tournaments help us identify a time related performance of forecasters?

4.2.1 General

In this Working Paper (to be published), we analyze the effectiveness of the method on the performance of the forecasters at the early stages (1st ten

days) of each forecasting question. To achieve that, we use the established experimental structure and we perform a comparative analysis of the forecasters' performance during the 1st ten days of each question. In order to achieve that, we have instructed all participating forecasters to imperatively submit a forecast in the aforementioned timeframe and proceed to subsequent forecasts, at their own discretion, while taking under consideration the evolution of the situation and the respective information flow. B. Mellers et al. (2015) have performed a similar approach but focused primarily on the early forecasts provided during the 1st day a new question was introduced to the tournament. Their reasoning behind this was to capture forecasts that were made quickly without profound research. In our experiment we've expanded the timeframe of the initial forecast to 10 days, aiming to detect superforecasting competences deriving from the analysis of the available information. In both cases, the hypotheses were verified: Superforecasters prevail both in 1st day forecasts (B. Mellers et al., 2015), but also in 1st ten days forecasts.

After having analyzed the collected results, we can clearly state that the participating forecasters that were following the structured analogies approach were almost constantly providing more accurate forecasts than the control sample.

The above finding is considered of great importance for the decision-making process. In other words, decision makers can feel more confident in

justifying their decisions, when the supporting information flow derives from forecasts that were elicited through the use of a structured approach, and in particular, the proposed modified version of structured analogies.

4.2.2 Hypothesis

We list our main hypothesis in the form of a traditional “null hypothesis” with the aim of examining, on the basis of our data, whether it should be rejected or not. Respectfully our hypothesis is stated as follows:

H1: Under the strict constraint of sample size (number of forecasters) and time (duration of experiment) structured superforecasting does not aid in identifying forecasters with superior early performance.

4.2.3 Project design

The present research was performed in the framework of the experiment described in Section [8.1.6](#).

4.2.4 Results-Analysis and Hypotheses Testing

In order to counterbalance the experiment's resource constraints, all participants were instructed to provide an initial forecast within the first ten days that the question was set, and then update it at their own discretion,

given the information flow. Keeping in mind that the present approach tests the feasibility of applying the GJP concept in SMEs, an exploitation of all available resources was considered as indispensable.

Given the average forecast update, which was estimated at 1.42 per question per person, we would anticipate to have similar results, in terms of performance.

The above assumption is verified by the below descriptive statistics:

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.32	-0.036	-0.294	0.024	0.043	-0.138	0.029	0.037
mean	0.596	0.034	0.127	0.336	0.656	0.103	0.316	0.418
SE.mean	0.022	0.014	0.041	0.034	0.069	0.043	0.055	0.072
CI.mean.0.95	0.043	0.027	0.081	0.068	0.136	0.085	0.109	0.141
var	0.344	0.131	1.191	0.836	3.361	1.326	2.163	3.653
std.dev	0.587	0.361	1.091	0.915	1.833	1.151	1.471	1.911
coef.var	0.984	10.512	8.589	2.719	2.796	11.162	4.649	4.577
Team "B"								
median	0.32	-0.053	-0.386	0	0	-0.177	-0.016	-0.021
mean	0.466	-0.032	-0.115	0.167	0.334	-0.094	0.096	0.124
SE.mean	0.017	0.01	0.032	0.027	0.051	0.029	0.037	0.048
CI.mean.0.95	0.034	0.019	0.062	0.052	0.101	0.058	0.073	0.095
var	0.231	0.07	0.783	0.551	2.037	0.671	1.079	1.811
std.dev	0.48	0.265	0.885	0.742	1.427	0.819	1.039	1.346
coef.var	1.03	-8.399	-7.677	4.434	4.273	-8.753	10.865	10.889

Table 14: Cross team descriptive statistics

The provided descriptive statistics include all standardization methods used, and verify once again that there exists a small variability in the

comparative performance depending the metric being used. This leads us to the conclusion that the cross-tabulation of all metrics is considered as indispensable in order to avoid 'cherry picking' approaches, where one selects the most favourable metric in order to justify a statement.

The above statistics are also verified graphically through the below provided boxplots (indicative for sAvg_mean):

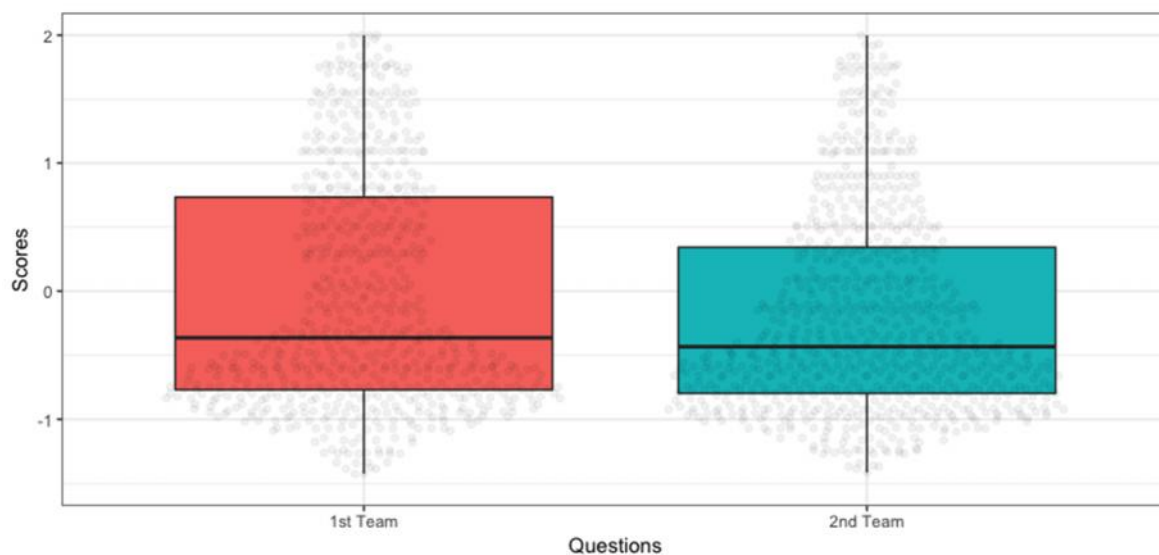


Figure 26: Standardized over the Mean Average Brier Scores per Teams Boxplots

The visual representation of the results, provides us with a valuable information. The additional training on analogies and forecast decomposition that the 2nd Team (B) received, had minimum to non-impact (depending on the metric being used) to the top performers (in comparison with Team A), but a significant impact on those bringing up the rear.

In similar vein to the findings described in Section [8.1.7.3](#), the hypothesis is verified stochastically as well with minor variation for the corresponding one conducted for the entire experiment:

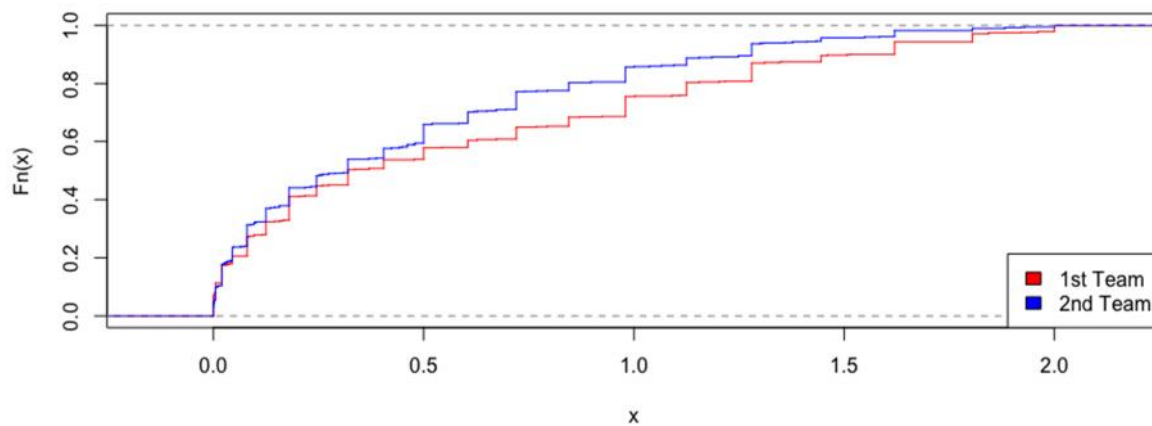


Figure 27: ECDF plots for 1st order Stochastic Dominance Test

4.2.5 General Discussion

Unfortunately, the value of the above findings can only be exploited on retrospective. This actually means that we first have to invest time to identify the competent forecasters, and then start using their early forecasts. The duration of this 'grace' period, depends on the forecasting horizon of each question being asked. E.g. in the case where a question requests for a forecast 6 months ahead, that means that we will have to wait until the closing of the question in order to be able to score the early forecasts.

In our experiment the frequency of new questions was one every week, and their average horizon span was 4-6 months. That means that we would

have to wait for approximately 6,5 months to receive sufficient data in order to judge the performance of the participants.

The present research does not provide an answer on the impact of the volatility in duration of each question and of their density on the forecasting performance. E.g. replicability –reproducibility of the results cannot be taken for granted if we were to place 6 forecasting questions, on a 6-day period, asking for results one month ahead. It is assumed that different organizations will have different needs in terms of how far in the future they would like to look with their forecasting questions. That would mean that there is a need to reproduce an experimental cycle, similar to the present one, in order to test and validate their hypothesis. Unfortunately, sometimes tradeoffs will be indispensable.

4.3 “PESCO - PM² - ESDC”: Could E-Learning Bring Closer Together EU’s Success Stories?

4.3.1 General

The present paper covers a focal point of the overall research, which is the training provision to the potential subject matter experts. It is a fact that training should be tailored to the addressed training audience and able to

cover the pre-identified learning outcomes. We use as an example the need to train member states representatives, dealing with the Permanent Structure Cooperation projects (PESCO), with the European Commission's new Project Management Methodology, namely, the PM² Methodology.

We propose various training approaches, all based on the profiles and needs of the trainees, and spanning from F2F to e-mentoring, to Synchronous and Asynchronous e-training.

4.3.2 Abstract

Security and defense are among the main topics currently being discussed at EU level. The starting point was the publication of the EU Global Strategy in 2016, which led to the development of several new instruments. Some of them, for example permanent structured cooperation (PESCO), had already existed for a decade (Treaty of Lisbon, 2009) but were waiting for the right moment to be implemented. PESCO can be considered the biggest project for European security to date, which brings together the EU Member States, the European Commission, the Council of the European Union and the European External Action Service, the main actors when it comes to the Union's foreign and security policy.

The authors argue that the Open PM² methodology should be used to manage PESCO. The necessary education and training for project managers and team leaders should be offered by the European Security and Defence

College (ESDC) and its 140 network partners, under the auspices of the Open PM² Centre of Excellence, by using the well-established and widely recognized e-learning management system of the number one CSDP training provider. The article provides concrete solutions and a detailed training needs analysis.

Keywords: PESCO; PM²; ESDC; E-Learning; Organizational Performance; LMS

4.3.3 Introduction

The European Union's (EU) security environment has faced serious challenges over the past few years. Most of these challenges were not foreseen at the time of its establishment and are starting to pose severe threats. Some examples are the numerous conflicts in the EU's neighbouring countries, several terrorist attacks on EU territory (US Department of State, 2018), the constant migration flow and finally, the turbulence and uncertainty which the Trump administration has caused concerning the contribution of EU Member States to the NATO mechanism (Trump, 2017).

4.3.3.1 Historical Background

Cooperation and integration were always at the core of the developments leading to the European Union in 1992, and from the very beginning, security and defence played a crucial role in preventing any future

armed conflict between European countries. While they were concluding the Treaty establishing the European Coal and Steel Community in 1951, the founding fathers were also discussing a European Defence Community (EDC, 1950). The plan on European defence was rejected by the French parliament; instead, a Western European Union was created as the European pillar of NATO. Nevertheless, security and defence remained within the long-term objective of an 'ever closer union'.

After several decades of dormant existence, the Western European Union was awakened through the Treaty of Maastricht (EUR-Lex, 1992), which established the European Union in 1992. The role of the WEU was 'to elaborate and implement decisions and actions of the [European] Union which have defence implications'. Only seven years later and on the basis of lessons learned during the disintegration process of Yugoslavia, the EU Member States decided to establish a European Security and Defence Policy for civilian missions and military operations abroad, including the development of civilian and military capabilities.

Another 10 years later, in 2009, the Treaty of Lisbon (EUR-Lex, 2009) introduced both a mutual assistance clause (Article 42(7) TEU) and a solidarity clause (Article 222 TFEU) into the Treaties of the European Union, which again brought cooperation and integration in the field of security and defence into the spotlight. With these two articles, security forces received a mandate to become active on the territory of the European Union in the event of (a) armed

aggression (Article 42(7) TEU), (b) natural or man-made disaster and (c) terrorism (both Article 222 TFEU), with the latter including both prevention and consequence management.

A European Council in 2013 under the slogan 'defence matters' gave strategic guidance on the work to be undertaken in the coming years. In 2015, after a series of terrorist attacks in Europe, France asked for the support of all EU Member States and in response Article 42(7) TEU was activated. The discussion again showed that a response by all EU Member States together is stronger than one by a nation state alone. The publication of the EU's Global Strategy in 2016 (Mogherini, 2016), which emphasised the need for greater security and defence for European citizens, paved the way for the implementation of permanent structured cooperation (PESCO) supported by a European Defence Fund (EDF) and a coordinated annual review on defence (CARD), as already agreed on in 2009 in the Treaty of Lisbon.

4.3.3.2 Entry into the PESCO era

Permanent structured cooperation has its legal basis in Article 42(6) of the Treaty of Lisbon TEU (EUR-Lex, 2009), which states that:

'those Member States whose military capabilities fulfil higher criteria and which have made more binding commitments to one another in this area

with a view to the most demanding missions shall establish permanent structured cooperation within the Union framework.'

Details for this cooperation model were laid down in Protocol No 10 on permanent structured cooperation established by Article 42 of the Treaty on European Union (see Official Journal of the European Union: [Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union](#), C 115, 9 May 2008). But it took another 10 years until the time was ripe for implementation. The words of the High Representative/Vice-President (HR/VP) Federica Mogherini, on 12 December 2017, are considered a landmark (EEAS, 2017):

'We did it. In the most ambitious and inclusive manner, with 25 Member States, we launched PESCO together. The 25 have taken binding commitments to improving their cooperation, and we will start with a first set of 17 very concrete projects spanning from common military training, to providing medical support to our operations. The possibilities of PESCO are immense.'

The difference between PESCO and former forms of cooperation is the fact that it is fortified by a binding commitment clause for all participating members. Nevertheless, Member States retain their sovereignty since they have the right to opt-out upon notification (or be suspended for underperformance). The binding nature of PESCO commitments is reinforced by the annual regular assessment that will be conducted by the High

Representative of the Union for Foreign Affairs and Security Policy (Council of the European Union, 2017).

PESCO has a two-layer structure, one at Council and one at project level. The former serves as a policy and decision making authority and is responsible for putting in place an assessment mechanism to track Member States' performance on the assumed commitments. The latter is at Member State level and deals with the management of the approved and assigned projects. Furthermore, and in order to facilitate functionality, a PESCO secretariat will be set up, comprised of personnel from the European Union Military Staff (EUMS) and the European Defence Agency (EDA).

The first 17 collaborative projects, which have already been identified and acknowledged by the 25 participating Member States, are based in the area of capability development and range from the establishment of a European Medical Command, an EU Training Mission Competence Centre and Cyber Rapid Response Teams and Mutual Assistance in Cyber Security, to military disaster relief and an upgrade of maritime surveillance (for further details see Appendix D. In support of the above projects, several other mechanisms have been put in place, namely:

- The Coordinated Annual Review on Defence (CARD) *'to develop, on a voluntary basis, a more structured way to deliver identified capabilities*

based on greater transparency, political visibility and commitment from Member States' (EDA, 2017)



Figure 28: CARD formation approach

- The European Defence Fund (EDF) 'to provide financial incentives to foster defence cooperation from research to the development phase of capabilities including prototypes'. (EEAS, 2017)



Figure 29: EDF breakdown

The participating Member States are required to develop national implementation plans (NIP), in which they must describe their approaches

towards achieving the commonly set goals. The approved NIPs will be communicated to all Member States and will constitute the baseline for performance measurement.

4.3.4 Is there a link between PESCO and a structured project management approach?

There is no clear answer to this question. When reading all the relevant (unclassified) documentation concerning the projects under the PESCO umbrella, the general impression is that there is a lot of wishful thinking but very few concrete steps towards a structured project management approach. In particular the EEAS's PESCO factsheet (EEAS, 2017) states that:

'...the general rules for project management are to be developed at overarching level'.

By analysing the above sentence, three obstacles for a structured management approach can be observed:

- General: No precise guidance will be provided.
- Are to be developed: The schedule is somewhat vague.
- Overarching: No reference to lower management levels.

Meanwhile, all Member States are under strict pressure to deliver their NIPs, within which they will be describing their high level commitments, not just

towards PESCO in general, but particularly towards each project in which they take part, either as leading or participating nations. Those high level commitments include high level budgets and schedules which are obligatory and non-negotiable.

Coming back to the PESCO projects and taking into consideration Article 5(6) of the Council's decision (Council of the European Union, 2017), which states that '*... the participating Member States, taking part in a project, shall agree among themselves on the arrangements for, and the scope of, their cooperation, and the management of that project...*' In terms of efficient and effective multi-project coordination, supervision and management, this management approach can be called a '*Babel Tower approach*'.

4.3.4.1 The Council's provisions for project governance

The December 2017 decision (Council of the European Union, 2017) provided several high level governance rules including the following characteristic examples:

- Commitment to drawing up harmonised requirements for all capability development projects agreed by participating Member States.
- Commitment to consider the joint use of existing capabilities in order to optimise the available resources and improve their overall effectiveness.

- Aiming for fast-tracked political commitment at national level, including possibly reviewing national decision-making procedures.
- Commitment to agree on common technical and operational standards of forces acknowledging that they need to ensure interoperability with NATO.
- Ensure that all projects with regard to capabilities led by participating Member States make the European defence industry more competitive via an appropriate industrial policy which avoids unnecessary overlap.
- The Member States taking part in a project will agree among themselves on the arrangements for, and the scope of, their cooperation, and the management of that project.
- The Member States taking part in a project will regularly inform the Council about the development of the project, as appropriate.

Operating expenditure arising from projects undertaken within the framework of PESCO will be supported primarily by the Member States taking part in the individual project. Contributions from the general budget of the Union may be made to such projects in compliance with the Treaties and in accordance with the relevant Union instruments.

Before further analysing the abovementioned governance rules, it is crucial to provide some information on the recognised best practices for project stratification. Every organisation, whether public or private, breaks down its management approach into discrete levels, such as the ones depicted below (taken from PM² Guide (Kourounakis & Maraslis, 2016))



Figure 30: Project Management levels

Portfolio Management is a collection of projects, programmes and other activities which are grouped together for better control over financial and other resources, and to facilitate their effective management in terms of meeting strategic objectives.

Programme Management is a group of related projects grouped together to facilitate a level of management which will make it possible to achieve additional objectives and benefits that would not have been possible if these projects were managed individually.

Project Management includes the activities of planning, organising, securing, monitoring and managing the necessary resources and work to deliver specific project goals and objectives in an effective and efficient way.

While keeping in mind the above stratification, the 17 PESCO projects could be regrouped and programmes could be formed (not all projects would have to be included in a specific programme). The various programmes and independent projects would form the PESCO Project Portfolio (see Appendix D). This would create discrete levels of authority, thus enhancing project governance. The Council's governance rules, as presented in its December 2017 decision (Council of the European Union, 2017), are at Project Portfolio level.

4.3.4.2 The road towards a better project management approach. The case of PM2.

Prior to launching a new project, initial research should be undertaken in order to identify an optimal solution (given the level of information available at that time). This research is normally presented in a 'feasibility study' (ProjectManagementDocs, n.d.-b) or a 'business case' (Kourounakis & Maraslis, 2016; Project Management Docs, n.d.-a) in which multiple aspects of relevance to the project are taken into consideration, e.g. alternative

approaches, technological limitations, marketplace conditions, staffing requirements, schedule and budgetary projections. This helps to determine whether or not a project justifies the organisation's investment (following a thorough cost/benefit analysis) and whether it is aligned with its strategic plans.

To date, although only a vague description of the 17 PESCO projects exists, formal initial approval has already been granted by the European Commission to proceed with their implementation (see Appendix D). An alternative to the abovementioned way forward could have been to use one of the best practices, namely the Open PM² methodology (Kourounakis & Maraslis, 2016) developed by the European Commission.

In point 1.1 of the PM² Guide we get a clear view of what PM² has to offer:

'PM² is a Project Management Methodology developed by the European Commission. Its purpose is to enable Project Managers (PMs) to deliver solutions and benefits to their organisations by effectively managing project work.'

PM² has been created considering the environment and needs of EU institutions and projects, in order to facilitate the management of projects' complete lifecycle.'

PM² incorporates elements from a wide range of globally accepted project management best practices, described in standards and methodologies, as well as relevant European Commission communications and operational experience from various internal and external projects.'

The above methodology is by no means a binding set of rules and procedures, but rather an amalgam of best practices, made available for adaptation to individual project needs. This adaptation is performed through the 'tailoring procedure'. In this phase, project managers decide which elements of methodology are to be used for a particular project depending on its nature, the characteristics of the performing organisation, existing levels of expertise, etc.

Some of the most critical aspects of project management that could be covered through the adaptation of the PM² methodology are the following¹⁹:

- A solid project governance structure, providing support and insight both at Council level and at Member State level.
- A flexible set of guidelines for project planning, throughout the project's lifecycle.
- A balanced approach towards tackling project constraints, namely: scope, time, cost and quality.
- A firm risk management approach that helps diminish ambiguity and minimise the impact/probability of negative risks (threats), or maximise the impact/probability of positive risks (opportunities).

¹⁹ **The list is indicative and not exhaustive**

- A ready-to-use set of artefacts (document templates), along with thorough guidelines concerning their usage.
- A structured communication framework enabling efficient and effective dissemination of project information to project stakeholders.
- A formal framework for grouping and tackling stakeholders based on their individual characteristics and needs.

Well-designed monitoring and control activities necessary for managing the project.

4.3.5 Is there room for the European Security & Defence College (ESDC) between PESCO & PM²?

The European Security and Defence College (ESDC) was established in 2005 as a network college, comprised of 140 national entities, with the aim of providing strategic-level education and training for the Common Security and Defence Policy (CSDP). The training audience includes civil servants, diplomats, police officers and military personnel from the EU Member States and staff from EU institutions/agencies involved in CSDP. In most cases, partner countries and organisations are invited to send participants to attend ESDC courses.

Having trained more than 20 000 personnel so far, the ESDC has become a recognised training and education provider within the EU framework. It offers courses for legal and political advisers, as well as courses on human rights,

mediation and negotiation, maritime security, cyber-security and the fight against corruption, to name just a few.

With its wealth of experience in providing high-quality training, the ESDC might be the ideal place to bring together PESCO and PM², and this would help to overcome the problem of there not yet being a defined project management approach within PESCO. Hence, the ESDC will provide training for PM² (under the auspices of the Open PM2 Centre of Excellence), for the personnel tasked with implementing the 17 PESCO projects. This new ESDC task will facilitate capability development within CSDP, is therefore fully in line with the current mandate and would create the following benefits:

- PESCO and the 17 projects are an integral part of CSDP and will support the CSDP missions and operations in terms of efficiency and effectiveness
- the ESDC, through its role as a network college, will help bring together subject matter experts (SMEs) in the field of PM² and the PESCO project managers
- the ESDC can use its ILIAS learning management system (LMS), launch e-training modules and thus minimise expenses and time loss (Rehrl & Cammel, 2017).

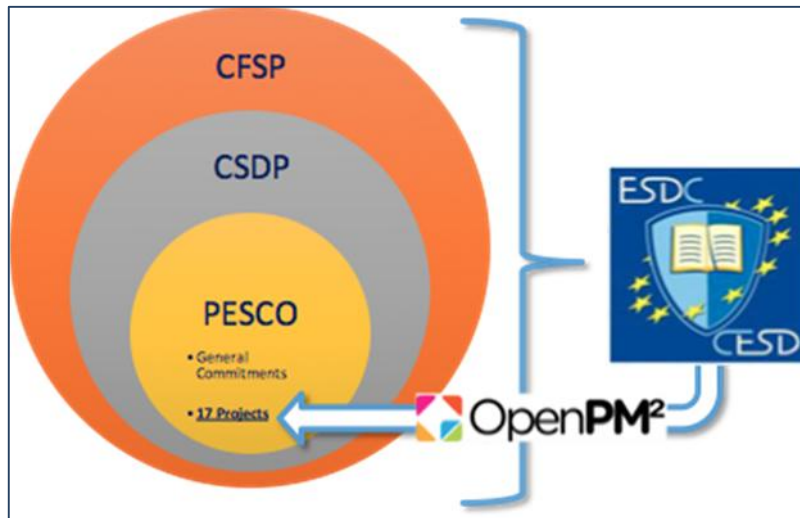


Figure 31: ESDC contribution

4.3.5.1 Win-Win-Win! A multiple-gain approach

The proposed way forward would create a win-win-win environment for PESCO, the ESDC and PM².

- For PESCO projects
 - Structured, streamlined approach in terms of project management, based on global standards
 - Knowledge spill-over
 - Participating personnel with in-depth domain knowledge
 - Stratification of projects (portfolio/programmes/projects), providing more efficient and effective governance
 - Access to numerous templates and guides for all personnel participating in the projects

- Acquisition of domain-specific knowledge in critical project management areas, such as: a) stakeholder management, b) scope-time-cost management, c) quality management, d) human resource management, e) procurement management and, last but not least, f) risk management
- Better governance through structured approaches to monitoring and controlling project work
- Early identification and handling of potential risks (opportunities and/or threats)
- Improved communication and information dissemination
- Stakeholder satisfaction
- For PM²
 - Boost visibility and reputation
 - Increase 'market' penetration (within EU institutions and cooperating organisations)
 - Knowledge spill over/dissemination
 - Acquire new feedback from the performing organisations and thus apply new improvements to the methodology and its supporting framework
- For the ESDC

- Broaden the spectrum of the training activities and audiences
- Increase its visibility to new areas and organisations
- Support the establishment of a European security culture
- Enrich training content through the accumulation of supplementary knowledge
- Participate in the 'requirements collection' process for the PESCO projects
- Use synergies to transfer knowledge in the CSDP area

4.3.6 Training Method

The project management training, which is to be provided to all receiving organisations of the 17 PESCO projects, cannot take a 'one size fits all' form. The training approach should be adapted to the authority level and the needs/requirements of the receiving party. An overview of the proposed training stratification is presented in the table below:

RECEIVING PARTY	NUMBER OF PARTICIPANTS	FORM OF SERVICE	METHOD	TECHNOLOGY
Portfolio manager	1	Mentoring	F2F or e-mentoring	Video conference PM ² wiki
Programme managers	7	Training	F2F	Classroom PM ² wiki
Project managers	17	Training	Synchronous e-training	ILIAS LMS PM ² wiki
Project team members	Unknown	Training	Asynchronous e-training	ILIAS LMS PM ² wiki
External cooperating entities (e.g. subcontractors)	Unknown	Training	Asynchronous e-training	ILIAS LMS PM ² wiki

Table 15: Proposed Stratification

4.3.6.1 Portfolio Manager

The portfolio manager could be an assigned entity from within the European Defence Agency (EDA) or the European Union Military Staff (EUMS) (= PESCO Secretariat) acting under the strategic guidance of the Council and the HR/VP. In the case of the portfolio manager, the mentoring approach is considered the best-suited methodology, rather than the teaching/training one. This is based on the assumption that high-level executive officers already possess a great deal of domain knowledge and expertise; the complementary approach should aim to structure existing skill sets and enhance their

applicability in ways that promote performance. There are multiple definitions of mentoring (Bozeman & Feeney, 2007; Chao, 1997), but we could encompass them as follows:

'Mentoring is the process involved with the diffusion of knowledge (occasionally bilaterally), social capital (and even psychological support), perceived by the recipient as relevant to his/her work, from within a relationship of mutual trust.'

Traditional mentoring and e-mentoring differ only in the communication method used. Face-to-face (F2F) mentoring takes place in personal meetings where participants interact in person and synchronously. e-Mentoring is performed via technology (virtually), either synchronously or asynchronously. In order to be comprehensive, hybrid or blended mentoring should be mentioned as well, where the mentor and the mentee interact either face to face or virtually (Murphy, 2011). The only caveat to the above approach is the challenge of identifying and hiring the right mentor, capable of delivering in line with expectations.

4.3.6.2 Programme Managers

Programme management is a way of achieving strategic goals and objectives through the coordinated management of related projects. The same benefits could not be attained when the projects are individually managed. Through the proposed training, the programme manager should

develop skills which will help him or her to perform relevant tasks (Zein Omar, 2010), such as:

- Defining programme governance
- Planning overall programme management
- Managing the programme's budget and schedule
- Managing risks and issues while taking corrective measurements
- Coordinating the projects and their interdependencies
- Managing and utilising resources across projects
- Managing stakeholder communications
- Aligning individual project deliverables (outputs) with the programme's 'outcome'.

Given that programmes include projects performed from more than one country, the role of the programme manager should be performed by SMEs from within the EDA. Communication and central coordination will thereby be enhanced and governance facilitated. The classroom training for programme managers can be provided by the ESDC within its or the EDA's premises.

Other positive side effects would be:

- The small number of participants
- Proximity to the PM² Centre of Excellence
- Minimisation of related costs

4.3.6.3 Project Managers

Project management in general and the selection of the participating personnel rests in the hands of the Member States leading and participating in each project (Council of the European Union, 2017). Although it is clearly set out in the Council Decision that the list of the project members of each individual project is to be attached to the corresponding Council Decision, given the time horizon of the projects we cannot exclude the possibility that a certain rotation of personnel will take place. In any case, we strongly believe that such rotation should not affect the project managers, given the importance of their roles in the successful evolution of the projects (excluding situations of substantiated underperformance on their part).

After having identified the project manager's role, the focus must be on the training in order to guarantee a relatively uniform level of skills and performance. The training programme should aim to build the following skills (Project Management Institute, 2017):

- Technical project management
- Leadership
- Strategic and business management.

Apart from the technical skills, the Project Management Institute (PMI) identified two other skill sets: a) *strategic and business management skills*, which will convey a clear image of the performing organisation's strategic goals in order to be able to effectively negotiate and implement the decisions

supporting strategic alignment and innovation; b) *leadership skills*, which involve the abilities of the project manager to effectively guide, motivate and direct the project team.

Assuming that there are 17 project managers, the synchronous e-training approach is the best solution. In general, e-training involves using technology to educate and train. This can be done face to face or via remote computer-mediated communication (CMC) or purely online training (Mohsin & Sulaiman, 2013); in this specific case, remote, purely online and synchronous e-training could be used. This approach minimises the impact of other factors such as setting up training locations, increased set-up/overhead costs, loss of time in commuting, schedule constraints, etc. (Loh, Lo, Wang, & Mohd-Nor Rohaya, 2013).

4.3.6.4 Project team members and external cooperating entities

This group consists of an undefined number of trainees. They will be the principal workforce of the various PESCO projects. Their training should focus on building detailed skill sets, in particular for technical project management and artefacts handling. Unfortunately, the possible large number and dispersion of trainees would make both F2F and synchronous e-training

methods unsuitable. However, the evolution of technology, especially in the field of LMS, could help us to overcome these obstacles.

Some of the principal ground rules to be considered when setting up an asynchronous e-training module are as follows:

- Motivate participants (Law et al., 2010): We need to substitute the missing interaction with the instructor with other actions that promote learner engagement. A best practice is to communicate tangible course goals up-front that highlight the usability of the course content. Another effective approach is to use realistic scenario training, with the scenarios being similar to the situations that the trainee is going to face in his or her day-to-day job. Furthermore, having a clear view of Keller's ARCS model (Keller, 1987) of instructional design helps us understand the major influences on motivation to learn. (A concise summary of the model can be found on the Learning Theories website ("ARCS MODEL OF MOTIVATIONAL DESIGN THEORIES (KELLER)," n.d.).)
- User-friendly interface: The graphical user interface (GUI) is of tremendous significance in an e-learning course (Ahmad et al., 2004; Zhang & Nunamaker, 2003). We should aim for the best first impression and active engagement by using clear navigation schemes and well-structured content.

- Keep participants interested by incorporating variety in the learning activities: The list of potential tools is endless and could include interactive simulations, case studies, quizzes and games.
- Content chunking (Clark & Mayer, n.d.; Mayer & Moreno, 2003; Mödritscher, 2006): The human brain, which is capable of storing a quadrillion bytes of data and performing extremely complex operations, slows down to the speed of a snail when asked to recall 10 numbers or repeat just a few simple words. This has to do with the actual working memory of a human brain (similar to the RAM in our PCs). In e-learning, content chunking is the process of presenting content in the form of crisp sentences and bulleted or numbered lists. Instructional designers break down long strings of information into bite-sized absorbable pieces, helping learners to stay focussed.
- Include effective assessment strategies (Roberts, 2006; Wang, 2007): Constructive feedback as an outcome of a well-structured assessment has multiple positive effects over the training experience. Apart from helping learners identify their weak points, it improves training effectiveness by boosting memory retention. The effort of retrieving information (no matter the outcome) makes it easier to retrieve when needed (Lahey, 2014; Roediger & Karpicke, 2006)

4.3.6.5 The common denominator

The PM² initiative already has in place several supporting functions, one of them being the PM² wiki (online resources). After acquiring access to it through EU Login Registration, (the European Commission's Authentication Service website ("EU Login Registration - European Commission," n.d.)), the trainee will find him/herself immersed into a plethora of resources and supporting material that can guide him/her through the project life cycle. The specific wiki can further expand and become the solid foundation for all the proposed training approaches presented in the table through the addition of the microlearning approach (Gassler, Hug, & Glahn, 2004).

Microlearning is a relatively new concept in the e-training world and deals with providing small learning units and short-term learning activities. Ideally, it promotes repetitive learning through embedding learning patterns into the receiver's daily routine, by making use of communication supporting devices (Gassler et al., 2004) (e.g. tablets, laptops, mobiles/smartphones, etc.) (Statista, n.d.).

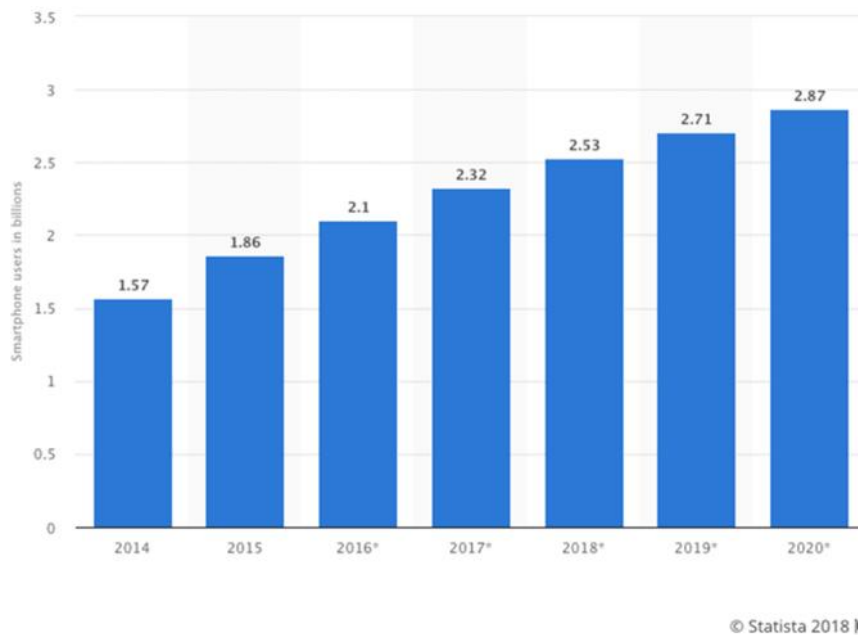


Figure 32: Smartphones in Billions

Distinct examples of microlearning are flashcards, mini expert video tutorials, games, quizzes, short podcasts, road maps, etc. Some other terms used interchangeably with microlearning are: a) learning chunks, b) learning nuggets, c) bite-sized learning and d) snackable content (ELDRIDGE, 2017).

The most prominent device for microlearning is the smartphone, followed by tablets (both mobile devices). The table below, which is constantly updated by the UN International Telecommunications Union (ITU), sets out the number of active mobile broadband subscriptions per year from 2007 to 2017 (International Telecommunications Union, 2017b, 2017a)

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017*
Developed	225	336	450	554	712	829	927	1 015	1 118	1 189	1 227
Developing	43	86	165	253	471	721	1 032	1 645	2 179	2 676	2 993
World	268	422	615	807	1 184	1 550	1 959	2 660	3 297	3 864	4 220

Figure 33: Active mobile broadband subscriptions (in millions) * estimation

The above growth in the use of mobile devices (with internet access) has formed an area of further research in the field of microlearning. Mobile devices provide a power platform for delivering personalised and just-in-time training content. Some other statistics to be considered are those on video usage over the internet (research performed by Syndacast for the year 2014 with projections for the year 2015) (Syndacast, 2014; Vinu, Sherimon, & Krishnan, 2011):

- Online video accounts for half of all mobile traffic
- 65 % of video viewers watch at least two thirds of a video
- More than 80 % of senior executives watch more online video now than they did a year ago
- 75 % of business executives watch work-related videos at least once a week
- 59 % of senior executives say that if both text and video are available, they prefer to watch the video version
- 96 % of business to business (B2B) companies are planning to use video in content marketing over the next year
- Using the word 'video' in an email subject line boosts open rates by 19 %

- 78 % of people watch videos online every week
- By 2018, video will make up 79 % of consumer internet traffic.

Based on the above statistics, the incorporation of video microlearning content in the PM² wiki (and any other LMS in general, including the ESDC's ILLIAS), would boost participants' interaction, enhance knowledge absorption, and further promote its content.

4.3.7 Conclusion

In the present paper, we have tried to present the potential that exists in the current highly volatile situation regarding PESCO projects, by applying the Open PM2 methodology and the ESDC's training expertise. The proposed 'glue' to bring these three elements together is e-learning in all of its discrete aspects, while taking into consideration the participants' characteristics and training needs.

Numerous studies have demonstrated that e-learning within an organisation is a win-win situation for all parties involved: the trainee, the trainer, the organisation and its customers, and in specific cases even the wider public. We believe that the proposed approach will confirm the above findings.

Quoting the Chair of the European Union Military Committee during his recent speech at the College of Europe, we can say with certainty that the challenges are ahead of us (Kostarakos, 2018):

'The real work in every project begins when the conceptual phase is over.

When whatever has been decided and agreed upon will be up for implementation. This will be the real test for the honesty of purposes and the validity of commitments.'

4.4 On the M4.0 forecasting competition: can you tell a 4.0 earthquake from a 3.0?

4.4.1 General

In this short, but yet very sharp paper, we present and highlight the need for a stricter framework when it comes to new quantitative methods performance evaluation. In our first paper we set the example by using as a benchmark a trained group (in a set of skills), to compare it with the group that has received some additional training in structured analogies and forecast decomposition. We knew all the way from the beginning that we were risking not to be able to achieve a significant difference in performance but nevertheless, we did not give in and planned ahead. Our courage paid off, and our super trained group exhibited super-performance.

But we were measuring and comparing performance over single –point forecasts, in a judgmental forecasting framework. The need to go the extra mile and provide forecasting tournaments bringing together judgmental and pure quantitative forecasts has been recently highlighted through IARPA's new forecasting tournament, namely the Hybrid Forecasting Competition (HFC)(IARPA, n.d.).

The HFC program takes the ACE program(IARPA, 2010) one step ahead and aims to develop and test hybrid geopolitical forecasting systems. The desired forecasting systems will be able to bring together human and machine forecasts in order to increase forecasting accuracy. Both systems (human and machines) come with their pros and cons. E.g. human generated forecasts are lenient to biases, whereas machine generated forecasts strive to provide accurate forecast whenever the historical data is either limited on non –existent.

Through the identification of an optimal hybrid approach, IARPA aims to maximize the effect of the strengths and minimize the impact of the weaknesses of the two different methods. These systems will be evaluated through a multi-year competition to identify approaches that may enable the

Intelligence Community (IC) to profoundly improve the accuracy and timeliness of geopolitical forecasts.

The machine generated forecasts that will be provided throughout the tournament, and for each of the posed questions will be most probably based on quantitative data in the form of time series. What we propose is that the performance of the machine generated forecasts are benchmarked against the already established and globally acknowledged 'fast & cheap' benchmarks, like the Theta Method, ARIMA, Damped ES and ETS, not to mention more advanced methods like the awarder MAPA method (Kourentzes, Petropoulos, & Trapero, 2014).

4.4.2 Abstract

Twenty years on from the publication of the results of the well-celebrated M3 competition, and right about the time we got used to the idea that there will be no more M-type competitions, the M4 competition came in 2019. A 4.0 earthquake is 10 times 'stronger' than a 3.0, and that was what M4.0 was aspiring to; mission accomplished?

Keywords: M4 competition; Hybrid method; Combination; Benchmark; Intermittence;

4.4.3 First cut is the deepest

First cut is the deepest and probably no new forecasting competition will ever have the impact of M1 (S. Makridakis et al., 1982). There are 1360 citations to date in that article and many IIF members argue that the whole discipline is practically an offspring of M1.

The first time you say the story that simplicity matters, and simple models can be as accurate, more robust than complex ones, it breaks the waves: you definitely feel that 'scientific earthquake'. Nevertheless, as Makridakis himself admitted in an interview to (Fildes & Nikolopoulos, 2006) : "I don't know if there is more work to be done on this type of competitions".

Strangely enough, the M2 competition that was very different (Spyros Makridakis et al., 1993): focusing on non-disguised data and comparing real experts, working in real series, been able to search for whatever information they wanted and even using their judgment to forecast; that is the one that got the least attention (288 citations to date).

4.4.4 Thinner, Lighter, Faster

M1 was almost ten times bigger than its predecessor was. The M3 competition was three times bigger than M1, with more methods and metrics employed. It was indeed impossible to run in real time 3003 long series in the

early 80s, but that was definitely doable (over a weekend actually) in one expensive PC in late 90s; today you can probably do that in less than a minute in a 100\$ laptop. However, a forecasting competition is not meant to be like a new iPad: thinner, lighter, and faster; it must every time redefine expectations on how empirical forecasting evaluations should be performed.

4.4.5 A new competition

A new forecasting competition cannot just be ten times bigger than the previous one (Spyros Makridakis & Hibon, 2000). In order to claim the 4.0 in the long history of forecasting competitions (Hyndman, 2019), M4 brought in new things: far more series, more categories, prediction intervals, replicability, and full transparency. In addition, industry participation for the first time was a major plus; and an open invitation to the machine learning community to really take part in M4.

4.4.6 Reality matters and more can be done

One fundamental question remains unanswered: does M4 represents reality? How do companies really produce forecasts? There is evidence (Fildes & Goodwin, 2007), that forecasts are prepared in practically no time, for thousands of time series, with forecasters being familiar with only a few SKUs, in outdated systems that users often do not trust and override continuously.

Reality matters and our personal take is that blind and static competitions are not fit-for-purpose any more. We need competitions with real series, for products and services known to the participants. We need participants to provide point forecasts and prediction intervals regularly every 2-3 months. That needs commitment, but people in real life do that at much higher frequencies, and they are committed, so it is definitely doable.

It is also very important to focus on the series that really matter in real life. We tend to forecast in vacuum and think that it does not matter if we forecast 'apples' or 'oranges': but it does. For example, in finance for an investment bank to take investment decisions a set of time series needs to be forecasted regularly. From personal communications with an investment bank based in London we know that many economic and monetary series are monitored in a financial forecasting context. Every trading house uses obviously more or less series, but this is the common denominator in the financial sector. Therefore, size does not matter in the design of the 'finance' subset of the next forecasting competition; we need less and named series if we are to move forward, rather than more (and collinear) anonymous series.

4.4.7 Sins of commission

What else real life is? Real life is intermittent: 60% of any inventory consists of spare parts, and these are not cheap to stock. So we do have 60% of SKUs in any warehouse that present intermittent demand patterns but we have decided to ignore such series from our forecasting competitions for the last 40 years. There must be a rational for not including such series, but it looks more like a sin of commission rather than one of omission.

4.4.8 The winner takes it all

The team from Uber led by Smyl is the winner, by a good margin (see the results in table 4 of (S. G. Makridakis, Spiliotis, & Assimakopoulos, 2019)). From second to sixth position we find five different combinations: this is something we expected, maybe not to that extend; in fact, in the top-25 positions we find 15 combinations.

In the past M-competitions big private organizations have not participated. They have had in other types of competitions, but not the M-type ones. This time M4 got the attention of the likes of Uber and Amazon and Microsoft, even if not all of them formally participated. The win of Uber also advocates for the fact that there is a lot of forecasting expertise in the practitioners' community. This expertise and research taking place in industry is not scholarly reported in IJF. Uber's method was impressive by itself – a hybrid method, state of the art technically; and intuitively appealing as it exploits

properties of the entire dataset every time forecasts are produced for an individual time series.

We also notice that Forecast pro outperforms all benchmarks including the Theta method (Assimakopoulos & Nikolopoulos, 2000), the latter being the only method that performed better than it in M3. There were no articles or announcements in the recent years about any change in the core algorithm of Forecast Pro. The more forecasts needed, the more accurate Forecast Pro becomes, and the selection algorithm it employs eventually outperforms individual methods – even the ones not included in its engine. This is a sign of robustness and consistency and this is all good news for the Forecast Pro team. This is also good news for the entire commercial Forecasting Support Systems development community. We also must congratulate the company for always been willing to test their software in real blind competitions, and face the respective publicity that comes with it.

4.4.9 Omelets and eggs

Given that there were so many submissions in the 'combinations' category, and performed so well, it is inevitable to ask the obvious question: who gets the credit? So if someone does an equal weighted combination of Theta method, ARIMA and ETS should the credit go to the one combining, or

to those developed those three methods, to both, or none? As the famous football manager Jose Mourinho¹ has nicely once put it:

“ ‘Omelets and Eggs’: you cannot make a good omelet without good eggs...”

4.4.10 Time is of the Essence

Despite the cloud services and the unlimited computing power than one can buy nowadays, time is still of the essence. It was more of an issue 20 years ago for the M3. Nevertheless, if a method needs 3 days to run in an i7 laptop, while another method runs in 7 minutes or 7 seconds, this arguably constitutes a competitive advantage. A major retailer has only a window a few hours every night in order to forecasts 100K to 150K SKUs. Of the M4 more advanced benchmarks, Theta method seems to have the edge running in 12.7 mins for the entire 100K series of M4 dataset in Amazon Web Services with 8 cores, ETS coming second with 888 mins, and ARIMA third with 3030 mins.

4.4.11 The one to beat

Over the years, the IIF community has seen many forecasting studies that proposed new methods that could only outperform Naïve, a moving average or just ETS; this is methodologically wrong, and we should as an academic community work towards banishing this phenomenon. It has been obvious for the last two decades that there is a series of very accurate methods, which

are computationally cheap and free in R and Python packages for example Hyndmans's forecast package.

M4 results corroborated grossly to this; in any empirical forecasting investigation, the following methods should be employed as benchmarks - in order of performance in M4 (table 4, Makridakis et al., 2019): The Theta method – even the basic model used in M3 and not one of the advanced ones (K. I. Nikolopoulos & Thomakos, 2019), ARIMA, Damped ES and ETS. In addition, combinations should be employed starting with the average of Simple, Holt, and Damped exponential smoothing.

We also propose that we should also use the mean and median of the combination of: Theta method, ARIMA, ETS, and Damped ES. Any newly proposed forecasting method, in order to be publishable, should be on par or better than these 'fast and cheap' benchmarks – and probably even more advanced methods like the awarded MAPA method (Kourentzes et al., 2014); c'est la vie!

Verdict: We really felt this 4.0 'scientific earthquake'.

Chapter 5 Discussion

5.1 Summary

This research aimed at expanding 'by containing' the findings of the GJP. We have narrowed down significantly the applicable methods that the forecasters could use in order to provide their judgemental point forecasts. In particular, we prompted the additionally trained group to make use of a modified version of structured analogies by combining it with forecast decomposition.

The use of analogies is an acknowledged forecasting method, that assumes that two different kinds of phenomena share similar behaviours. It is considered a very convenient technique especially in the cases where no actual historical data exist for the question at hand. We thus considered forecasting by analogies a prominent methodology for forecasting geopolitical events with limited historical data, within the framework of a forecasting tournament.

We focused on the above approach having in mind SMEs and their limited capacity in resources. In general, in order to be classified as an SME an

entity should have a somewhat small number of employees, usually spanning from 10 to 250 (the actual number depends on the country where the company is registered). SMEs, are considered the backbone of Europe's economy²⁰. They represent 99% of all businesses in the EU and have created approximately 85% of new jobs that being two-thirds of the total private sector employment in the EU.

'The European Commission (EC) considers SMEs and entrepreneurship as key to ensuring economic growth, innovation, job creation, and social integration in the EU'.

According to the EC, in order for an entity to be characterised as an SME, it should comply with the following characteristics:

²⁰ https://ec.europa.eu/growth/smes_en

Company category	Staff headcount	Turnover	or	Balance sheet total
Medium-sized	< 250	≤ € 50 m		≤ € 43 m
Small	< 50	≤ € 10 m		≤ € 10 m
Micro	< 10	≤ € 2 m		≤ € 2 m

Table 16: SME determining factors according to the EC²¹

The limited staff within the SMEs is responsible for many (and sometimes all) tasks relating to innovation, production, marketing, sales and accounting, for the entire business. This can be occasionally a major handicap given employees might not have the required dexterities to perform everything equally proficiently, but in the long run, they develop 'umbrella skills' that broaden their view and boost their performance. In that vein, we believe that all personnel should be considered an asset of high value and should participate in the organizational decision making procedure, even just through the provision of forecasts.

When it comes to forecasting, most organizations and SMEs in particular use both judgemental and quantitative forecasting techniques, but in the long run there is heavier weight being placed on the judgemental forecasting/adjustment, especially when that comes from the higher levels of the organizational hierarchy. Judgemental forecasting (although inferior to

²¹ https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en

quantitative methods when adequate historical data exists) can be considered efficient and effective when the following set of conditions apply:

- The forecaster has access to information that cannot be included or represented within the statistical model (Fildes et al., 2009).
- The forecaster has verified that there is not adequate quantitative data to feed a model.
- The forecaster has exhibited a steady and reliable performance in providing forecasts for a set of similar events, within a relatively steady environment (M. Lawrence et al., 2006).

In the case where all the above apply, then the forecaster can proceed with providing a judgemental forecast, but even in this case, he/she should comply with some ground rules:

- If the problem is too complicated to be handled as one, break it down into smaller more manageable components. This approach should be handled with prudence:
 - Make sure that the sub-components when summed up together again constitute 100% of the initial statement.
 - Clearly distinguish 'or' events and 'and' events. In the former case the probabilities should be added, whereas in the latter they should be multiplied.

- Identify analogies for each of the identified events and score them for relevance prior using them for forecast elicitation. The applicable rules should be clearly communicated to all prior being put in place.
- Use the identified analogies as the well justified anchoring values and build your forecast from that point on.
- Balance the internal and the external view in a problem (Philip Eyrickson Tetlock & Gardner, 2015).
- Justify your forecast and keep records of your mental path in order to use it afterwards for post forecast analysis and personal calibration.
- Challenge estimates. A devil's advocate may be one's best friend.
- Update forecasts cautiously. Try to avoid under reacting or over reacting to the new evidence. Thomas Bayes is our friend and we should strive to follow the key concepts implied by the Bayes Theorem.

All the above apply at individual level. The question now is the following: What should companies do to collectively exploit the mental and forecasting power of their personnel? Should they just randomly focus on the ones with the best CVs and ignore all the others? Our opinion and answer to the above question is 'NO'. Forecasting tournaments have been proven to be precious tools that help us identify, train, group and exploit the collective power of the participating forecasters!

As mentioned in Section [3.2.5](#), the entities that will adopt the procedure proposed in the present thesis will benefit from:

- Improvement of H-R management procedures
 - Enhancement of recruiting methodologies
 - Enhancement of appraisal procedures
 - Better allocation of resources
 - Contribution in creating well “calibrated” job descriptions
- Improvement in forecasting abilities:
 - Enhancement of planning and decision making procedures
 - Enhancement of risk management procedures
 - Effective and efficient use of resources (primarily monetary)
- Insight for future development.

5.2 The way ahead

Research is an ever ending story. One can never stop being on the lookout for new methods, tools and techniques that will contribute to the

improvement of forecasting and decision making in general. In one of my recent LinkedIn articles²² I wrote:

'Risk Management is a high impact trend, but if not applied cautiously can easily turn into a curse!

The famous British historian Thomas Babington Macaulay once said:

'Half knowledge is worse than ignorance.'

*The above statement is valid in many fields and not just in the one of history. The market nowadays is saturated with experts/pundits claiming to possess the risk-tackling elixir! Unfortunately, no such elixir exists, especially when it comes to risk management. Solon, the Athenian statesman, lawgiver, and poet said **'I grow old ever learning many things'** (Γηράσκω δ' αἰεὶ πολλὰ διδασκόμενος). All risk managers and risk consultants should be encouraged to embrace that notion and be engaged in a constant quest for new best practices and knowledge.'*

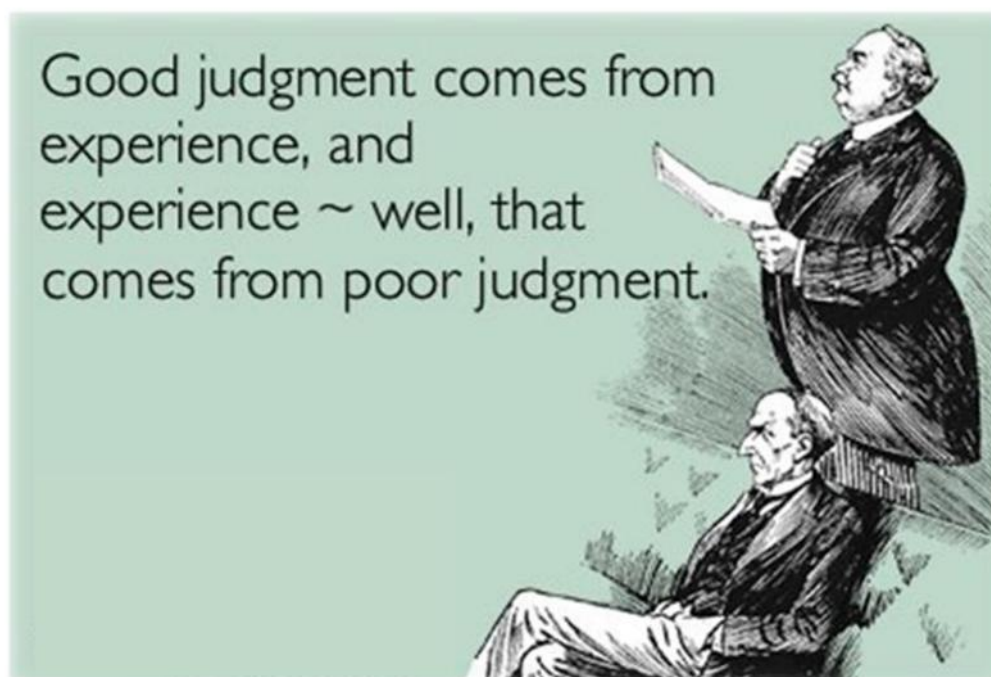
Forecasting is an intrinsic part of Risk Management given a risk is defined as a two element notion, one being the probability of occurrence and the other the impact if the risk was to actually become an event / fact.

Forecasting geopolitical, and not only, events should not be done in a vacuum. An inter-organizational forecasting tournament can be an indispensable tool especially if combined with the need of enterprise and

²² <https://www.linkedin.com/feed/update/urn:li:activity:6539981461704048640>

project risk management. That would mean that we should also be forecasting probabilities for inter-organizational events as well. So far we're not aware of any research that aims to implement a forecasting tournament inter-organizationally and challenge all internal resources to engage in providing probabilistic forecasts to questions relating to the respective internal decisions.

We're looking forward to our next research endeavor...



References

- Ahmad, A.-R., Basir, O., & Hassanein, K. (2004). Adaptive User Interfaces for Intelligent E-Learning: Issues and Trends. In *The Fourth International Conference on Electronic Business (ICEB2004)*. Beijing. Retrieved from [http://watnow.uwaterloo.ca/pub/rahim/AUI-ICEB-2004\(Dec2004\)EN060-paper.pdf](http://watnow.uwaterloo.ca/pub/rahim/AUI-ICEB-2004(Dec2004)EN060-paper.pdf)
- ARCS MODEL OF MOTIVATIONAL DESIGN THEORIES (KELLER). (n.d.). Retrieved February 24, 2018, from <https://www.learning-theories.com/kellers-arcs-model-of-motivational-design.html>
- Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, 110(3), 486–498. <https://doi.org/10.1037/0033-2909.110.3.486>
- Armstrong, J. S. (2001). *Principles of Forecasting* (Vol. 30). <https://doi.org/10.1007/978-0-306-47630-3>
- Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting*, 16(4), 521–530. [https://doi.org/10.1016/S0169-2070\(00\)00066-2](https://doi.org/10.1016/S0169-2070(00)00066-2)
- Bozeman, B., & Feeney, M. K. (2007). Toward a useful theory of mentoring: A conceptual analysis and critique. *Administration and Society*, 39(6), 719–739. <https://doi.org/10.1177/0095399707304119>
- Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Carbone, R., & Gorr, W. L. (1985). Accuracy of Judgmental Forecasting of Time Series. *Decision Sciences*, 16(2), 153–160. <https://doi.org/10.1111/j.1540-5915.1985.tb01480.x>
- Chang, W., Atanasov, P., Patil, S., Mellers, B. A., & Tetlock, P. E. (2017). Accountability and adaptive performance under uncertainty: A long-term view. *Judgment and Decision Making*, 12(6), 610–626. Retrieved from <http://journal.sjdm.org/17/17630/jdm17630.pdf>
- Chang, W., Chen, E., & Mellers, B. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*,

11(5), 509–526.

Chao, G. T. (1997). Mentoring Phases and Outcomes. *Journal of Vocational Behavior*, 51(1), 15–28. <https://doi.org/10.1006/jvbe.1997.1591>

Charles William King. (1908). *Plutarch's Morals: Theosophical Essays: On the E at Delphi*, tr. by Charles William King. Retrieved from <http://www.sacred-texts.com/cla/plu/pte/pte07.htm>

Clark, R. C., & Mayer, R. E. (n.d.). *E-learning and the science of instruction : proven guidelines for consumers and designers of multimedia learning*. Retrieved from https://books.google.be/books?hl=en&lr=&id=v1uzCgAAQBAJ&oi=fnd&pg=PR17&dq=Bite+Size+Content+e-learning&ots=TMuMIGbJhn&sig=uYPOLKtoDrA8y3y3z79x4SKU-x0&redir_esc=y#v=onepage&q=Bite+Size+Content+e-learning&f=false

Council of the European Union. (2017). COUNCIL DECISION establishing Permanent Structured Cooperation (PESCO) and determining the list of Participating Member States. Brussels. Retrieved from <http://www.consilium.europa.eu/media/32000/st14866en17.pdf>

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving Intelligence Analysis with Decision Science (In Press). *Perspectives on Psychological Science*, (June), 1–7.

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, 33(1), 264–277. <https://doi.org/10.1016/j.joep.2011.10.009>

EDA. (2017). Coordinated Annual Review on Defence (CARD) Fact Sheet. Retrieved from https://www.eda.europa.eu/docs/default-source/eda-factsheets/2017-10-05-factsheet_card.pdf

EDC. (1950). Treaty establishing the European Defence Community. Retrieved from <http://aei.pitt.edu/5201/1/5201.pdf>

EEAS. (2017). Permanent Structured Cooperation (PESCO) - Factsheet. Bruxelles. Retrieved from <https://eeas.europa.eu/headquarters/headquarters-homepage/34226/permanent-structured-cooperation-pesco->

factsheet_en

- ELDRIDGE, B. (2017). DEVELOPING A MICROLEARNING STRATEGY WITH OR WITHOUT AN LMS. In *The 13th International Scientific Conference eLearning and Software for Education*.
- EU Login Registration - European Commission. (n.d.). Retrieved February 24, 2018, from https://ec.europa.eu/europeaid/funding/about-grants/how-apply-grant/applicant-registration-pador/eu-login-registration_en
- EUR-Lex. (1992). Treaty of Maastricht on European Union. Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM%3Axy0026>
- EUR-Lex. (2009). Treaty of Lisbon. Lisbon. Retrieved from <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12007L%2FTXT>
- European Council. (n.d.). Permanent Structured Cooperation (PESCO) first collaborative PESCO projects -Overview. Retrieved February 24, 2018, from <http://www.consilium.europa.eu/media/32082/pesco-overview-of-first-collaborative-of-projects-for-press.pdf>
- Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6), 570–576. <https://doi.org/10.1287/inte.1070.0309>
- Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1), 3–23. <https://doi.org/10.1016/j.ijforecast.2008.11.010>
- Fildes, R., & Nikolopoulos, K. (2006). Spyros Makridakis: An interview with the International Journal of Forecasting. *International Journal of Forecasting*, 22(3), 625–636. <https://doi.org/10.1016/J.IJFORECAST.2006.04.008>
- Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3), 331–340. <https://doi.org/10.1002/for.1129>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2015). Why Quantitative Probability Assessments Are Empirically Justifiable in Foreign Policy Analysis. *Working Paper*, 1–33.

- Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament. *International Studies Quarterly*, (April), 1–13. <https://doi.org/10.1093/isq/sqx078>
- Gardner, D. (2011). *Future Babble: Why Expert Predictions Are Next to Worthless, and You Can Do Better*. New York: Penguin Group (USA) Inc. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Future+Babble#1%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Future+Babble:+Why+Expert+Predictions+Are+Next+to+Worthless,+and+You+Can+Do+Better%230>
- Gassler, G., Hug, T., & Glahn, C. (2004). Integrated Micro Learning—An outline of the basic method and first results. *Interactive Computer Aided Learning*, 4, 1–7. Retrieved from <http://www.academia.edu/download/8323082/gassler.pdf>
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)
- Gentner, D. (2002). Analogy in Scientific Discovery: The Case of Johannes Kepler. In *Model-Based Reasoning* (pp. 21–39). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4615-0605-8_2
- Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind : advances in the study of language and thought*. MIT Press. Retrieved from <https://mitpress.mit.edu/books/language-mind>
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, 103(3), 592–596. <https://doi.org/10.1037/0033-295X.103.3.592>
- Goleman, D. (2011). *The brain and emotional intelligence: New insights*. (A. Satpute, Ed.), *More Than Sound LLC* (1st Digital). Northampton MA: More Than Sound LLC.
- Graber, M. (2003). Metacognitive training to reduce diagnostic errors: ready for prime time? *Academic Medicine: Journal of the Association of American Medical Colleges*, 78(8), 781. <https://doi.org/10.1097/00001888->

200308000-00004

- Green, K. C., & Armstrong, J. S. (2007). Structured analogies for forecasting. *International Journal of Forecasting*, 23(3), 365–376. <https://doi.org/10.1016/j.ijforecast.2007.05.005>
- Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge University Press.
- Hadar, J., & Russell, W. R. (1969). Rules for Ordering Uncertain Prospects. *American Economic Review*, 59(1), 25. <https://doi.org/10.1126/science.151.3712.867-a>
- Hanoch, G., & Levy, H. (1969). The Efficiency Analysis of Choices Involving Risk. *Review of Economic Studies*, 36(3), 107–335. <https://doi.org/10.2307/2296431>
- Hernandez, D. (2017). How Our Company Learned to Make Better Predictions About Everything. *Harvard Business Review*, (May 15, 2017).
- Horowitz, M., Stewart, B., Tingley, D., Bishop, M., Resnick, L., Roberts, M., ... Tetlock, P. (2016). What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance At Geopolitical Forecasting, 1–54.
- Hug, T. (2007). *Didactics of Microlearning* (Vol. 7). Munster: Waxmann Publishing Co.
- Hyndman, R. J. (2019). A brief history of forecasting competitions. *Journal of Forecasting* (Forthcoming). Retrieved from <https://robjhyndman.com/papers/forecasting-competitions.pdf>
- IARPA. (n.d.). Hybrid Forecasting Competition (HFC). Retrieved May 27, 2019, from <https://www.iarpa.gov/index.php/research-programs/hfc?highlight=WyJoZmMiXQ==>
- IARPA. (2010). Aggregative Contingent Estimation (ACE). Retrieved May 27, 2019, from <https://www.iarpa.gov/index.php/research-programs/ace>
- International Telecommunications Union. (2017a). GLOBAL AND REGIONAL ICT DATA: Active mobile-broadband subscriptions. Retrieved February 24, 2018, from <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>
- International Telecommunications Union. (2017b). *Measuring the Information Society Report 2017*. Tunisia. Retrieved from https://www.itu.int/en/ITU-D/Statistics/Documents/events/wtis2017/Plenary3_Skhirtladze.pdf
- Kahneman, D. (2013). *Thinking, fast and slow*. Macmillan.

- Katsagounos, I., & Rehr, J. (2018). "PESCO - PM2 - ESDC" Could E-Learning Bring Closer Together EU's Success Stories? Bucharest: eLSE.
- Keller, J. M. (1987). Development and use of the ARCS model of instructional design. *Journal of Instructional Development*, 10(3), 2–10. <https://doi.org/10.1007/BF02905780>
- Khong, Y. F. (1992). *Analogies at war: Korea, Munich, Dien Bien Phu, and the Vietnam decisions of 1965*. New Jersey 08540: Princeton University Press. Retrieved from <https://press.princeton.edu/titles/5008.html>
- Kostarakos, G. M. (2018). Speech at the College of Europe "PESCO, Possible Game Changer for Security and Defence Cooperation in the EU"; Bruges. Retrieved from https://eeas.europa.eu/sites/eeas/files/20180213_ceumc_speech_at_the_college_of_europe.pdf
- Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, 30(2), 291–302. <https://doi.org/10.1016/J.IJFORECAST.2013.09.006>
- Kourounakis, N., & Maraslis, A. (2016). *PM2 Project Management Methodology - Open Edition (0.9)*. Brussels: European Commission, DIGIT Centre of Excellence in Project Management (CoEPM2). <https://doi.org/10.2799/957700>
- Lahey, J. (2014). Students Should Be Tested More, Not Less - The Atlantic. Retrieved February 24, 2018, from <https://www.theatlantic.com/education/archive/2014/01/students-should-be-tested-more-not-less/283195/>
- Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books. Retrieved from [http://sr-ix.com/CA/ATEK635/readings/Philosophy in the Flesh.pdf](http://sr-ix.com/CA/ATEK635/readings/Philosophy%20in%20the%20Flesh.pdf)
- Laplace, P. S., & Dale, A. I. (1995). *Philosophical essay on probabilities*. Springer-Verlag. Retrieved from https://books.google.be/books/about/Pierre_Simon_Laplace_Philosophical_Essay.html?id=vDZzuGcM4DUC&redir_esc=y

- Law, K. M. Y., Lee, V. C. S., & Yu, Y. T. (2010). Learning motivation in e-learning facilitated computer programming courses. *Computers & Education*, 55(1), 218–228. <https://doi.org/10.1016/J.COMPEDU.2010.01.007>
- Lawrence, M., Goodwin, P., O'Connor, M., & Önköl, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22(3), 493–518. <https://doi.org/10.1016/j.ijforecast.2006.03.007>
- Lawrence, M. J., Edmundson, R. H., & O'Connor, M. J. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting*, 1(1), 25–35. [https://doi.org/10.1016/S0169-2070\(85\)80068-6](https://doi.org/10.1016/S0169-2070(85)80068-6)
- LITSA, A., PETROPOULOS, F., & NIKOLOPOULOS, K. (2012). Forecasting the Success of Governmental "Incentivized" Initiatives: Case Study of a New Policy Promoting the Replacement of Old Household; Air-conditioners. *Journal of Knowledge Management, Economics and Information Technology*, 2(1), 1–15. Retrieved from <https://ideas.repec.org/a/spp/jkmeit/1262.html>
- Liu, C., Vlaev, I., Fang, C., Denrell, J., & Chater, N. (2017). Strategizing with Biases: Making Better Decisions Using the Mindspace Approach. *California Management Review*, 59(3), 135–161. <https://doi.org/10.1177/0008125617707973>
- Loh, P. Y.-W., Lo, M.-C., Wang, Y.-C., & Mohd-Nor Rohaya. (2013). Improving the level of competencies for Small and Medium Enterprises in Malaysia through enhancing the effectiveness of e-Training: a conceptual paper. *Labuan E-Journal of Muamalat and Society*, 7, 1–16. Retrieved from <http://www.myjurnal.my/public/article-view.php?id=76282>
- Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., ... Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2), 111–153. <https://doi.org/10.1002/for.3980010202>
- Makridakis, S. G., Hogarth, R. M., & Gaba, A. (2010). Why forecasts fail. What to do instead. *MIT Sloan Management Review*, 51(2), 83–90. <https://doi.org/10.1038/35037613>
- Makridakis, S. G., Spiliotis, E., & Assimakopoulos, V. (2019). The M4 Competition: 100,000 time series and 61 forecasting methods (forthcoming). *International Journal of Forecasting*.
- Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting*:

Methods and Applications. *Journal of Forecasting*, 1.
<https://doi.org/10.1017/CBO9781107415324.004>

Makridakis, Spyros, Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, 9(1), 5–22.
[https://doi.org/10.1016/0169-2070\(93\)90044-N](https://doi.org/10.1016/0169-2070(93)90044-N)

Makridakis, Spyros, & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4), 451–476.
[https://doi.org/10.1016/S0169-2070\(00\)00057-1](https://doi.org/10.1016/S0169-2070(00)00057-1)

Mayer, R. E., & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1), 43–52.
https://doi.org/10.1207/S15326985EP3801_6

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., ... Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, 10(3), 267–281. <https://doi.org/10.1177/1745691615577794>

Mellers, Barbara, Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., ... Tetlock, P. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. *Journal of Experimental Psychology. Applied*, 21(1), 1–14. <https://doi.org/10.1037/xap0000040>

Mellers, Barbara, Tetlock, P., & Arkes, H. R. (2019). Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition*, 188, 19–26.
<https://doi.org/10.1016/J.COGNITION.2018.10.021>

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, 3(1), 1–19. <https://doi.org/10.1037/dec0000032>

Mödritscher, F. (2006). E-learning theories in practice: A comparison of three methods. *Journal of Universal Science and Technology*, 28, 3–18. Retrieved from
http://www.jucs.org/justl_0_0/elearning_theories_in_practice/justl_0_0_0003_0018_moedritscher.html

Mogherini, F. (2016). Shared Vision, Common Action: A Stronger Europe A

Global Strategy for the European Union's Foreign And Security Policy. Retrieved from https://eeas.europa.eu/archives/docs/top_stories/pdf/eugs_review_web.pdf

Mohsin, M., & Sulaiman, R. (2013). A STUDY ON E-TRAINING ADOPTION FOR HIGHER LEARNING INSTITUTIONS. *International Journal of Asian Social Science*, 3(9), 2006–2018. Retrieved from [http://www.aessweb.com/pdf-files/ljass si 3\(9\), 2006-2018.pdf](http://www.aessweb.com/pdf-files/ljass%203(9),%2006-2018.pdf)

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and Motivational Biases in Decision and Risk Analysis. *Risk Analysis*, 35(7), 1230–1251. <https://doi.org/10.1111/risa.12360>

Murphy, W. M. (2011). From e-mentoring to blended mentoring: Increasing students' developmental initiation and mentors' satisfaction. *Academy of Management Learning and Education*, 10(4), 606–622. <https://doi.org/10.5465/amle.2010.0090>

Nikolopoulos, K. I., & Thomakos, D. D. (2019). *Forecasting with the Theta method: theory and applications*. Wiley. Retrieved from [https://www.wiley.com/en-ad/Forecasting+With+The+Theta+Method%3A+Theory+and+Applications -p-9781119320760](https://www.wiley.com/en-ad/Forecasting+With+The+Theta+Method%3A+Theory+and+Applications-p-9781119320760)

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukos, V., & Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of Business Research*, 68(8), 1785–1791. <https://doi.org/10.1016/j.jbusres.2015.03.037>

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, 9(2), 163–172. [https://doi.org/10.1016/0169-2070\(93\)90002-5](https://doi.org/10.1016/0169-2070(93)90002-5)

Office of the Director of National Intelligence. (2017). IARPA Announces Publication of Data from the Good Judgment Project. <https://doi.org/doi:10.7910/DVN/BPCDH5>

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(02), 109–130; discussion 130-178. <https://doi.org/10.1017/S0140525X08003543>

Project Management Institute. (2017). *A Guide to the Project Management Body of Knowledge*. Project Management Institute.

- ProjectManagementDocs. (n.d.-a). Business Case Template. Retrieved February 23, 2018, from <http://www.projectmanagementdocs.com/project-initiation-templates/business-case.html#axzz57xiuLDhq>
- ProjectManagementDocs. (n.d.-b). Feasibility Study Template. Retrieved February 23, 2018, from <http://www.projectmanagementdocs.com/project-initiation-templates/feasibility-study.html#axzz57xiuLDhq>
- Raphals, L. (2013). *Divination and prediction in early China and ancient Greece*. Retrieved from <https://books.google.be/books?hl=en&lr=&id=ObP1AAAAQBAJ&oi=fnd&pg=PR10&dq=Raphals+2013&ots=ff5YiUDdBp&sig=QaDRaJjXhhTWNy6zGz9uRuEdF8>
- Rehrl, J., & Cammel, A. (2017). THE SECURITY POLICY DIMENSION OF TRAINING AND ELEARNING. In *The 13th International Scientific Conference eLearning and Software for Education*. Bucharest.
- Roberts, T. S. (2006). *Self, peer, and group assessment in e-learning*. Information Science Pub. Retrieved from https://books.google.gr/books?hl=el&lr=&id=oV-9AQAQBAJ&oi=fnd&pg=PR1&dq=e-learning+self+assessment&ots=0a7aGfCIX3&sig=DqsxkG6-qQ_0PJ8_jGVjjDxu7N4&redir_esc=y#v=onepage&q=e-learning+self+assessment&f=false
- Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Sadler-Smith, E. (2015). Wallas' Four-Stage Model of the Creative Process: More Than Meets the Eye? *Creativity Research Journal*, 27(4), 342–352. <https://doi.org/10.1080/10400419.2015.1087277>
- Sanders, N. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, 20(3), 353–364. [https://doi.org/10.1016/0305-0483\(92\)90040-E](https://doi.org/10.1016/0305-0483(92)90040-E)
- Savio, N. D., & Nikolopoulos, K. (2013). A strategic forecasting framework for governmental decision-making and planning. *International Journal of Forecasting*, 29(2), 311–321.

<https://doi.org/https://doi.org/10.1016/j.ijforecast.2011.08.002>

Savio, N., & Nikolopoulos, K. (2010). Forecasting the effectiveness of policy implementation strategies. *International Journal of Public Administration*, 33(2), 88–97. <https://doi.org/10.1080/01900690903241765>

Schoemaker, P. J. H., & Tetlock, P. E. (2016). Superforecasting: How to upgrade your company's judgment. *Harvard Business Review*, 2016(May). <https://doi.org/10.1017/CBO9781107415324.004>

Simon, H. A. (1979). Rational Decision Making in Business Organizations. *The American Economic Review*, 69, 493–513. <https://doi.org/10.2307/1808698>

Statista. (n.d.). Number of smartphone users worldwide 2014-2020. Retrieved February 24, 2018, from <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>

Syndacast. (2014). Video Marketing Statistics and Trends 2015. Retrieved February 24, 2018, from <http://syndacast.com/video-marketing-statistics-trends-2015>

Tetlock, P. E., Mellers, B. a., Rohrbaugh, N., & Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science*, 23(4), 290–295. <https://doi.org/10.1177/0963721414534257>

Tetlock, Philip E. (2005). *Expert Political Judgment. How good is it? How can we know?* Princeton University Press. New Jersey 08540: NJ Princeton.

Tetlock, Philip Eyrikson, & Gardner, D. (2015). *Superforecasting: the art and science of prediction*. <https://doi.org/10.1017/CBO9781107415324.004>

Trump, D. (2017). Transcript - Speech: Donald Trump Holds a Make America Great Again Rally in Pensacola, Florida. Retrieved February 23, 2018, from <https://factba.se/transcript/donald-trump-speech-make-america-great-again-pensacola-december-8-2017>

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, New Series*, 185(4157), 1124–1131.

Tweney, R. D. (1991). Faraday's notebooks: the active organization of creative science. *Physics Education*, 26(5), 301–306. <https://doi.org/10.1088/0031-9120/26/5/008>

Ungar, L., Mellors, B., Satopaä, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). The good judgment project: A large scale test of different methods

of combining expert predictions. *Aaai*, FS-12-06, 37–42. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-84875584655&partnerID=tZOtx3y1>

US Department of State. (2018). Country Reports on Terrorism. Retrieved February 23, 2018, from <https://www.state.gov/j/ct/rls/crt/>

Vinu, P. V., Sherimon, P. C., & Krishnan, R. (2011). Towards pervasive mobile learning – the vision of 21st century. *Procedia - Social and Behavioral Sciences*, 15, 3067–3073. <https://doi.org/10.1016/J.SBSPRO.2011.04.247>

Wallas, G. (1926). *Art of thought*. Solis Press.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning*, 23(3), 171–186. <https://doi.org/10.1111/j.1365-2729.2006.00211.x>

Whitmore, G. A. (1970). Third Degree Stochastic Dominance. *American Economic Review*, 60(3), 457–459.

Zein Omar. (2010). Roles, responsibilities, and skills in program management. In *PMI® Global Congress*. Milan: Project Management Institute. Retrieved from <https://www.pmi.org/learning/library/roles-responsibilities-skills-program-management-6799>

Zhang, D., & Nunamaker, J. F. (2003). Powering E-Learning In the New Millennium: An Overview of E-Learning and Enabling Technology. *Information Systems Frontiers*, 5(2), 207–218. <https://doi.org/10.1023/A:1022609809036>

Appendix A Per Team Performance Analysis

1. Standardized (Median IQR) Average Brier Scores per Teams

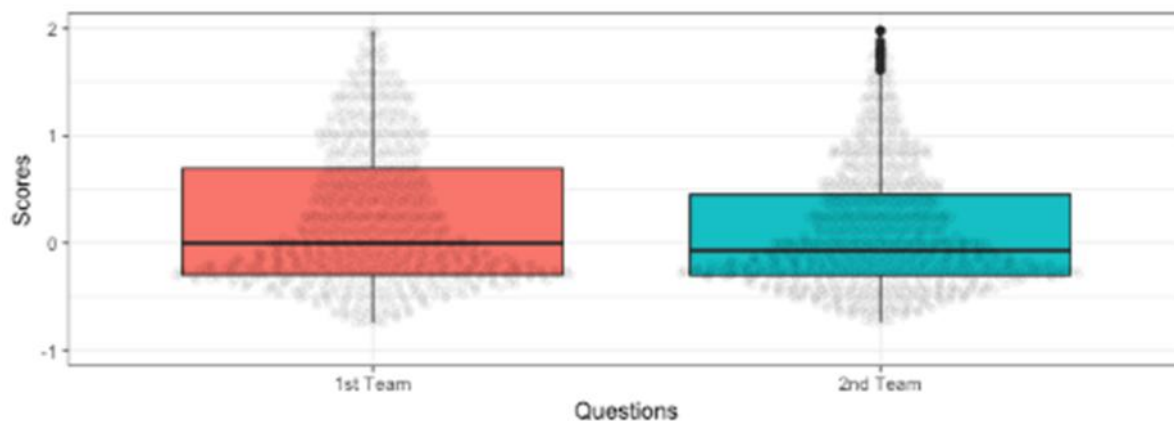


Figure 34: Standardized (Median IQR) Average Brier Scores per Teams boxplots

Perf Measure	Team "A"	Team "B"
min	-0.74583796	-0.74065853
max	8.04444855	8.04444855
median	0.06775545	-0.05475397
mean	0.37695643	0.18942133
SE.mean	0.03293767	0.02543213
CI.mean.0.95	0.06465058	0.04990695
var	0.90479821	0.64226557
std.dev	0.95120881	0.80141473
coef.var	2.52339194	4.23085789

Table 17: Descriptive statistics for Standardized (Median IQR) Average Brier Scores per Teams

2. Standardized (MedianIQR)Net Brier Scores per Teams

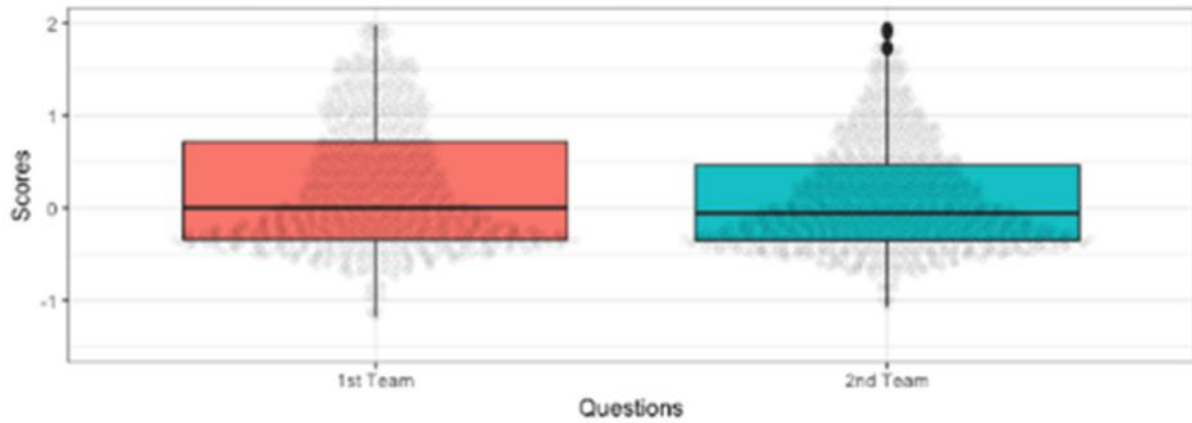


Figure 35: Standardized (MedianIQR)Net Brier Scores per Teams Boxplots

Perf Measure	Team "A"	Team "B"
min	-1.18027031	-1.0755105
max	9.00972305	8.51655865
median	0.03718775	-0.0320698
mean	0.37342261	0.1771715
SE.mean	0.03562906	0.02597273
CI.mean.0.95	0.06993328	0.0509678
var	1.05870438	0.66986067
std.dev	1.02893361	0.81845017
coef.var	2.7554133	4.61953615

Table 18: Descriptive Statistics for Standardized (MedianIQR)Net Brier Scores per Teams

3. Standardized (MedianMAD) Average Brier Scores per Teams

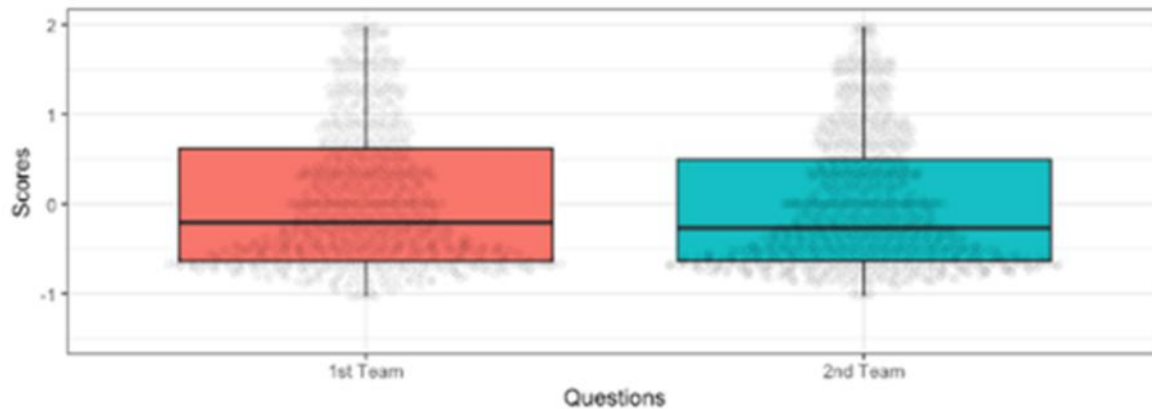


Figure 36: Standardized (MedianMAD) Average Brier Scores per Teams Boxplots

Perf Measure	Team "A"	Team "B"
min	-1.03789986	-1.03069223
max	16.98325464	16.98325464
median	0.12338928	-0.08359563
mean	0.77344142	0.40244725
SE.mean	0.06868602	0.05179935
CI.mean.0.95	0.13481801	0.10164889
var	3.93461921	2.66439062
std.dev	1.98358746	1.63229612
coef.var	2.56462532	4.05592566

Table 19: Descriptive Statistics for Standardized (MedianMAD) Average Brier Scores per Teams

4. Standardized (MediaMAD)Net Brier Scores per Teams

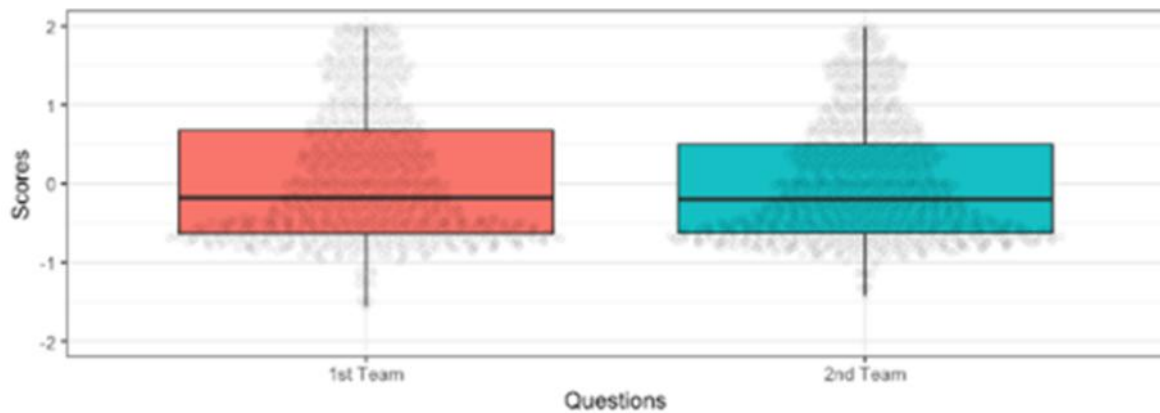


Figure 37 : Standardized (MediaMAD)Net Brier Scores per Teams Boxplots

Perf Measure	Team "A"	Team "B"
min	-1.56190342	-1.42327018
max	15.85736799	14.98938467
median	0.07551205	-0.06511975
mean	0.68922549	0.32806514
SE.mean	0.06567298	0.04583341
CI.mean.0.95	0.12890396	0.08994157
var	3.59699175	2.08599638
std.dev	1.89657369	1.44429789
coef.var	2.75174629	4.40247295

Table 20: Descriptive Statistics for Standardized (MediaMAD)Net Brier Scores per Teams

Appendix B Descriptive Statistics per Question

1. 1st Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.080	-0.126	-0.358	0.007	0.014	-0.376	-0.006	-0.011
mean	0.209	-0.003	-0.023	0.548	1.158	-0.008	0.628	1.106
SE.mean	0.041	0.038	0.107	0.172	0.364	0.115	0.198	0.349
Cl.mean.0.95	0.082	0.076	0.212	0.343	0.724	0.228	0.394	0.693
var	0.149	0.130	1.004	2.617	11.663	1.159	3.457	10.710
std.dev	0.386	0.360	1.002	1.618	3.415	1.077	1.859	3.273
coef.var	1.846	-128.834	-43.909	2.950	2.950	-128.834	2.960	2.960
Team "B"								
median	0.067	-0.123	-0.393	-0.049	-0.103	-0.367	0.008	0.014
mean	0.231	0.006	0.034	0.641	1.353	0.019	0.675	1.188
SE.mean	0.038	0.031	0.100	0.161	0.340	0.093	0.161	0.284
Cl.mean.0.95	0.076	0.062	0.198	0.320	0.675	0.185	0.320	0.563
var	0.154	0.102	1.036	2.700	12.033	0.908	2.710	8.395
std.dev	0.392	0.319	1.018	1.643	3.469	0.953	1.646	2.897
coef.var	1.696	50.745	29.572	2.564	2.564	50.745	2.439	2.439

Table 21: Descriptive Statistics for 1st Question

2. 2nd Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.731	-0.005	-0.041	0.011	0.016	-0.014	0.010	0.013
mean	0.767	0.016	0.030	0.049	0.068	0.040	0.047	0.063
SE.mean	0.067	0.051	0.129	0.069	0.097	0.131	0.091	0.120
Cl.mean.0.95	0.134	0.103	0.258	0.139	0.193	0.261	0.181	0.239
var	0.319	0.188	1.189	0.343	0.664	1.214	0.583	1.021
std.dev	0.565	0.433	1.091	0.586	0.815	1.102	0.764	1.011
coef.var	0.737	27.215	36.650	11.945	11.945	27.215	16.123	16.123
Team "B"								
median	0.720	-0.017	-0.062	0.000	0.000	-0.043	-0.010	-0.014
mean	0.737	-0.015	-0.028	0.018	0.025	-0.038	-0.007	-0.009
SE.mean	0.053	0.040	0.103	0.055	0.077	0.101	0.070	0.093
Cl.mean.0.95	0.106	0.079	0.205	0.110	0.153	0.202	0.140	0.185
var	0.230	0.129	0.857	0.247	0.479	0.834	0.400	0.701
std.dev	0.480	0.359	0.926	0.497	0.692	0.913	0.633	0.837
coef.var	0.651	-24.091	-32.721	27.853	27.853	-24.091	-90.838	-90.838

Table 22: Descriptive Statistics for 2nd Question

3. 3rd Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.245	-0.138	-0.409	0.102	0.244	-0.296	0.110	0.208
mean	0.512	0.048	0.115	0.518	1.242	0.103	0.424	0.800
SE.mean	0.065	0.060	0.128	0.102	0.243	0.130	0.102	0.192
CI.mean.0.95	0.129	0.120	0.255	0.202	0.485	0.259	0.203	0.384
var	0.312	0.269	1.207	0.762	4.385	1.246	0.769	2.741
std.dev	0.559	0.519	1.099	0.873	2.094	1.116	0.877	1.655
coef.var	1.092	10.845	9.534	1.685	1.685	10.845	2.070	2.070
Team "B"								
median	0.180	-0.219	-0.537	0.000	0.000	-0.471	-0.028	-0.052
mean	0.404	-0.040	-0.097	0.349	0.838	-0.087	0.275	0.519
SE.mean	0.049	0.044	0.096	0.077	0.184	0.095	0.074	0.140
CI.mean.0.95	0.097	0.087	0.191	0.152	0.365	0.188	0.148	0.279
var	0.211	0.171	0.817	0.516	2.966	0.789	0.487	1.736
std.dev	0.460	0.413	0.904	0.718	1.722	0.888	0.698	1.317
coef.var	1.139	-10.264	-9.325	2.055	2.055	-10.264	2.540	2.540

Table 23: Descriptive Statistics for 3rd Question

4. 4th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.180	-0.099	-0.438	0.000	0.000	-0.293	0.000	0.000
mean	0.388	0.039	0.115	0.355	0.801	0.115	0.320	0.467
SE.mean	0.048	0.044	0.129	0.083	0.186	0.131	0.103	0.151
CI.mean.0.95	0.096	0.088	0.257	0.165	0.371	0.262	0.205	0.300
var	0.171	0.144	1.211	0.498	2.534	1.261	0.775	1.656
std.dev	0.413	0.379	1.100	0.706	1.592	1.123	0.880	1.287
coef.var	1.065	9.785	9.531	1.988	1.988	9.785	2.755	2.755
Team "B"								
median	0.129	-0.101	-0.573	-0.087	-0.196	-0.300	-0.005	-0.008
mean	0.305	-0.035	-0.105	0.213	0.481	-0.105	0.148	0.216
SE.mean	0.037	0.033	0.100	0.064	0.144	0.097	0.076	0.111
CI.mean.0.95	0.075	0.065	0.199	0.127	0.287	0.193	0.151	0.221
var	0.112	0.086	0.797	0.328	1.667	0.752	0.462	0.987
std.dev	0.335	0.293	0.893	0.573	1.291	0.867	0.680	0.994
coef.var	1.099	-8.279	-8.473	2.683	2.683	-8.279	4.607	4.607

Table 24: Descriptive Statistics for 4th Question

5. 5th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.980	0.197	0.343	0.248	0.325	0.341	0.261	0.351
mean	0.935	0.154	0.269	0.205	0.268	0.267	0.217	0.292
SE.mean	0.079	0.075	0.128	0.075	0.098	0.130	0.077	0.103
Cl.mean.0.95	0.157	0.150	0.256	0.150	0.197	0.259	0.154	0.207
var	0.410	0.372	1.085	0.373	0.640	1.107	0.392	0.707
std.dev	0.640	0.610	1.041	0.611	0.800	1.052	0.626	0.841
coef.var	0.685	3.947	3.864	2.984	2.984	3.947	2.878	2.878
Team "B"								
median	0.500	-0.264	-0.437	-0.210	-0.275	-0.456	-0.212	-0.285
mean	0.617	-0.142	-0.247	-0.098	-0.129	-0.244	-0.086	-0.116
SE.mean	0.065	0.061	0.106	0.062	0.081	0.105	0.062	0.084
Cl.mean.0.95	0.130	0.121	0.211	0.124	0.162	0.209	0.124	0.167
var	0.305	0.265	0.807	0.277	0.476	0.790	0.279	0.504
std.dev	0.552	0.515	0.898	0.527	0.690	0.889	0.529	0.710
coef.var	0.895	-3.637	-3.637	-5.362	-5.362	-3.637	-6.115	-6.115

Table 25: Descriptive Statistics for 5th Question

6. 6th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.213	-0.096	-0.257	0.090	0.151	-0.256	0.088	0.146
mean	0.413	0.092	0.252	0.645	1.081	0.247	0.639	1.063
SE.mean	0.061	0.058	0.155	0.169	0.284	0.156	0.171	0.284
Cl.mean.0.95	0.122	0.116	0.310	0.338	0.567	0.312	0.341	0.568
var	0.231	0.210	1.494	1.776	4.984	1.507	1.805	5.005
std.dev	0.480	0.458	1.222	1.333	2.232	1.228	1.343	2.237
coef.var	1.164	4.971	4.848	2.066	2.066	4.971	2.104	2.104
Team "B"								
median	0.125	-0.146	-0.480	-0.153	-0.256	-0.391	-0.060	-0.099
mean	0.228	-0.080	-0.217	0.134	0.224	-0.213	0.135	0.225
SE.mean	0.033	0.031	0.083	0.090	0.151	0.082	0.090	0.150
Cl.mean.0.95	0.065	0.061	0.165	0.180	0.302	0.164	0.180	0.299
var	0.073	0.065	0.474	0.563	1.579	0.467	0.559	1.551
std.dev	0.270	0.255	0.688	0.750	1.257	0.684	0.748	1.246
coef.var	1.185	-3.205	-3.176	5.598	5.598	-3.205	5.543	5.543

Table 26: Descriptive Statistics for 6th Question

7. 7th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.405	-0.112	-0.241	0.087	0.126	-0.245	0.014	0.019
mean	0.601	0.079	0.168	0.422	0.609	0.172	0.364	0.486
SE.mean	0.069	0.067	0.145	0.118	0.171	0.146	0.123	0.164
Cl.mean.0.95	0.139	0.134	0.290	0.237	0.342	0.293	0.245	0.328
var	0.289	0.268	1.257	0.841	1.752	1.285	0.903	1.609
std.dev	0.538	0.517	1.121	0.917	1.324	1.133	0.950	1.269
coef.var	0.894	6.584	6.663	2.173	2.173	6.584	2.610	2.610
Team "B"								
median	0.320	-0.166	-0.418	-0.058	-0.084	-0.364	-0.086	-0.114
mean	0.442	-0.077	-0.165	0.149	0.216	-0.168	0.079	0.105
SE.mean	0.050	0.047	0.105	0.086	0.124	0.103	0.086	0.115
Cl.mean.0.95	0.100	0.094	0.209	0.171	0.247	0.205	0.172	0.230
var	0.169	0.147	0.734	0.491	1.023	0.707	0.497	0.886
std.dev	0.411	0.384	0.857	0.701	1.011	0.841	0.705	0.941
coef.var	0.930	-5.001	-5.192	4.688	4.688	-5.001	8.955	8.955

Table 27: Descriptive Statistics for 7th Question

8. 8th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.328	-0.162	-0.328	0.034	0.076	-0.326	0.029	0.059
mean	0.603	0.090	0.187	0.321	0.716	0.181	0.328	0.665
SE.mean	0.078	0.072	0.146	0.081	0.182	0.145	0.085	0.174
Cl.mean.0.95	0.157	0.144	0.293	0.163	0.364	0.291	0.171	0.347
var	0.355	0.300	1.242	0.385	1.917	1.221	0.424	1.746
std.dev	0.596	0.548	1.115	0.621	1.384	1.105	0.651	1.321
coef.var	0.987	6.109	5.962	1.933	1.933	6.109	1.986	1.986
Team "B"								
median	0.180	-0.287	-0.604	-0.120	-0.267	-0.580	-0.120	-0.245
mean	0.400	-0.094	-0.194	0.109	0.243	-0.190	0.109	0.222
SE.mean	0.055	0.052	0.104	0.058	0.129	0.105	0.062	0.125
Cl.mean.0.95	0.111	0.104	0.207	0.115	0.257	0.210	0.123	0.251
var	0.199	0.176	0.697	0.216	1.075	0.715	0.248	1.022
std.dev	0.446	0.419	0.835	0.465	1.037	0.846	0.498	1.011
coef.var	1.116	-4.457	-4.306	4.268	4.268	-4.457	4.550	4.550

Table 28: Descriptive Statistics for 8th Question

9. 9th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.283	-0.222	-0.444	-0.048	-0.089	-0.456	-0.051	-0.091
mean	0.552	0.039	0.082	0.298	0.549	0.080	0.310	0.561
SE.mean	0.082	0.079	0.161	0.106	0.195	0.162	0.109	0.197
Cl.mean.0.95	0.165	0.159	0.323	0.213	0.392	0.325	0.219	0.395
var	0.339	0.311	1.292	0.560	1.899	1.307	0.592	1.933
std.dev	0.582	0.558	1.137	0.748	1.378	1.143	0.769	1.390
coef.var	1.055	14.213	13.901	2.510	2.510	14.213	2.478	2.478
Team "B"								
median	0.328	-0.161	-0.356	0.010	0.018	-0.330	0.035	0.063
mean	0.478	-0.030	-0.063	0.203	0.374	-0.062	0.215	0.388
SE.mean	0.056	0.053	0.110	0.072	0.133	0.109	0.073	0.133
Cl.mean.0.95	0.112	0.106	0.219	0.144	0.266	0.218	0.147	0.265
var	0.206	0.184	0.783	0.340	1.152	0.772	0.350	1.142
std.dev	0.453	0.429	0.885	0.583	1.073	0.879	0.591	1.069
coef.var	0.949	-14.201	-14.070	2.873	2.873	-14.201	2.754	2.754

Table 29: Descriptive Statistics for 9th Question

10. 10th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.213	-0.132	-0.291	0.150	0.522	-0.288	0.156	0.509
mean	0.521	0.161	0.354	0.677	2.361	0.351	0.699	2.284
SE.mean	0.083	0.080	0.174	0.142	0.496	0.174	0.148	0.484
Cl.mean.0.95	0.167	0.161	0.350	0.286	0.999	0.351	0.298	0.974
var	0.332	0.306	1.452	0.972	11.825	1.459	1.053	11.249
std.dev	0.577	0.553	1.205	0.986	3.439	1.208	1.026	3.354
coef.var	1.107	3.444	3.406	1.456	1.456	3.444	1.469	1.469
Team "B"								
Median	0.097	-0.243	-0.533	-0.048	-0.168	-0.530	-0.049	-0.161
Mean	0.221	-0.124	-0.274	0.163	0.570	-0.271	0.170	0.556
SE.mean	0.043	0.041	0.089	0.073	0.254	0.089	0.076	0.247
Cl.mean.0.95	0.085	0.082	0.178	0.146	0.509	0.178	0.151	0.494
Var	0.113	0.103	0.493	0.330	4.015	0.491	0.355	3.787
std.dev	0.336	0.321	0.702	0.574	2.004	0.701	0.595	1.946
coef.var	1.523	-2.581	-2.564	3.516	3.516	-2.581	3.500	3.500

Table 30: Descriptive Statistics for 10th Question

11. 11th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
Median	0.245	-0.236	-0.465	-0.106	-0.169	-0.457	-0.097	-0.154
Mean	0.522	0.027	0.050	0.286	0.454	0.053	0.302	0.482
SE.mean	0.090	0.087	0.168	0.128	0.203	0.168	0.132	0.210
Cl.mean.0.95	0.182	0.175	0.338	0.257	0.409	0.338	0.265	0.422
Var	0.383	0.354	1.322	0.768	1.937	1.324	0.813	2.068
std.dev	0.619	0.595	1.150	0.877	1.392	1.151	0.902	1.438
coef.var	1.186	21.873	22.970	3.063	3.063	21.873	2.982	2.982
Team "B"								
Median	0.405	-0.088	-0.167	0.120	0.191	-0.171	0.127	0.203
Mean	0.474	-0.021	-0.039	0.219	0.347	-0.041	0.229	0.366
SE.mean	0.060	0.058	0.112	0.085	0.136	0.112	0.088	0.140
Cl.mean.0.95	0.121	0.116	0.224	0.171	0.271	0.224	0.176	0.280
Var	0.222	0.205	0.766	0.445	1.122	0.765	0.470	1.194
std.dev	0.471	0.452	0.875	0.667	1.059	0.874	0.685	1.093
coef.var	0.993	-21.573	-22.691	3.052	3.052	-21.573	2.987	2.987

Table 31: Descriptive Statistics for 11th Question

12. 12th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.720	0.052	0.090	0.233	0.339	0.093	0.247	0.374
mean	0.828	0.156	0.275	0.347	0.506	0.278	0.365	0.551
SE.mean	0.100	0.097	0.171	0.106	0.155	0.172	0.109	0.164
Cl.mean.0.95	0.202	0.195	0.344	0.214	0.311	0.347	0.219	0.331
var	0.452	0.421	1.314	0.506	1.074	1.336	0.533	1.217
std.dev	0.672	0.649	1.146	0.711	1.037	1.156	0.730	1.103
coef.var	0.811	4.156	4.176	2.047	2.047	4.156	2.002	2.002
Team "B"								
Median	0.403	-0.237	-0.451	-0.103	-0.150	-0.423	-0.078	-0.118
Mean	0.547	-0.117	-0.206	0.049	0.072	-0.209	0.057	0.087
SE.mean	0.063	0.059	0.107	0.066	0.096	0.105	0.066	0.100
Cl.mean.0.95	0.125	0.118	0.213	0.132	0.193	0.210	0.133	0.201
Var	0.235	0.209	0.682	0.263	0.558	0.663	0.265	0.605
std.dev	0.484	0.457	0.826	0.513	0.747	0.815	0.514	0.778
coef.var	0.886	-3.905	-4.012	10.400	10.400	-3.905	8.979	8.979

Table 32: Descriptive Statistics for 12th Question

13. 13th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.405	-0.120	-0.203	0.118	0.208	-0.235	0.086	0.166
mean	0.592	0.072	0.143	0.377	0.666	0.141	0.359	0.693
SE.mean	0.091	0.086	0.168	0.126	0.223	0.169	0.123	0.237
Cl.mean.0.95	0.183	0.174	0.339	0.254	0.448	0.341	0.247	0.477
var	0.387	0.349	1.333	0.747	2.330	1.348	0.707	2.635
std.dev	0.622	0.591	1.154	0.864	1.526	1.161	0.841	1.623
coef.var	1.052	8.213	8.072	2.290	2.290	8.213	2.343	2.343
Team "B"								
Median	0.320	-0.180	-0.361	0.000	0.000	-0.354	0.000	0.000
Mean	0.464	-0.047	-0.094	0.200	0.354	-0.093	0.189	0.365
SE.mean	0.059	0.055	0.110	0.082	0.145	0.109	0.079	0.152
Cl.mean.0.95	0.118	0.111	0.219	0.164	0.290	0.217	0.158	0.304
Var	0.216	0.190	0.744	0.417	1.300	0.734	0.385	1.434
std.dev	0.465	0.436	0.862	0.646	1.140	0.856	0.620	1.198
coef.var	1.002	-9.232	-9.211	3.225	3.225	-9.232	3.278	3.278

Table 33: Descriptive Statistics for 13th Question

14. 14th Question Statistics

	Avg	Net	sAvg_mean	sAvg_medianIQR	sAvg_medianMAD	sNET_mean	sNET_medianIQR	sNET_medianMAD
Team "A"								
median	0.980	0.161	0.274	0.236	0.313	0.262	0.232	0.313
mean	0.885	0.076	0.127	0.150	0.198	0.124	0.150	0.203
SE.mean	0.097	0.094	0.150	0.088	0.117	0.153	0.090	0.121
Cl.mean.0.95	0.196	0.189	0.302	0.178	0.236	0.308	0.181	0.244
var	0.425	0.397	1.010	0.351	0.616	1.048	0.363	0.660
std.dev	0.652	0.630	1.005	0.592	0.785	1.024	0.602	0.812
coef.var	0.737	8.286	7.918	3.962	3.962	8.286	4.006	4.006
Team "B"								
median1	0.520	-0.270	-0.435	-0.182	-0.241	-0.439	-0.181	-0.244
mean1	0.737	-0.060	-0.100	0.016	0.021	-0.098	0.020	0.027
SE.mean1	0.085	0.080	0.132	0.078	0.103	0.130	0.076	0.103
Cl.mean.0.951	0.171	0.160	0.264	0.155	0.206	0.260	0.153	0.206
var1	0.415	0.363	0.987	0.343	0.602	0.958	0.332	0.604
std.dev1	0.644	0.602	0.994	0.586	0.776	0.979	0.576	0.777
coef.var1	0.874	-10.039	-9.918	37.414	37.414	-10.039	28.398	28.398

Table 34: Descriptive Statistics for 14th Question

Appendix C R code for data processing

All data manipulations and analysis was performed using the 'R' statistical programming language²³. The vast amount of collected information required a robust and automated approach that would allow the researcher to provide timely and accurate results to:

- The participating forecasters, during the experimental procedure
- The academic society during the analysis of the collected information

In the following sub-appendices, and for the sake of reproducibility, we provide you with the R code being used:

C.1 Data collection per question

```
#MANIPULATE WORKSPACE
rm(list=ls()) #clear workspace
library(xlsx) # in order to be able to save into working directory as xlsx
q1_url<-"exported URL from google forms"
q1.csv<-read.csv(url(q1_url))

#rename columns
names(q1.csv)<-c("date","email","prob", "drop4", "conf", "drop6", "drop7",
"team", "drop9", "analogies", "dur")
new.q1<-q1.csv[,c(1,2,3,5,8,10)] # I made a new df including only the
columns with quantitative data
```

²³ More information on R: <https://www.r-project.org/>

```

new.q1$date<-strptime(new.q1$date,'%d/%m/%Y %H:%M:%S') # transform
time column data to typical POSIXlt
new.q1$date<- strptime(new.q1$date, format="%Y-%m-%d") # strip time from
time stamp
new.q1[is.na(new.q1)] <- 0 # I do this iot replace NAs with "0"
# Add new column in new.q1, named "outcome" with values "0" (didn't
happen) or "100" (happened) depending on the outcome of the question
new.q1$outcome<-rep(0,nrow(new.q1))
# Add new column in new.q1, named "startdate" defining the opening of the
question
new.q1$startdate<-rep(as.Date("2016-11-14"),nrow(new.q1))
# Add new column in new.q1, named "finishdate" defining the closing date
of the question (always put the next of the actual in order to make sure that
all values are included)
new.q1$finishdate<-rep(as.Date("2017-01-08"),nrow(new.q1))
# Add new column with "singlebrier" per question per user
# I use Brier 1950, which is the one that tetlock uses.[(outcome-
p)^2+(outcome'-p')^2] It actually gives me a range of values from 0 to 1
singlebrier<-((new.q1$outcome-new.q1$prob)/100)^2 + (((100-
new.q1$outcome)-(100-new.q1$prob))/100)^2
new.q1<-cbind(new.q1,singlebrier) #attach single brier column

temp<-new.q1
dates <- unique(temp[,1])
Ndates<- length(dates) #No of unique dates
idx <- unique(temp[,2]) # Unique emails (referring to 2nd column)
Nidx<- length(idx) # No of unique emails
upd.q1 <- NULL
end.date <- new.q1[1,9] # I've changed it myself here, in order to avoid re-
entering the finishdate (eg "2017-12-31")
#c.dates <-
seq.Date(from=as.Date(dates[1]),to=as.Date(dates[length(dates)]),by="1
day")

```

```

c.dates <- seq.Date(from=as.Date(dates[1]),to=as.Date(end.date),by="1
day")
Nc <- length(c.dates)
for (i in seq(Nc)) { upd.q1 <-
rbind(upd.q1,data.frame(idx,rep(c.dates[i],Nidx),rep(NA,Nidx))) }
colnames(upd.q1) <- c("idx","c-dates","singlebrier")
for (i in seq(Nidx))
{
  #cat("i = ",i,"\n")
  i.rows <- which(upd.q1$idx == idx[i])
  i.date <- upd.q1[i.rows,"c-dates"]

  old <- subset(new.q1,as.Date(new.q1$date) == i.date[1])
  sbs <- subset(old,old$email == idx[i]),"singlebrier")
  # Reversed the lines...
  if (length(sbs) > 1) { sbs <- mean(sbs) }
  if (length(sbs) > 0) { upd.q1[i.rows,"singlebrier"][1] <- sbs }

  for (j in seq(2,length(i.date)))
  {
    old <- subset(new.q1,as.Date(new.q1$date) == i.date[j])
    sbs <- subset(old,old$email == idx[i]),"singlebrier")
    # Reversed the lines here too
    if (length(sbs) ==0) { upd.q1[i.rows,"singlebrier"][j] <-
upd.q1[i.rows,"singlebrier"][j-1] }
    if (length(sbs) > 1) { sbs <- mean(sbs) }
    if (length(sbs) > 0) { upd.q1[i.rows,"singlebrier"][j] <- sbs }
  }
}
# Keep the NA indices
na.idx <- which(is.na(upd.q1[, "singlebrier"]))
# INITIALIZE
ALL <- NULL
# Renaming

```

```

Ndates <- length(c.dates)
temp <- upd.q1

#COMPUTATIONS IN NESTED LOOPS
#I use function seq(from=, to=, by=, ), where seq=sequence
for (i in seq(Ndates)){
  temp.i <- subset(temp,temp[,"c-dates"] == c.dates[i])
  store.i <- matrix(0,nrow=Nidx,ncol=1)
  rownames(store.i) <- idx
  for (j in seq(Nidx))
  {
    forc.ij <- subset(temp.i,temp.i[,"idx"] == idx[j]) # By using Dr T's
method, I do not get the lenght issue, given I subset data.frame temp.i
    store.i[j,1] <- forc.ij[,"singlebrier"]
  }

  # Compute the average of the day and the net Brier
  avg <- mean(store.i,na.rm=TRUE)
  net.i <- store.i - avg
  date.i <- data.frame(rep(c.dates[i],Nidx),idx,store.i,net.i)
  colnames(date.i) <- c("Date","ID","SBS","Net")
  ALL <- rbind(ALL,date.i)
}

q1.all<-ALL
q1.all<-q1.all[,c(2,3,4)]

q1.all[na.idx,3] <- 0
# change the row names
rownames(q1.all) <- NULL

```

```

avg.brier <- matrix(0,nrow=Nidx,ncol=2)
rownames(avg.brier) <- idx
colnames(avg.brier) <- c("Avg","Net")
for (i in seq(Nidx))
{
  qi <- subset(q1.all,q1.all[,"ID"] == idx[i])
  avg.brier[i,] <- c(mean(qi[,"SBS"],na.rm=TRUE),mean(qi[,"Net"]))
}
df.avg.brier <- data.frame(idx,avg.brier,row.names=NULL)

# I will add a 4th column with question ID (character). eg 1st, 2nd etc
df.avg.brier$qlD<-rep("1st",nrow(df.avg.brier))

# I will import a new dataframe with the team of each forecaster (it's in
popurateR googlesheet, named "teams")
teams<-"https://docs.google.com/spreadsheets/d/e/2PACX-
1vTecDsfKw3SwPYIx9FswYzlLoryFv3lctv7CA1Q3lOuQ3TEW71eAP8HF5H23SDK_
EbPr--9a68PYs/pub?gid=0&single=true&output=csv"
teams.csv<-read.csv(url(teams))
# Workaround
## Step 1: Remove white spaces in idx column [I can spot them by using
paste(df.avg.brier$idx)]
df.avg.brier$idx<-trimws(df.avg.brier$idx)
## Step 2: Replace uppercase characters with lowercase
df.avg.brier$idx<-tolower(df.avg.brier$idx)
## Step 3: Use pmatch function, to avoid the "removing NAs" issue
team.idx <- match(df.avg.brier[,"idx"],teams.csv[,"idx"]) #will give the row
index in the teams.csv where the emails of df.avg.brier match the emails in
the teams.csv
teams.csv <- teams.csv[team.idx,"team"] #will give you the actual teams.
Then, merge as:
df.avg.brier <- data.frame(df.avg.brier,teams.csv)
colnames(df.avg.brier) <- c("idx", "Avg", "Net", "qlD" , "teams")
# Save
library(googlesheets)

```

```

all_my_sheets_in_drive <- gs_ls() # read google drive contents
gs <- gs_title("populateR") # load specific googlesheet from google drive (has
to be there IOT recall it)
#if I use head(gs) i can get the sheet key, plus some other info
gs_ws <- gs_ws_new(gs, ws_title = "q1sheetv2") # creates a new sheet (TAB), by
the name "q1sheet" in the google sheet "populateR"
gs <- gs_title("populateR")
# upload dataframe to worksheet "q1sheet" on google cloud in order to pass
it by URL to the forecasting application (http://c-the-future.boards.net/)
gs_edit_cells(gs, ws="q1sheetv2", input = df.avg.brier, trim = TRUE)

```

C.2 Data standardization

```

rm(list=ls()) #clear workspace
#library(xlsx) replaced it due to errors. See line 555
#write.xlsx(df.name,file = "name.xlsx")
library(google Sheets)
library(dplyr)
gs<-gs_ls() #Returns a data frame of the sheets you would see in your Google
Sheets home screen.
gs_auth()
populateR<-gs_title("populateR")
gs_ws_ls(populateR)

q1 <- as.data.frame(gs_read(ss=populateR, ws = "q1sheetv2"))
Sys.sleep(10)
q2 <- as.data.frame(gs_read(ss=populateR, ws = "q2sheetv2"))
Sys.sleep(10)
.

```

```

.
.
q14 <- as.data.frame(gs_read(ss=populateR, ws = "q14sheetv2"))

q.all<-rbind(q1,q2,q3,q4,q5,q6,q7,q8,q9,q10,q11,q12,q13,q14)
q.all$teams <- factor(q.all$teams,
                      labels = c("1st Team", "2nd Team"))

##### Data standardization #####
standardize.data <- function(x,s.type=c("mean","median","yard"),y.value=0)
{ # Different types of standardization
  if (s.type == "mean") { y <- (x-mean(x))/sd(x) }
  if (s.type == "median/IQR") { y <- (x-median(x))/IQR(x) }
  if (s.type == "median/MAD") { y <- (x-median(x))/mad(x) } # I added this
  one (I believe this is the default type of median stand/ion)
  if (s.type == "yard") { y <- x - y.value }
  return(y)
}

# I start standardizing per question #
s.q1<- cbind(q1,standardize.data(q1$Avg,s.type="mean"),
             standardize.data(q1$Avg,s.type="median/IQR"),
             standardize.data(q1$Avg,s.type="median/MAD"),
             standardize.data(q1$Avg,s.type="yard",y.value=0),
             standardize.data(q1$Net,s.type="mean"),
             standardize.data(q1$Net,s.type="median/IQR"),
             standardize.data(q1$Net,s.type="median/MAD"),
             standardize.data(q1$Net,s.type="yard",y.value=-2))

s.q2<- cbind(q2,standardize.data(q2$Avg,s.type="mean"),
             standardize.data(q2$Avg,s.type="median/IQR"),
             standardize.data(q2$Avg,s.type="median/MAD"),
             standardize.data(q2$Avg,s.type="yard",y.value=0),
             standardize.data(q2$Net,s.type="mean"),

```

```

standardize.data(q2$Net,s.type="median/IQR"),
standardize.data(q2$Net,s.type="median/MAD"),
standardize.data(q2$Net,s.type="yard",y.value=-2))
.
.
.
.
s.q14<- cbind(q14,standardize.data(q14$Avg,s.type="mean"),
              standardize.data(q14$Avg,s.type="median/IQR"),
              standardize.data(q14$Avg,s.type="median/MAD"),
              standardize.data(q14$Avg,s.type="yard",y.value=0),
              standardize.data(q14$Net,s.type="mean"),
              standardize.data(q14$Net,s.type="median/IQR"),
              standardize.data(q14$Net,s.type="median/MAD"),
              standardize.data(q14$Net,s.type="yard",y.value=-2))

# Rename all the new columns
dfs <- c("s.q1", "s.q2","s.q3", "s.q4","s.q5", "s.q6","s.q7", "s.q8","s.q9",
"s.q10","s.q11","s.q12","s.q13","s.q14")
for(df in dfs) {
  df.tmp <- get(df)
  names(df.tmp) <- c("idx", "Avg", "Net", "qID", "teams",
"sAvg_mean","sAvg_medianIQR","sAvg_medianMAD","sAvg_yard",
"sNET_mean","sNET_medianIQR","sNET_medianMAD","sNET_yard")
  assign(df, df.tmp)
}
# Bind them all together in one dataframe
s.q.all<-
rbind(s.q1,s.q2,s.q3,s.q4,s.q5,s.q6,s.q7,s.q8,s.q9,s.q10,s.q11,s.q12,s.q13,s.q14)
s.q.all$teams <- factor(s.q.all$teams,

```



```

        labels = c("1st Team", "2nd Team")) # I renamed the entries in
column "team"
s.q.all <- na.omit(s.q.all) # I drop NAs from q.all (intruders not allocated in
teams)
# Create data frames per team
s.q.all.t1<-subset(s.q.all, s.q.all$teams=="1st Team")
s.q.all.t1<- s.q.all.t1[,-5] # removed team column
s.q.all.t2<-subset(s.q.all, s.q.all$teams=="2nd Team")
s.q.all.t2<- s.q.all.t2[,-5] # removed team column

##### I start standardizing per team #####

q.all.s.t1<-subset(q.all, q.all$teams=="1st Team")
q.all.s.t1<- q.all.s.t1[,-5] # removed team column
q.all.s.t2<-subset(q.all, q.all$teams=="1st Team")
q.all.s.t2<- q.all.s.t2[,-5] # removed team column

q.all.s.t1<- cbind(q.all.s.t1,standardize.data(q.all.s.t1$Avg,s.type="mean"),
  standardize.data(q.all.s.t1$Avg,s.type="median/IQR"),
  standardize.data(q.all.s.t1$Avg,s.type="median/MAD"),
  standardize.data(q.all.s.t1$Avg,s.type="yard",y.value=0),
  standardize.data(q.all.s.t1$Net,s.type="mean"),
  standardize.data(q.all.s.t1$Net,s.type="median/IQR"),
  standardize.data(q.all.s.t1$Net,s.type="median/MAD"),
  standardize.data(q.all.s.t1$Net,s.type="yard",y.value=-2))

q.all.s.t2<- cbind(q.all.s.t2,standardize.data(q.all.s.t2$Avg,s.type="mean"),
  standardize.data(q.all.s.t2$Avg,s.type="median/IQR"),
  standardize.data(q.all.s.t2$Avg,s.type="median/MAD"),
  standardize.data(q.all.s.t2$Avg,s.type="yard",y.value=0),
  standardize.data(q.all.s.t2$Net,s.type="mean"),
  standardize.data(q.all.s.t2$Net,s.type="median/IQR"),
  standardize.data(q.all.s.t2$Net,s.type="median/MAD"),
  standardize.data(q.all.s.t2$Net,s.type="yard",y.value=-2))

```

```

# Rename all the new columns
dfs <- c("q.all.s.t1", "q.all.s.t2")
for(df in dfs) {
  df.tmp <- get(df)
  names(df.tmp) <- c("idx", "Avg", "Net", "qID",
"sAvg_mean","sAvg_medianIQR","sAvg_medianMAD","sAvg_yard",
"sNET_mean","sNET_medianIQR","sNET_medianMAD","sNET_yard")
  assign(df, df.tmp)
}

```

C.3 Stochastic Dominance calculations

```

source("SD thom.R")
# I produce the stochastic dominance results for each and every one of the
standardized values
#Avg
x1 <- s.q.all.t1[,"Avg"]
x2 <- s.q.all.t2[,"Avg"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the H0: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#

```

```

for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

  if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
  if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

  xxb <- cbind(x1b,x2b)
  #out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
  pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
  res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
Avg<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put it within
the format function.

#Net
x1 <- s.q.all.t1[,"Net"]
x2 <- s.q.all.t2[,"Net"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

```

```

if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
Net<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put it within
the format function.
Net

#sAvg_mean
x1 <- s.q.all.t1[,"sAvg_mean"]
x2 <- s.q.all.t2[,"sAvg_mean"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{

```

```

cat("Now doing resample",br,"\n")

if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientif notation
sAvg.mean<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put
it within the format function.
sAvg.mean

#sAvg_medianIQR
x1 <- s.q.all.t1[,"sAvg_medianIQR"]
x2 <- s.q.all.t2[,"sAvg_medianIQR"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

```

```

if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
sAvg.median.IQR<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",")
# I put it within the format function.
sAvg.median.IQR

#sAvg_medianMAD
x1 <- s.q.all.t1[,"sAvg_medianMAD"]
x2 <- s.q.all.t2[,"sAvg_medianMAD"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

```

```

if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
sAvg.medianMAD<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",")
# I put it within the format function.
sAvg.medianMAD

#sAvg_yard
x1 <- s.q.all.t1[,"sAvg_yard"]
x2 <- s.q.all.t2[,"sAvg_yard"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

  if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }

```

```

if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
sAvg.yard<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put it
within the format function.
sAvg.yard

#sNet_mean
x1 <- s.q.all.t1[,"sNET_mean"]
x2 <- s.q.all.t2[,"sNET_mean"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

```



```

if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
sNet.mean<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put
it within the format function.
sNet.mean

#sNet_medianIQR
x1 <- s.q.all.t1[,"sNET_medianIQR"]
x2 <- s.q.all.t2[,"sNET_medianIQR"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

  if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
  if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

```

```

    xxb <- cbind(x1b,x2b)
    #out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
    pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
    res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientif notation
sNet.medianIQR<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") #
I put it within the format function.
sNet.medianIQR

#sNet_medianMAD
x1 <- s.q.all.t1[,"sNET_medianMAD"]
x2 <- s.q.all.t2[,"sNET_medianMAD"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
    cat("Now doing resample",br,"\n")

    if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
    if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }
}

```

```

xxb <- cbind(x1b,x2b)
#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientif notation
sNet.medianMAD<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",")
# I put it within the format function.
sNet.medianMAD

#sNet_yard
x1 <- s.q.all.t1[,"sNET_yard"]
x2 <- s.q.all.t2[,"sNET_yard"]
N1 <- length(x1)
N2 <- length(x2)
NN <- min(c(N1,N2))
BR <- 100 # Number of iterations for bootstrap samples. Should take values from
100 to 500. At this point I set it to 100
BP <- 20 # Botstrap for P values for each of the produced blocks. Should take
values from 20 to 50. At this point I set it to 20
bench <- 2 # It defines the No: T2 does not dominate over T1 in terms of
performance.
res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the results
colnames(res) <- c("1SD","2SD", "3SD")
#
for (br in seq(BR))
{
  cat("Now doing resample",br,"\n")

  if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
  if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

  xxb <- cbind(x1b,x2b)

```

```

#out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
res[br,] <- pvl$p.values[,1]
}
# apply(res,2,mean) # I change it in order to avoid scientific notation
sNet.yard<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I put it
within the format function.
sNet.yard

#Bind all stochastic dominance results
all.sd.p<-rbind(Avg, sAvg.mean, sAvg.median.IQR, sAvg.medianMAD,
sAvg.yard, Net, sNet.mean, sNet.medianIQR, sNet.medianMAD, sNet.yard )
all.sd.p<- as.data.frame(all.sd.p)

library("WriteXLS",
lib.loc="/Library/Frameworks/R.framework/Versions/3.4/Resources/library")
WriteXLS(all.sd.p, ExcelFileName = "all.sd.p.xls", SheetNames = NULL, perl =
"perl",
+ verbose = FALSE, Encoding = c("UTF-8", "latin1", "cp1252"),
+ row.names = TRUE, col.names = TRUE,
+ AdjWidth = FALSE, AutoFilter = FALSE, BoldHeaderRow = FALSE,
+ na = "",
+ FreezeRow = 0, FreezeCol = 0,
+ envir = parent.frame())
# For more info on the command arguments "https://cran.r-
project.org/web/packages/WriteXLS/WriteXLS.pdf"

#Stochastic dominance per question
source("SD per question 3SD.R")
# 1st question
x<-"Avg"

```

```

y<-"1st"
Q1Avg<-SD.per.Q(x,y)
x<-"sAvg_mean"
y<-"1st"
Q1sAvg.mean<-SD.per.Q(x,y)
x<-"sAvg_medianIQR"
y<-"1st"
Q1sAvg.medianIQR<-SD.per.Q(x,y)
x<-"sAvg_medianMAD"
y<-"1st"
Q1sAvg.medianMAD<-SD.per.Q(x,y)
x<-"sAvg_yard"
y<-"1st"
Q1sAvg.yard<-SD.per.Q(x,y)
x<-"Net"
y<-"1st"
Q1Net<-SD.per.Q(x,y)
x<-"sNET_mean"
y<-"1st"
Q1sNet.mean<-SD.per.Q(x,y)
x<-"sNET_medianIQR"
y<-"1st"
Q1sNet.medianIQR<-SD.per.Q(x,y)
x<-"sNET_medianMAD"
y<-"1st"
Q1sNet.medianMAD<-SD.per.Q(x,y)
x<-"sNET_yard"
y<-"1st"
Q1sNet.yard<-SD.per.Q(x,y)

# Put all Q scores together
q1.sd.p<-rbind(Q1Avg, Q1sAvg.mean, Q1sAvg.medianIQR,
Q1sAvg.medianMAD, Q1sAvg.yard, Q1Net, Q1sNet.mean,
Q1sNet.medianIQR, Q1sNet.medianMAD, Q1sNet.yard )

```

```

q1.sd.p<- as.data.frame(q1.sd.p)
write.csv(q1.sd.p, file = "q1.sd.p.csv") # need to open it with excel and change
ti to xls (with data validation)

.
.
.
.
# 14th question
x<-"Avg"
y<-"14th"
Q14Avg<-SD.per.Q(x,y)
x<-"sAvg_mean"
y<-"14th"
Q14sAvg.mean<-SD.per.Q(x,y)
x<-"sAvg_medianIQR"
y<-"14th"
Q14sAvg.medianIQR<-SD.per.Q(x,y)
x<-"sAvg_medianMAD"
y<-"14th"
Q14sAvg.medianMAD<-SD.per.Q(x,y)
x<-"sAvg_yard"
y<-"14th"
Q14sAvg.yard<-SD.per.Q(x,y)
x<-"Net"
y<-"14th"
Q14Net<-SD.per.Q(x,y)
x<-"sNET_mean"
y<-"14th"
Q14sNet.mean<-SD.per.Q(x,y)
x<-"sNET_medianIQR"

```

```

y<-"14th"
Q14sNet.medianIQR<-SD.per.Q(x,y)
x<-"sNET_medianMAD"
y<-"14th"
Q14sNet.medianMAD<-SD.per.Q(x,y)
x<-"sNET_yard"
y<-"14th"
Q14sNet.yard<-SD.per.Q(x,y)

# Put all Q scores together
q14.sd.p<-rbind(Q14Avg, Q14sAvg.mean, Q14sAvg.medianIQR,
Q14sAvg.medianMAD, Q14sAvg.yard, Q14Net, Q14sNet.mean,
Q14sNet.medianIQR, Q14sNet.medianMAD, Q14sNet.yard )
q14.sd.p<- as.data.frame(q14.sd.p)
write.csv(q14.sd.p, file = "q14.sd.p.csv")

```

C.3.1 Source code for 'SD thom' function

```

# Make a grid of values for a vector or matrix
make.grid <- function(x,nout=1000)
{
  # Get actual endpoints
  xmin <- min(x)
  xmax <- max(x)
  # Compute the grid endpoints
  if (xmin < 0) { xmin <- 1.618*xmin }
  else { xmin <- 0.618*xmin }
  if (xmax > 0) { xmax <- 1.618*xmax }
  else { xmax <- 0.618*xmax }
  # Make the grid
  grid <- seq(xmin,xmax,length.out=nout)
  # Return

```

```

    return(grid)
}
# Given a matrix of data, a benchmark variable for comparison and
# the orders of stochastic dominance, compute the required statistics
stochastic.dominance <- function(x,bmark,maxs=1,...)
{
  # Get the dimensions of the data matrix
  N <- nrow(x)
  K <- ncol(x)
  # Get the grid of values for evaluating the distribution function
  grid <- make.grid(x,...)
  # Length of grid
  ng <- length(grid)
  # Extract the benchmark variable and put the rest in another variable
  u <- x[,bmark]
  y <- as.matrix(x[,-bmark])
  # Initialize storage of the two tests at each order s
  tests <- matrix(0,nrow=maxs,ncol=2)
  # Outer loop over the dominance order
  for (s in seq(1,maxs,1))
  {
    # For each s compute the distribution of the benchmark
    # First replicate the benchmark data into a matrix
    umat <- matrix(u,nrow=N,ncol=ng)
    # Then put the grid also into a matrix
    gmat <- t(matrix(grid,nrow=ng,ncol=N))
    # Compute directly the statistic over the domain of u
    Du <- (umat <= gmat)*((gmat - umat)^(s-1))
    Du <- apply(Du,2,sum)/(N*factorial(s-1))
    # Initialize storage of the statistics of the free variables

```



```

Dstore <- matrix(0,nrow=ng,ncol=K-1)
# Inner loop over the free variables
for (j in seq(1,K-1,1))
{
  # Now do the same for each of the rest of the variables
  jmat <- matrix(y[,j],nrow=N,ncol=ng)
  # Compute directly the statistic over the domain of y[,j]
  Dj <- (jmat <= gmat)*((gmat - jmat)^(s-1))
  Dj <- apply(Dj,2,sum)/(N*factorial(s-1))
  # Store them
  Dstore[,j] <- Dj
  # Done with inner loop!
}
# Now compute the various statistics
# Replicate Du into a matrix
Dumat <- matrix(Du,nrow=ng,ncol=K-1)
# Compute the component that applies to the tests
# and note that the benchmark goes in first
intest <- apply(sqrt(N)*(Dumat-Dstore),2,max)
#print(intest)
# The benchmark dominates all others at order s
test1 <- max(intest)
# The benchmark dominates at least one of the others at order s
test2 <- min(intest)
# Save the tests
tests[s,] <- c(test1,test2)
}
# Return the tests
return(tests)
}
# Subsampling computation of the p-values for testing stochastic dominance
get.pvalues <- function(x,bmark,maxs=1,b,...)
{

```

```

# Get dimensions of data set
N <- nrow(x)
K <- ncol(x)
# Use default number of blocks?
if (b <= 0) { b <- 10*sqrt(N) }
# Number of subsamples
Nb <- N-b+1
# Storage
store1 <- matrix(0,nrow=Nb,ncol=maxs)
store2 <- matrix(0,nrow=Nb,ncol=maxs)
# Now roll over the blocks
for (i in seq(1,Nb,1))
{
  # Extract the ith block
  xi <- x[seq(i,b+i-1,1),]
  # Get the statistics
  fi <- stochastic.dominance(xi,bmark,maxs,...)
  # Extract and store the results
  store1[i,] <- fi[,1]
  store2[i,] <- fi[,2]
}
# Get the full sample statistics
f <- stochastic.dominance(x,bmark,maxs,...)
f1 <- f[,1]
f2 <- f[,2]
# Now compute the appropriate p-values
p1 <- apply(store1 > t(matrix(f1,nrow=maxs,ncol=Nb)),2,mean)
p2 <- apply(store2 > t(matrix(f2,nrow=maxs,ncol=Nb)),2,mean)
# Return everything
return(list(tests=f,p.values=cbind(p1,p2)))

```

```
}
```

C.3.2 Source code for 'SD per question 3SD' function

```
SD.per.Q<-function(x,y) { # x is type of score, y is No of question
  x1<-subset(s.q.all.t1,s.q.all.t1$qID==y)
  x1 <- x1[,x]
  x2<-subset(s.q.all.t2,s.q.all.t2$qID==y)
  x2 <- x2[,x]
  N1 <- length(x1)
  N2 <- length(x2)
  NN <- min(c(N1,N2))
  BR <- 100 # Number of iterations for bootstrap samples. Should take
values from 100 to 500. At this point I set it to 100
  BP <- 20 # Botstrap for P values for each of the produced blocks. Should
take values from 20 to 50. At this point I set it to 20
  bench <- 2 # It defines the H0: T2 does not dominate over T1 in terms of
performance.
  res <- matrix(0,nrow=BR,ncol=3) # The matrix wihtin which we will put the
results
  colnames(res) <- c("1SD","2SD", "3SD")
  #
  for (br in seq(BR))
  {
    cat("Now doing resample",br,"\n")

    if (N1 < N2) { x1b <- x1; x2b <- x2[sample.int(N2,N1,replace=TRUE)] }
    if (N2 < N1) { x1b <- x1[sample(N2)]; x2b <- x2 }

    xxb <- cbind(x1b,x2b)
    #out <- stochastic.dominance(xxb,bench,maxs=2,nout=floor(NN/4))
    pvl <- get.pvalues(xxb,bench,maxs=3,b=BP,nout=floor(NN/4))
    res[br,] <- pvl$p.values[,1]
  }
}
```

```

    # apply(res,2,mean) # I change it in order to avoid scientific notation
    outcome<-format(apply(res,2,mean),scientific = FALSE,big.mark = ",") # I
put it within the format function.
    print(outcome)
}
#x<-"Avg"
#y<- "1st"
#print(per.question(x,y))

```

C.4 Comparative performance of forecasters

```

xxx <- s.q.all
# We run the code per team and save the produced results
#xxx <- subset(xxx,xxx[,"teams"] == "1st Team")
xxx <- subset(xxx,xxx[,"teams"] == "2nd Team")
xxx
perc <- 0.25 # <--
qid <- c("1st","2nd","3rd",paste(seq(4,14),"th",sep=""))
rank.id <- "sAvg_mean" # <--
uid <- unique(xxx[,"idx"])
Nfirst <- 6 # <--
zi <- matrix(NA,nrow=length(uid),ncol=Nfirst)
rownames(zi) <- uid
for (i in seq(Nfirst))
{
  xi <- subset(xxx,xxx[,"qid"] == qid[i])
  mi <- na.omit(match(xi[,"idx"],uid))
  zi[mi,i] <- xi[,rank.id]
}

```

```

zi <- as.matrix(apply(zi,1,function(foo) { ifelse(length(na.omit(foo)) >=
5,mean(foo,na.rm=TRUE),NA) })))
zs <- zi[order(zi[,1]),,drop=FALSE]
ps <- floor(NROW(zs)*perc)
topp <- rownames(zs)[1:ps]
contains <- matrix(0,nrow=ps,ncol=14)
rownames(contains) <- topp
for (i in seq(14))
{
  xi <- subset(xxx,xxx[,"qid"] == qid[i])
  mi <- na.omit(match(topp,xi[,"idx"]))
  contains[na.omit(match(xi[mi,"idx"],topp)),i] <- 1
}
pp <- apply(contains,1,function(foo) { ifelse(sum(foo) >= 11,sum(foo)/14,NA)}
)*100
pp<- na.omit(as.data.frame(pp))
pp$names <- rownames(pp) #transfor row names into column and add it as
column at the end
pp<-pp[,c(2,1)] #reorder columns
zi <- matrix(NA,nrow=NROW(contains),ncol=14)
rownames(zi) <- rownames(contains)
for (i in seq(14))
{
  xi <- subset(xxx,xxx[,"qid"] == qid[i])
  mi <- na.omit(match(rownames(contains),xi[,"idx"]))
  zi[na.omit(match(xi[mi,"idx"],topp)),i] <- xi[mi,rank.id]
}
id11 <- apply(zi,1,function(foo) { length(na.omit(foo)) >= 11})
zi <- as.matrix(apply(zi[id11,,drop=FALSE],1,mean,na.rm=TRUE))
zi<-as.data.frame(zi)
zi$names <- rownames(zi) #transfor row names into column and add it as
column at the end
zi<-zi[,c(2,1)] #reorder columns
DrT_relult<-as.data.frame(c(pp,zi))

```

```
DrT_relult<-DrT_relult[,c(1,2,4)]  
names(DrT_relult)<-c('idx','Qualifying Percentage','Metric Score')
```

Appendix D PESCO Projects Stratification

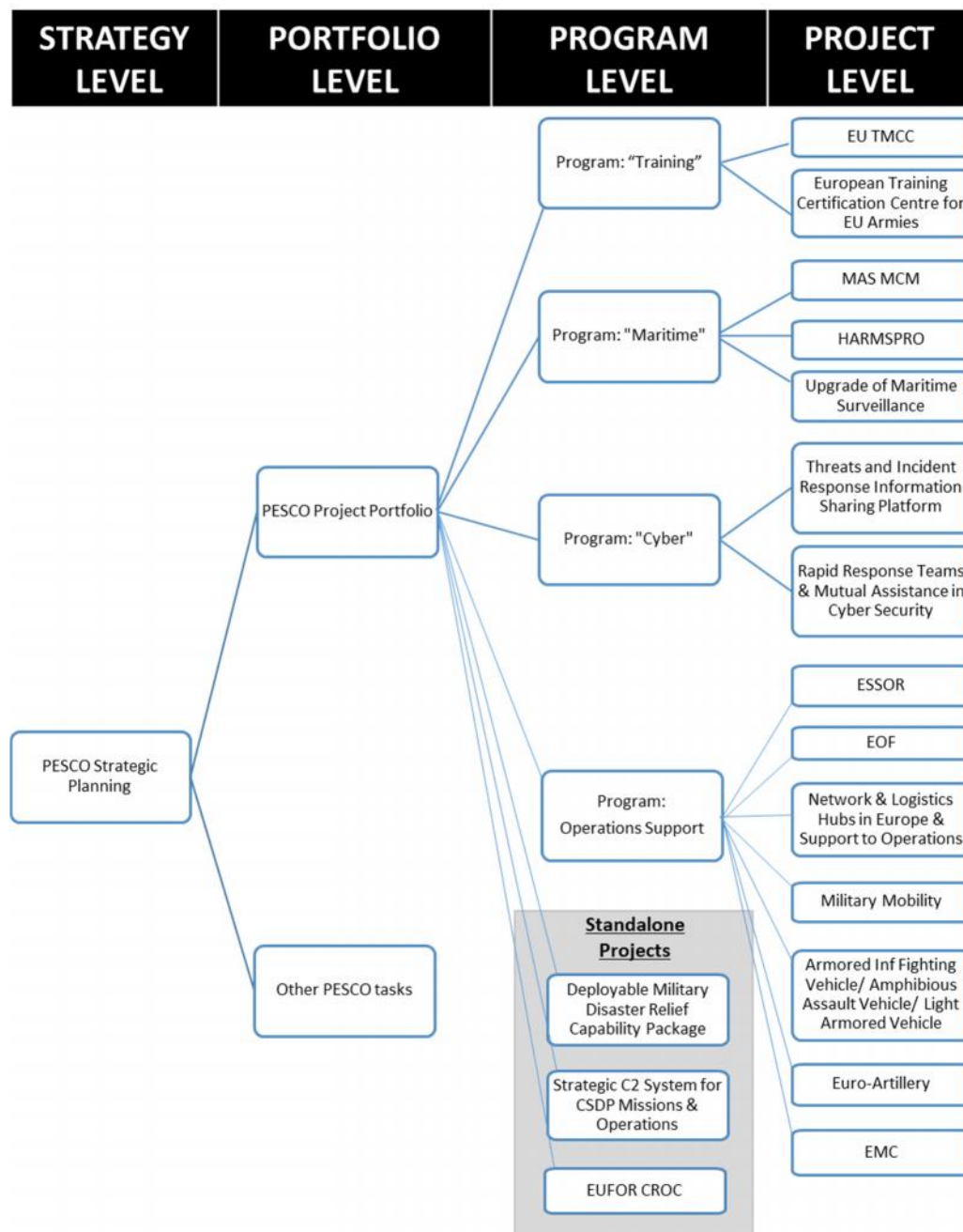


Figure 38: PESCO Project Stratification

D.1 Short description of the above PESCO projects

(Content herein is unclassified and retrieved from European's Council Official website (European Council, n.d.))

European Medical Command

"The European Medical Command (EMC) will provide the EU with an enduring medical capability to support missions and operations on the ground. The project is expected to make progress the interoperability and the coherence of health care capabilities in Europe (standardization of concepts, training and certification)."

European Secure Software defined Radio (ESSOR)

"The European Secure Software Defined Radio aims to develop common technologies for European military radios. The adoption of these technologies as a standard will guarantee the interoperability of EU forces in the framework of joint operations, regardless which radio platforms are used, thereby reinforcing the European strategic autonomy. It is expected to enhance logistic planning and movement as well as to deliver common standards and procedures, that will greatly improve the EU's and NATO's capability to conduct even the most demanding missions."

Military Mobility

"This project will support Member States' commitment to simplify and standardize cross-border military transport procedures. The project should help to reduce barriers such as legal hurdles to cross-border movement, lingering bureaucratic requirements (such as passport checks at some border crossings) and infrastructure problems, like roads and bridges that cannot accommodate large military vehicles."

European Union Training Mission Competence Centre (EU TMCC)

"The European Union Training Mission Competence Centre (EU TMCC) will improve the availability, interoperability, specific skills and professionalism of personnel (trainers) for EU training missions across participating Member States. Moreover, it will accelerate the provision for EU training missions due to a higher situational awareness regarding trained, educated and available personnel for current and future EU training missions."

European Training Certification Centre for European Armies

"The European Training Certification Centre for European Armies aims to promote the standardisation of procedures among European Armies and enable the staff, up to Division level, to practice the entire spectrum of the command and control (C2) functions at land, joint and interagency levels in a simulated training environment."

Energy Operational Function (EOF)

"The Energy Operational Function" aims to develop new systems of energy supply for camps deployed in the framework of joint operations and for soldier connected devices and equipment while ensuring that the energy issue is taken into account from the conceiving of combat systems to the implementation of the support in operations.

Deployable Military Disaster Relief Capability Package

"The Deployable Military Disaster Relief Capability Package will deliver a multi-national specialist military package for the assistance to EU and other States, which can be deployed within both EU-led and non EU-led operations. The new EU capability will manage a range of emergencies including natural disasters, civil emergencies, and pandemics. The project aims to include the establishment of a new EU Disaster Relief Training Centre of Excellence, and ultimately a Disaster Relief Deployable Headquarters."

Maritime (semi-) Autonomous Systems for Mine Countermeasures (MAS MCM)

"The Maritime (semi-) Autonomous Systems for Mine Countermeasures (MAS MCM) will deliver a world-class mix of (semi-) autonomous underwater, surface and aerial technologies for maritime mine countermeasures. The project will enable Member States to protect maritime vessels, harbours and off shore installations, and to safeguard freedom of navigation on maritime trading

routes.

Harbour & Maritime Surveillance and Protection (HARMSPRO)

"The Harbour & Maritime Surveillance and Protection (HARMSPRO) will deliver a new maritime capability which will provide Member States with the ability to conduct surveillance and protection of specified maritime areas, from harbours up to littoral waters, including sea line of communications and choke points, in order to obtain security and safety of maritime traffic and structures.

Upgrade of Maritime Surveillance

"The project on Upgrade of Maritime Surveillance will integrate land-based surveillance systems, maritime and air platforms in order to distribute real-time information to Member States, so as to provide timely and effective response in the international waters.

Cyber Threats and Incident Response Information Sharing Platform

"Cyber Threats and Incident Response Information Sharing Platform will develop active defence measures, potentially moving from firewalls to more active measures. This project aims to mitigate risks by focusing on the sharing of cyber threat intelligence through a networked Member State platform

Strategic Command and Control (C2) System for CSDP Missions and Operations

"The project aims to improve the command and control systems of EU missions and operations at the strategic level. Once implemented, the project will enhance the military decision-making process, improve the planning and conduct of missions, and the coordination of EU forces. "

Armoured Infantry Fighting Vehicle / Amphibious Assault Vehicle / Light Armoured Vehicle

"The project will develop and build a prototype European Armoured Infantry Fighting Vehicle / Amphibious Assault Vehicle / Light Armoured Vehicle. The vehicles would be based on a common platform and would support fast deployment manoeuvre, reconnaissance, combat support, logistics support, command and control, and medical support. "

Indirect Fire Support (Euro-Artillery)

"The Indirect Fire Support (Euro-Artillery) will develop a mobile precision artillery platform, which would contribute to the EU's combat capability requirement in military operations. This project aims at procuring a new capability / platform of a key mission component for land forces in the short to mid-term."

EUFOR Crisis Response Operation Core (EUFOR CROC)

"EUFOR Crisis Response Operation Core (EUFOR CROC) will decisively contribute to the creation of a coherent full spectrum force package, which could accelerate the provision of forces. EUFOR CROC will improve the crisis management capabilities of the EU. In phase 1 the project will start with an implementation study."

Appendix E Demographics Survey

Δημογραφικά Superforecasters

*Required

Χώρα προέλευσης *

Ελλάδα

Κύπρος

Other: _____

Φύλο *

Choose ▾

Διεύθυνση Ηλεκτρονικού Ταχυδρομείου (e-mail) *

Η διεύθυνση email που θα καταχωρήσετε σε αυτό το πεδίο, θα αποτελεί και το αναγνωριστικό σας στην εφαρμογή "Superforecasters", την οποία θα χρησιμοποιούμε για να καταχωρούμε τις προβλέψεις μας. Για να έχετε απόλυτη ανωνυμία, χρησιμοποιήστε ή φτιάξτε ένα λογαριασμό e-mail ο οποίος δε θα παραπέμπει στην ταυτότητά σας και θα έχει την παρακάτω μορφή "γράμμα u ή w+6 νούμερα @domain name" (πχ u542698@gmail.com ή w784599@yahoo.gr κ.ο.κ.)

Your answer

Εκπαιδεύσεις - Πrouπηρεσία

Παρούσα κατάσταση

Επιλέξτε όσα ισχύουν

- Προπτυχιακός φοιτητής
- Μεταπτυχιακός Φοιτητής
- Διδακτορικός Φοιτητής
- Σπουδαστής ΕΣΔΔΑ
- Εργαζόμενος στον ευρύτερο Δημόσιο τομέα
- Εργαζόμενος στον ευρύτερο Ιδιωτικό τομέα
- Άνεργος
- Συνταξιούχος
- Other: _____

Στην περίπτωση που είστε φοιτητής-τρια, επιλέξτε το πανεπιστήμιό σας

Choose

Γράψτε το Τμήμα του Πανεπιστημίου στο οποίο φοιτάτε

Your answer

Γράψτε το έτος φοίτησης

Your answer

Χρόνος προπηρεσίας σε καθήκοντα αντίστοιχα με αυτά που σας έχουν ανατεθεί την παρούσα περίοδο

Αφορά μόνο στους εργαζόμενους.

Choose

Εκπαιδεύσεις

Ακαδημαϊκές Εκπαιδεύσεις (επιλέξτε τον υψηλότερο κατεχόμενο τίτλο σπουδών) *

ΠΕ: Πανεπιστημιακής Εκπαίδευσης, ΤΕ: Τεχνικής Εκπαίδευσης

Choose

Οι σπουδές μου στην ανωτέρω σχολή περιλάμβαναν στο ωρολόγιο πρόγραμμα και το μάθημα της Στατιστικής Ανάλυσης ή/και Πιθανοθεωρίας

Συμπληρώνεται μόνο από αυτούς που αφορά

Choose

Έχω φοιτήσει στην ΕΣΔΔΑ

Συμπληρώνεται μόνο από αυτούς που αφορά

Choose

Επίπεδα γλωσσομάθειας

1. Επίπεδο Α
 - A1 «Στοιχειώδης γνώση της ξένης γλώσσας»
 - A2 «Βασική γνώση της ξένης γλώσσας»
2. Επίπεδο Β
 - B1 «Μέτρια γνώση της ξένης γλώσσας»
 - B2 «Καλή γνώση της ξένης γλώσσας»
3. Επίπεδο Γ
 - Γ1 «Πολύ καλή γνώση της ξένης γλώσσας»
 - Γ2 «Άριστη γνώση της ξένης γλώσσας»

Επίπεδο γλωσσομάθειας Αγγλικών

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Επίπεδο γλωσσομάθειας Γαλλικών

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Επίπεδο γλωσσομάθειας Γερμανικών

1: Στοιχειώδης γνώση (A1), 2: Βασική Γνώση (A2)

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Επίπεδο γλωσσομάθειας Τουρκικών

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Επίπεδο γλωσσομάθειας Αραβικών

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Επίπεδο γλωσσομάθειας Ρωσικών

	1	2	3	
Στοιχειώδης ή βασική γνώση	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Πολύ καλή ή άριστη γνώση

Άλλες ξένες γλώσσες

Your answer

Προσωπική ενημέρωση

Πόσο συχνά ενημερώνεστε επί θεμάτων διεθνούς πολιτικής και εξωτερικών σχέσεων; *

- Καθημερινά
- Σε εβδομαδιαία βάση
- Σε μηνιαία βάση
- Πιο σπάνια
- Δεν με αφορά το συγκεκριμένο θέμα και δεν ενημερώνομαι σχετικά με αυτό.

Από ποιες πηγές ενημερώνεστε;

Η παρούσα ερώτηση δεν συμπληρώνεται από αυτούς επέλεξαν "Δεν με αφορά...." στην προηγούμενη ερώτηση. (Μπορείτε να επιλέξετε πλέον της μίας απάντησης)

- Έντυπο περιοδικό τύπο
- Ηλεκτρονικό περιοδικό τύπο
- Ανεξάρτητους ιστότοπους (πχ blogs, προσωπικούς ιστότοπους κλπ)
- Facebook ή αντίστοιχους ιστότοπους κοινωνικής δικτύωσης
- Other:

Ενημερώνομαι κατά κανόνα από έντυπα ή ηλεκτρονικά μέσα στα:

- Ελληνικά
- Αγγλικά
- Γαλλικά
- Γερμανικά
- Ρωσικά
- Αραβικά
- Other:

Πιστεύω ότι πιο σημαντικό ρόλο στην παροχή, κατά το δυνατόν, σωστών εκτιμήσεων σε ερωτήσεις γεωπολιτικού ενδιαφέροντος, παίζει η: *

- Γενική παιδεία
- Στοχευμένη εξειδικευμένη γνώση στο ειδικό αντικείμενο που πραγματεύεται η κάθε ερώτηση.

Appendix F Forecasting question example

10η Ερώτηση

Η ερώτηση θα παραμείνει ενεργή μέχρι την 15 Μαϊ 17

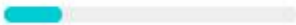
*Required

Διεύθυνση Ηλεκτρονικού Ταχυδρομείου (e-mail) *

Η διεύθυνση email που θα καταχωρήσετε σε αυτό το πεδίο, αποτελεί και το αναγνωριστικό σας στο πρόγραμμα "Superforecasters". Παρακαλούμε βεβαιωθείτε ότι είναι ορθά συμπληρωμένη. Ως επιβεβαίωση θα λάβετε και σχετικό email με την καταχώρηση της απάντησής σας.

Your answer

NEXT

 Page 1 of 5

Never submit passwords through Google Forms.

Ερώτηση

Μπορείτε να την απαντήσετε όσες φορές επιθυμείτε, ανανεώνοντας με αυτό τον τρόπο και επικαιροποιώντας την προηγούμενη εκτίμησή σας.

Με την ολοκλήρωση της συμπλήρωσης της παρούσας φόρμας θα ερωτηθείτε πόσο χρόνο καταναλώσατε. Οπότε κοιτάξτε το ρολόι σας τώρα.

Θα περάσει η τιμή του bitcoin τα 1.000 ευρώ μέχρι την 15 Μαΐ 17?

Ως σημείο αναφοράς για την επιβεβαίωση του γεγονότος, θα ληφθούν οι τιμές που αναγράφονται στον ακόλουθο ιστότοπο:

<http://bitcoinity.org/markets/kraken/EUR?theme=light>

Για περισσότερες πληροφορίες σχετικά με το bitcoin:

<https://en.wikipedia.org/wiki/Bitcoin>



Προσδιορίστε την εκτίμησή σας με τη μορφή ποσοστού *

Μπορείτε να ορίσετε τιμή από 0% (απόλυτα σίγουρος ότι δε θα γίνει το γεγονός), έως 100% (απόλυτα σίγουρος ότι θα γίνει.) Αναγράψτε μόνο το αριθμητικό μέρος της πιθανότητας, πχ 65

Your answer

Τεκμηριώστε την απάντησή σας. *

Αναλόγως σε ποια ομάδα εργασίας ανήκετε (με ή χωρίς Δομημένες Αναλογίες), προβείτε και σε ανάλογη τεκμηρίωση, συμφώνως με τις οδηγίες που σας έχουν δοθεί.

Your answer

Προσδιορίστε την προσωπική σας εμπιστοσύνη στην παραπάνω εκτίμηση (confidence to judgement) *

Η εμπιστοσύνη προκύπτει από την απάντηση στις παρακάτω ερωτήσεις: (1) Νιώθω ότι ήταν επαρκείς οι πληροφορίες που συνέλεξα? (2) Νιώθω ότι ανέλυσα σωστά και επαρκώς και τις παραπάνω πληροφορίες?

1 2 3 4 5

Χαμηλή εμπιστοσύνη

Υψηλή εμπιστοσύνη

Αναγράψτε τις διευθύνσεις (URLs) των ιστότοπων που επισκεφτήκατε κατά την αναζήτηση πληροφοριών

Μην αναγράφετε τους ιστότοπους τους οποίους, αν και επισκεφτήκατε, δεν αξιοποιήσατε κατά τη συλλογή πληροφοριών. Διαχωρίστε τις διάφορες διευθύνσεις μεταξύ τους με (,)

Your answer

Αναγράψτε συμπληρωματικές πηγές που χρησιμοποιήσατε (όχι URLs) κατά την αναζήτηση πληροφοριών

Διαχωρίστε και εδώ τις διάφορες πηγές με (,)

Your answer

BACK

NEXT

Page 2 of 5

Επιλέξτε σε ποιά ομάδα ανήκετε

Επιλέξτε τον ανάλογο κωδικό της ομάδας σας.

1. Ελεύθερη εργασία κατά μόνας (unaided personal work)
2. Ατομική εργασία με τη χρήση "Δομημένων Αναλογιών"
3. Ελεύθερη ομαδική εργασία (unaided team work)
4. Ομαδική εργασία με τη χρήση "Δομημένων Αναλογιών"

*

- 1
- 2
- 3
- 4

BACK

NEXT

Page 3 of 5

Whoever choses options 2 and 4 has the below additional question:

Δομημένες Αναλογίες

Επιλέξτε τα βήματα που ακολουθήσατε κατά τη χρήση της μεθόδου των Δομημένων Αναλογιών *

Επιλέξτε μόνο όσα υλοποιήσατε

- Αποδόμηση ερώτησης σε μικρότερα τμήματα
- Αναγνώριση και ανάλυση ανάλογων καταστάσεων
- Βαθμολόγηση ανάλογων καταστάσεων από πλευράς ομοιοτήτων-διαφορών με την αντίστοιχη υπό εξέταση
- Παροχή εκτίμησης
- Τις δομημένες αναλογίες τις χρησιμοποίησα μόνο για την αρχική μου πρόβλεψη, και τώρα απλά ανανεώνω την εκτίμησή μου σύμφωνα με τις νέες πληροφορίες.

Πόσες ανάλογες καταστάσεις αναγνώρισατε?

Choose ▾

BACK

NEXT

Page 4 of 5

Χρόνος απάντησης ερώτησης

Πόσο χρόνο αναλώσατε, από τη στιγμή που διαβάσατε την ερώτηση, μέχρι τώρα?

Choose ▾

BACK

SUBMIT

Page 5 of 5