Superforecasting reality check: evidence from a small pool of experts and expedited identification

Ilias Katsagounos, Dimitrios D. Thomakos
University of Peloponnese, Greece
Konstantia Litsiou
Manchester Metropolitan University, UK
Konstantinos Nikolopoulos*+
Durham University, UK

Abstract

Superforecasting has drawn the attention of academics - despite earlier contradictory findings in the literature, arguing that humans can consistently and successfully forecast over long periods. It has also enthused practitioners, due to the major implications for improving forecast-driven decision-making. The evidence in support of the superforecasting hypothesis was provided via a 4-year project led by Tetlock and Mellers, which was based on an exhaustive experiment with more than 5000 experts across the globe, resulting in identifying 260 superforecasters. The result, however, jeopardizes the applicability of the proposition, as exciting as it may be for the academic world; if every company in the world needs to rely on the aforementioned 260 experts, then this will end up an impractical and expensive endeavor. Thus, it would make sense to test the superforecasting hypothesis in real-life conditions: when only a small pool of experts is available, and there is limited time to identify the superforecasters. If under these constrained conditions the hypothesis still holds, then many small and medium-sized organizations could identify fast and consequently utilize their own superforecasters. In this study, we provide supportive empirical evidence from an experiment with an initial (small) pool of 314 experts and an identification phase of (just) 9 months. Furthermore - and corroborating to the superforecasting literature, we also find preliminary evidence that even an additional training of just 20 minutes, can influence positively the number of superforecasters identified.

Keywords: Forecasting; Superforecasting; Real-life conditions; Experts; Training;

* Corresponding author: kostas.nikolopoulos@durham.ac.uk
+ Professor Nikolopoulos thanks Professor Barbara Mellers (Wharton) for her critical feedback in the initial conception and write up of this article; he also thanks Professor Scott Armstrong (Wharton) and Professor Kesten Green (University of South Australia) for continuously sharing their thoughts and recommendations throughout this project, Professor Vasilis Assimakopoulos (NTUA) for his support, and Professor Fotios Petropoulos (Bath) for the ideas and advice he shared with the authoring team. Professor Nikolopoulos also thanks the audience in the following seminars for their feedback and influential questions and suggestions: Wisconsin School of Business at UW Madison, Gies College of Business at the University of Illinois at Urbana Champaign, Driehaus College of Business at DePaul University, University College London (QFF), Alliance Manchester Business School and Bangor Business School. Finally, the authors thank Professor Ruud Teunter, the Associate editor and two anonymous reviewers for their very constructive feedback during the review process of this article.

1. Introduction and motivation

Superforecasting (Tetlock & Gardner, 2015), has drawn the attention of both academics and practitioners. Academics were fascinated to see evidence that specific individuals can consistently over long periods of time, outperform other humans in forecasting for very difficult tasks, despite contradicting earlier findings in the literature. Practitioners were also very keen to follow this line of thought, due to the major implications for improving tactical forecast-driven decision-making. The evidence in support of the superforecasting hypothesis was provided via a large research project[1] led by Tetlock and Mellers, which was based on an exhaustive experiment over 4 years with more than 5000 experts across the globe, resulting in identifying 260 superforecasters.

These superforecasters managed to produce very difficult – mostly geopolitical – forecasts, better than anybody in the world for almost half a decade. This was an unexpected result as forecasting is not an easy task at all (Makridakis, Hogarth, & Gaba, 2010), and it is something that some rules may apply (Petropoulos, Markidakis, Assimakopoulos, & Nikolopoulos, 2014), but there is no universal solution, despite being one of the most – if not the most –ubiquitous scientific (sub-)discipline (Nikolopoulos, 2020).

Especially when it comes to needing to provide forecasts in the absence of hard data, needing to rely upon experts[2] to provide judgmental forecasts (Goodwin, Gönül & Önkal, 2017) or judgmental adjustments (De Baets & Harvey, 2020; Rekik, Glock & Syntetos, 2017; Syntetos, Kholidasari & Naim, 2016), usually for one-off events in the not-so-close future, the task becomes even more challenging (Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015; Savio & Nikolopoulos, 2013). And the task seems to have been tantalizing academia forever: Plutarch "On the 'E' at Delphi" (Plutarch, 1936 translation, p.231), noted:

> *'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.'*

[1] https://en.wikipedia.org/wiki/The_Good_Judgment_Project

[2] We do use the terms 'experts' and experiment 'participants' interchangeably in this study, as it is common practice in the superforecasting literature. We do acknowledge the literature on generalists versus specialists (Teodoridis, Bikard, & Vakili, 2018) and the difficulty to really assign the title 'expert' in an experiment participant; but in this context, the participants of the geopolitical forecasting experiments had either expertise in the context, or in the methods and skills needed to perform the task, and as such a certain level of expertise can be assumed.

The one main finding from Tetlock and Mellers' research project is that although superforecasters do exist, they are a rarity. It takes a lot of time and a very big initial pool of experts in order to identify them - all and all a few hundred across the globe. That end result, however, jeopardizes the applicability of the proposition, as exciting as it may be for the academic word. If each and every company in this world needs to rely on the aforementioned superforecasters, then this will end up being a very expensive and constrained endeavor for most interested parties.

Thus, it would make sense to test the superforecasting hypothesis in real-life conditions where the scarcity of resources is profound. What if – as what is the case very often – we have to select from a small pool of experts, and with limited time to identify the superforecasters: so a few hundred of experts, and an identification timeframe of just a few months in order to find the real superforecasters. If under these constrained conditions the hypothesis still holds, then many reasonably-sized organizations could identify fast and consequently utilize their own superforecasters, in order to produce forecasts to inform challenging tactical decisions. This is the main motivation for our study: *a superforecasting reality check*. A check that would inform theory, but more emphatically, it will release the full potential of Tetlock and Mellers' proportion for practitioners.

To that end, we present the empirical results from an experiment with an initial pool of 314 experts and an identification phase that lasted 9 months. Furthermore, we investigate the impact of increasing the amount of training provided to the experts, in creating superforecasters, via testing if a short 20-minute additional training (on top of the standard 40-minute training prescribed in the original project of Tetlock and Mellers), can help us find more superforecasters. This of course coming with all the caveats of the 'training' literature, that training does neither necessarily leads to 'learning', neither can guarantee that the taught methods (even if learned) have been put in practice during the aforementioned experiment.

The rest of the paper is structured as follows: in the next section, we provide a focused literature review, while section three presents the setup of our experiment. Section four provides the empirical results while in the last section we present the main conclusions, implications for theory and practice and a roadmap for future research.

2. Background Literature

Our study adopts a targeted in-depth literature review approach. The study focuses on a judgmental forecasting experiment. Within that literature, the focus remains in the very recent superforecasting literature. It also touches lightly on the impact of training, yet again within that expert-forecaster identification framework. Thus our literature follows the same sequence: a) judgmental forecasting, b) superforecasting, c) training (within superforecasting). For a broader literature review on forecasting in an Operational Research (OR_ context - that the audience of this journal's main interests lies, the reader can follow either Perera, Hurley, Fahimnia, & Reisi (2019) or Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos (2016).

Judgmental forecasting

For a thorough review of judgmental forecasting techniques across all application areas, the reader can revisit the seminal work of Lawrence, Goodwin, O'Connor, & Önkal (2006), or the more recent review of Arvan, Fahimnia, Reisi, & Siemsen (2019) that is more focused on OR applications and on the integration of human judgment into quantitative forecasting methods.

Over the years there have been many studies that compare the accuracy of judgmental and statistical forecasting on a case by case basis, with varying outcomes (Carbone & Gorr, 1985; Lawrence, Edmundson, & O'Connor, 1985; O'Connor, Remus, & Griggs, 1993; Sanders, 1992). In a corporate environment, there has been substantial evidence that expert judgment is important for companies' decision-making (Fildes & Goodwin, 2007), either in the form of adjustments to statistical forecasts (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009) or as pure forecasts (Franses & Legerstee, 2010). However, as helpful as judgmental approaches may often be, their relative effectiveness is entangled with several limitations, the most salient of which is the forecaster's inherent biases (Tversky & Kahneman, 1974; Makridakis, Wheelwright, & Hyndman, 1998). As a result, forecasters often inadequate forecasts, furthermore fail to acknowledge their poor performance, been surprised once they face their own true forecasting limits (Makridakis, Hogarth, & Gaba, 2010).

Superforecasting

The more recent and exciting research findings in the field of judgmental forecasting come from the superforecasting experiment (Tetlock & Gardner, 2015), a project initially called the 'Good Judgment Project' (GJP) sponsored by the Intelligence Advanced Research Programs Activity (IARPA), which took the form of a series of geopolitical forecasting tournaments (Tetlock & Gardner, 2015; Mellers, Stone, Atanasov, Rohrbaugh, Metz, Ungar, Bishop, & Horowitz, 2015; Tetlock, Mellers, Rohrbaugh, & Chen, 2014; Ungar et al., 2012).

The project was led by Tetlock and Mellers based at the Wharton School in the University of Pennsylvania and was based on an exhaustive forecasting experiment over 4 years, involving more than 5000 experts (voluntarily enrolling) across the world. An incentive was given as an annual honorarium of $150 ($250 for experts retained from the previous year) on the condition that 1/3 of forecasts were provided by experts (1/2 for the retained superforecasters).

The outcome after this lengthy multi-year process, was the identification of 260 superforecasters, and their imperative individual performance was attributed to: (a) cognitive abilities and styles, (b) task-specific skills, (c) motivation and commitment, and (d) enriched environments (Mellers, Stone, Murray, Minster, Rohrbaugh, Bishop, Chen, Baker, Hou, Horowitz, Ungar, & Tetlock, 2015).

Furthermore, key organizational aspects that led to the success of the whole experiment, were considered to be: (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 (Tetlock, Mellers, Rohrbaugh, & Chen, 2014).

Despite the recent results in favor of the superforecasting hypothesis, there has been over the years extensive evidence that humans do not forecast accurately in the long-run: from the more amusing experiments with the extraordinary – but yet so true – performance of primates picking portfolios (Malkiel, 1973, p.24): "*A blindfolded monkey throwing darts at a newspaper's financial pages could select a portfolio that would do just as well as one carefully selected by experts*". In fact, many practitioners take this argument even further[3]: The monkeys have done a much better job than both the experts and the stock market.". Even Tetlock himself, in his earlier studies, has presented empirical evidence against his very own recent success, from forecasting experiments with economists (Tetlock, 2005). In these earlier studies, experts are portrayed as no better at making long-term predictions than most people. Furthermore, he pinpoints a lack of accountability as a critical contributor to the propagation of bad forecasts (Gardner, 2011; Tetlock, 2005).

It is also worth noting that the recent evidence provided by Tetlock, Mellers and their team, came from an extensive multi-year project involving thousands of experts and a vast amount of dedicated resources, almost a supernatural experiment, that does not reflect the situation for real-life organizations, even for large multinational companies. As such a clear gap exists in the literature, as if a smaller scale experiment where there is less time to identify the experts and a much smaller initial pool of experts, would still render enough evidence in favor of the superforecasting hypothesis.

[3] https://www.forbes.com/sites/rickferri/2012/12/20/any-monkey-can-beat-the-market/#5c54d45b630a

Training

Chang, Chen, & Mellers, (2016) summarize the impact of various training practices on judgmental forecasting accuracy as a) didactic, b) process-based, c) feedback-based, and d) format-based. The above techniques are not all equally effective, but their combination can lead to enhanced levels of performance by mitigating biases. The specific training adopted during the GJP, by and large, lasts 40 minutes and includes the following four topics: introduction to biases and de-biasing techniques, basic statistics and probabilistic reasoning and introductory Bayesian analysis (Chang, Chen, & Mellers, 2016; Dhami, Mandel, Mellers, & Tetlock, 2015).

The emphasis on the latter topics comes from the fact that the human mind most of the time works in intuitive mode (Lakoff & Johnson, 1999), humans have a natural tendency to use heuristics in cognitive tasks (Kahneman, 2011; Tversky & Kahneman, 1974). Heuristics are commonly defined as rules of thumb. However, humans also are susceptible to systematic errors known as biases (Montibeller & von Winterfeldt, 2015; Kahneman, 2011). There is a lot of discourse in the literature on how to manage these biases (Liu, Vlaev, Fang, Denrell, & Chater, 2017; Dolan, Hallsworth, Halpern, King, Metcalfe, & Vlaev, 2012) as if left unattended can be quite costly for any decision-making process (Arkes, 1991).

The intention of training in any context, moreover within this superforecasting exercise, is to lead to learning and consequently use in-practice of the acquired knowledge. However, training comes with well know caveats, most notably that raining does not imply learning (Antonacopoulou, 1999). Furthermore, it does not guarantee that trained experts do apply in-practice what they have learned (Camp, 2012). Thus, claiming that experts used a specific method they have just been taught, could be perceived as a 'leap', that usually only with post-experiment questionnaires can be confirmed; and even then, the evidence is subjective, and comes from the very own testimony of the participants, so biased yet again. Nevertheless, the evidence from the superorecasting experiment is that short-training in that context did help in identifying superforecasters (Tetlock and Gardner, 2015), and as such the aforementioned 'leap' is one the authors feel comfortable to accept, defend and explore further and defend in this study.

One useful area of judgmental forecasting, that was not included in the training of the original superforecasting experiment, is that of forecasting by analogies (Armstrong, 2001). Analogical ability is intrinsic to human cognition (Reisberg, Gentner, & Smith, 2013). Analogical reasoning uses what is known about one case to infer new information about another (Gentner & Smith, 2012; Khong, 1992; Gentner, 1983). The importance of the ability to use relational similarity in cognitive tasks has often been emphasized (Gentner & Goldin-Meadow, 2003; Penn, Holyoak, & Povinelli, 2008), moreover has led to significant scientific discoveries, as is the case for Faraday (Tweney, 1991) and Kepler (Gentner, 2002).

A promising extension of this line of research was proposed by Green and Armstrong (2007) that formulated the structured analogies (SA) approach, which combines the positive aspects of analogic reasoning while minimizing potential biases via the use of an objective administrator, initially applied for conflict resolution. Savio and Nikolopoulos (2010, 2013), relaxed the need for an administrator in SA and proposed a simpler version (s-SA) with successful applications in environmental strategic making (Savio & Nikolopoulos, 2013), digital planning strategies (Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015; Litsa, Petropoulos, & Nikolopoulos, 2012), and forecasting the success of megaprojects (Litsiou, Polychronakis, Karami, & Nikolopoulos, 2020).

Given that forecasting by analogies can be delivered as short-training, even within 15-30 minutes, and it is intuitively appealing to practitioners, we do consider this a gap in the superforecasting literature.

Research Questions

In light of the above literature review and the identified gaps, we form our research questions as follows:

> **RQ1**: If we have a small pool of experts and limited time for the identification process, can we still find evidence supporting the superforecasting hypothesis?
>
> **RQ2**: Under the constraints set in RQ1, does an extra short-training, focusing on structured judgmental forecasting approaches, lead to identifying more superforecasters?

3. Methodology

The primary aim of our study is to explore if we can identify superforecasters in a constrained real-life business environment; constrained in terms of sample size and time. This primary aim is depicted by RQ1. A secondary aim is to investigate if via providing additional training (versus the original experiment of Tetlock and Mellers), we can identify more superforecasters – this is depicted in RQ2.

Of the two research questions, the first one could have been posed as a more formal statistical hypothesis, however, given it questions a small sample situation, we do consider it is best to be left and explored as a more loose research question. The second research question is, in fact, dependent to the first one, as it builds on the same constraints in terms of identification time and size of the pool of experts. Nevertheless, it investigates an additional treatment, that of extra training: this latter quest is by nature difficult to establish with a more positivistic approach, as training does not necessarily lead to learning, nonetheless the respective use of taught techniques.

We set up a forecasting tournament following the organization and practices of the GJP (as described in the 2015 book of Tetlock and Gardner). Our study was conducted between November 2016 and July 2017. Following the exact practices as in GJP, resulted in a study fully consistent with the original project of Tetlock and Mellers, plus the methods and metrics used in the analysis have already been peer-reviewed and scrutinized in the numerous published studies the GJP team (Tetlock et al., 2014; Tetlock & Gardner, 2015; Ungar et al., 2012).

Participants were asked to submit their forecasts for a variety of time-bound questions using a custom-designed web interface with Google forms. The forecasts collected were in the form of experts' "subjective probability" for events about to happen (or not), also known as "belief probability" or "personalist probability" (Hacking, 2001). This kind of forecast expresses a personal belief concerning the likelihood of an outcome and primarily relates to single events rather than repeated ones. One difference between the experimental procedure described in this paper and the GJP is that our participants were incentivized to answer almost all the questions, as fewer responses would have produced a lot of 'missing values' in an already much smaller initial sample, thus creating analysis challenges (Merkle, Steyvers, Mellers, & Tetlock, 2016).

Experts 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. Some key demographic characteristics of the pool of our experts are provided below:

Characteristic	Value	Comment
Number of experts	314 initially registered **195** *fully-engaged*	64% retention rate *Fully-engaged* are the ones that answered 5 out of 6 questions in the *identification* phase and 6 out of 8 questions in the *confirmation* phase
Origin of participants	EU	All from the same EU country
Gender	63.6% males 36.4% females	Out of 195
Sample stratification	67.7% academia 32.3% industry	'Industry' refers to both the public and private sector
Number of respondents per question	100–130	The number of respondents per question was variable, peaking during the first 3 weeks (160), and then varying from 100–130
Number of questions	14 6 in the *identification* phase 8 in the *confirmation* phase	Open for 2–6 months *Identification* phase: top-performing forecasters identified as potential superforecasters *Confirmation* phase: the initially identified top-performers, confirmed as **superforecasters** (via continuing to be top-performers)
Total number of responses	2,100	Forecasts registered in the system

Table 1. Key statistics of our experiment

Anonymity was preserved throughout the experiment, as experts have been provided with unique IDs and emails, thus limiting the potential for one expert to influence another; experts have been advised neither to contacting other experts nor to seek any information rather than through the internet. All participants had to fill in a survey with standard demographic information. Experts were randomly allocated at the start of the project in one of two groups: a) **Group A** with the exact same training as in GJP, and b) **Group B** that received an extra short-training in structured judgmental forecasting approaches.

From the initial pool of 314 experts, 195 (64%) were fully-engaged till the end of our study, 86 in Group A (n_A= 86) and 109 in group B (n_B= 109). For one expert to be classified as fully-engaged in the experiment, and have their forecasts included in our analysis, the expert should produce at least 5 out of the 6 forecasts in the identification phase (when we rank the performance and spot our potential superforecasters), and at least 6 out of the 8 forecasts in the *confirmation* phase (when we confirm who are - if any - our superforecasters).

Questions, scoring, and feedback

The fourteen questions used in our study have been created by the authoring team to be of geopolitical nature and of similar difficulty with the ones in GJP. In addition, for each question, a clear way to define the outcome of the forecasts was prescribed at the time the question was posed and was respectively communicated to the participants, in order to prevent potential "ex post facto" disputes (Mellers et al., 2015). For reference, one of the fourteen required forecasts is provided hereafter:

> *Question* (required forecast): Will the United States of America submit an official request by May 30th, 2017, to withdraw from the United Nations Framework Convention on Climate Change (UNFCCC)?
>
> *Verification of outcome*: According to Article 25 of the 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 the submission of the withdrawal via UN's official site: https://treaties.un.org/Pages/CNs.aspx?cnTab=tab1&clang=_en.

The questions provided had a forecast horizon of between two and six months, and all participants were instructed to provide an initial forecast within the first ten days, and to update it subsequently if they feel such a need arrives, based on information flow.

The participants were asked to provide their forecast as a probability estimate (0%-100%) in increments of '1' (Tetlock & Gardner, 2015). The traditional Brier score was used to measure forecasting accuracy (Brier, 1950). The Brier Score measures the squared deviations between probabilistic point forecasts and actual outcomes. A principal characteristic of the Brier score is that extremely wrong forecasts are heavily penalized. The actual outcomes have a binary expression: "0" if the event in question did not occur, "1" if the event took place. All participants received Brier scores and were ranked accordingly, receiving a score for each day a question was active, starting from the time they placed their first point forecast (Horowitz, Brandon, Stewart, Tingley, Bishop, Resnick-Samotin, Roberts, Chang, Mellers, & Tetlock, 2019). Furthermore, the average net Brier points for each question was calculated, in order to take into account the performance of other forecasters at the time each forecast was placed. We averaged all the Brier scores per day (benchmark Brier) for all the active participants for a question, and then subtract them from the forecaster's daily score; the same scoring method was used in GJP. For both these metrics, we calculated also the standardized versions of them.

The above scores were available to each participant, and they could also check their own ranking comparing to all the other participants who provided an answer to a specific question. This approach follows the 'outcome accountability' practice (Chang, Atanasov, Patil, Mellers, & Tetlock, 2017), where, within the framework of a geopolitical tournament, "accountable forecasters perform better than their non-accountable counterparts, in terms of forecasting accuracy'', and thus can boost the chances of finding evidence of the superforecasting hypothesis.

Furthermore, it should be emphasized that if forecasters were allowed to provide their estimates in an unconstrained numerical form (as discussed in a more recent study by Friedman, Baker, Mellers, Tetlock, & Zeckhauser (2018)), this can lead to skewed results, since it can be substantially impacted by grey or black swans that lead to enormous forecasting errors (Schoemaker & Tetlock, 2016). Our questions, however, were not related to any extreme events.

Incentive

The incentive offered to participants was to attend free of charge an intensive training course towards the Project Management Professional (PMI/PMP©) certification. The course had a fee of 650.00€ in 2017 and as such the monetary equivalent of our incentive to participants was a very significant one. To qualify for the free enrollment in the course, participants should be fully-engaged to the experiment, thus had to make a forecast for at least 11 out of the 14 questions in the two phases, at least 5 during the identification phase and at least 6 during the confirmation phase.

Training

The training design is detailed in Table 2:

Training Subject	**Duration**	**Group A**	**Group B**
The world of biases	8′	√	√
De-biasing techniques	12′	√	√
Basic statistics & probabilistic reasoning	11′	√	√
Practical Bayesian thinking	9′	√	√
Forecast decomposition	12′		√
Structured analogies and their applications	8′		√
		40′	60′

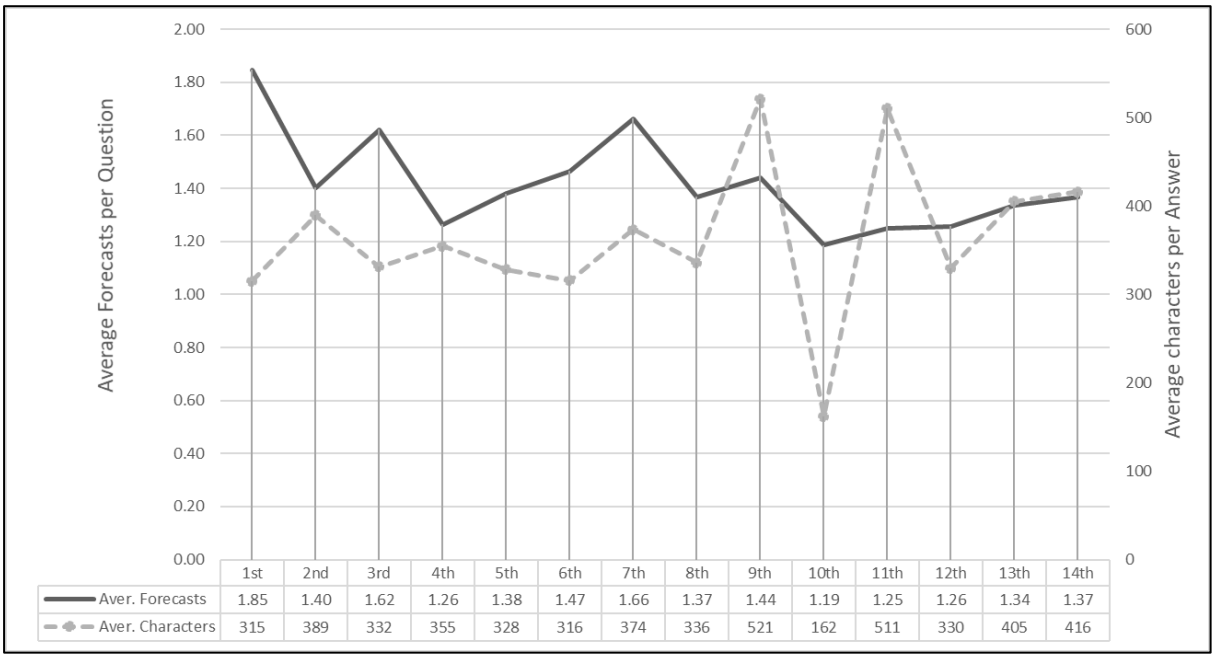
Table 2. Training per group

The new training material introduced for Group B, is focusing on: a) forecast decomposition (Armstrong, 2001), and b) Structured Analogies (Green & Armstrong, 2007) and the simplified version of the latter, semi-structured analogies (Savio & Nikolopoulos, 2010). Training effects were expected to last for the full 9 months of our experiment (Tetlock and Gardner, 2015). An example of the effectiveness of similar training in a corporate environment is provided by Hernandez (2017), where the forecasting performance improvements at TWITCH are discussed.

4. Empirical findings

The participant's engagement was similar to that of the GJP. The average forecast update per participant per question was 1.42; in the GJP, the corresponding frequency was 1.49 (Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2015). We have requested that justification should be provided for every forecast; the corresponding character count was relatively high, averaging 363 characters per person per question. The corresponding request in the GJP was much looser, that there should be at least one 50-word comment at some point during the calendar year (Mellers et al., 2015).

In figure 1 we illustrate the evolution (x-axis presents the sequence of the 14 questions) of the forecast update per participant per question (left vertical axis) and the character count per justification per participant per question (right vertical axis). A close look at the graph shows there is a change after the middle of the experiment. The correlation coefficient changes after the sixth question: for the first 6 questions it is -0.531, while for the last 8 questions it is 0.317[4]. The above finding may be considered a 'maturity indicator' which reveals the point at which forecasters adapt to the nature of the experiment and external information flow.



	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th	14th
Aver. Forecasts	1.85	1.40	1.62	1.26	1.38	1.47	1.66	1.37	1.44	1.19	1.25	1.26	1.34	1.37
Aver. Characters	315	389	332	355	328	316	374	336	521	162	511	330	405	416

Figure 1. Evolution of character count and number of forecasts per question, for the 14 questions of the study

[4] Similar result if we analyse the first seven questions versus the last seven

4.1. The effect of a small pool of experts and expedited identification

In order to examine RQ1, we set as a cut-off point the 6th question, and we consider this the *identification* phase. This was decided as participants knew they had to answer at least 11 questions, with the middle point been at 5.5, thus we selected the completion of the sixth question as the end of the identification phase5. At this stage, we identify the forecasters in various percentiles: the top 2%, which is the standard check for the validity of the superforecasting hypothesis; then 5%, 10%, and 25% (the top quartile). Those -if any, that fall in those percentiles constitute the potential superforecasters. We then move on to the *confirmation* phase, where we check and confirm who among the potential superforecasters, remains within the same percentile for the last eight questions: all those who do, are confirmed as a superforecasters at the respective percentile. There is no guarantee that there will be any.

We decided to be very strict in our definition of what constitutes a superforecaster. In the identification phase, we had to consider forecasting performance across six questions and respective forecasts. We could have ranked experts via their average performance across the six questions and then pick the top 2% of those (5%, 10%, 25% respectively). This gives an expert participant in our experiment an (a priory) 2% chance to be a potential superforecasters. To be confirmed in the next phase, the expert should be in the top 2% in the confirmation phase too, so the actual chances prima facie of been a superforecaster are just 0.04%. For the top quartile, the respective chances are 25% for the identification phase and 6.25% after the confirmation phase.

With our stricter criteria, for an expert to be a potential superforecaster, the condition is to be ranked amongst the top 2% (5%, 10%, 25% respectively) for at least 5 out of 6 questions in the identification phase. This gives only an a priori chance of $(2\%)^5$, a mere chance of 3.2e-7 %. In fact, the chances are slightly better as an expert can answer any 5 out of the first six questions, so six times more, yet again 1.92e-6 %, so mere impossible one could argue. And to be confirmed a superforecaster an a priori chance of $(2\%)^6$, a chance of 6.4e-9 % (or slightly more at 1.8e-7 % if we consider all possible combinations of 6,7 and 8 questions out of 8); So mission impossible?

[5] We find similar results in our analysis if we use as a cut-off point the 5th or the 7th question for the 2% (the superforecasters have answered all 14 questions) and 5%, and insignificant differences in the larger % (as some participants were skipping answering some questions). This was done as part of robustness checks during the analysis.

Here exactly lies the very essence and beauty of what a true superforecaster is claimed to be. Imagine if you had 100 runners and had them run 6 races and for a runner to be considered a super-runner, he/she would have to be in the top-2in five out of six races. This gives a random chance of 1.92e-6 %, But if Usain Bolt6 - the current world record holder in 100m and 200m was among those 100 runners, then one could argue that being in the top-2 in five out of six races would be a piece of cake for him. The same is the argument Tetlock and Mellers are doing in the superforecasting hypothesis: in a similar fashion that Usain Bolt is better genetically and better-trained to win those races, a superforecaster is better naturally, and trained better to forecast better, far better than anybody else, and consistently: a natural talent in forecasting. So to that end, probabilities become irrelevant, and abilities come to the forefront! The exact steps followed were as follows:

> *Identification* phase

- We calculated the scores per forecaster for the first 6 questions
- We ranked them (from the lowest Brier score to the highest) and created percentile 'benchmark bins' (2%, 5%, 10%, and 25%)
- We identified the top-forecasters per bin and per question, and these are identified as the potential superforecasters

> *Confirmation* phase

- We calculated the scores per forecaster for the remaining 8 questions
- We detected those that were belonging in the respective bins for a minimum of 6 out of the last 8 questions (11 out of 14 for the overall experiments): these are the confirmed superforecasters (per bin).

In figure 2, for the total number of the 195 fully-engaged participants in our study (Groups A and B together), for which formal analysis has been performed, we present the number of the confirmed superforecasters (first bar) at the 2%, 5%, 10% and 25% bins respectively, versus the maximum potential superforecasters in each bin (second bar – that is simply the 2%,5%, 10%, and "5% of the sample size n=195). The metric being used for the ranking of the performance of the participants is the standardized over the mean average Brier Score (sAvg_mean); nevertheless, similar results we get for all the four metrics employed in the study[7].

[6] https://en.wikipedia.org/wiki/Usain_Bolt

[7] The four performance metrics used in our analysis are the ones used in the GJP literature as well: Average Brier Scores (Avg), standardized over the mean Average Brier Score (sAvg_mean), Net Brier Points (Net), and Standardized over the mean Net Brier points (sNet_mean)

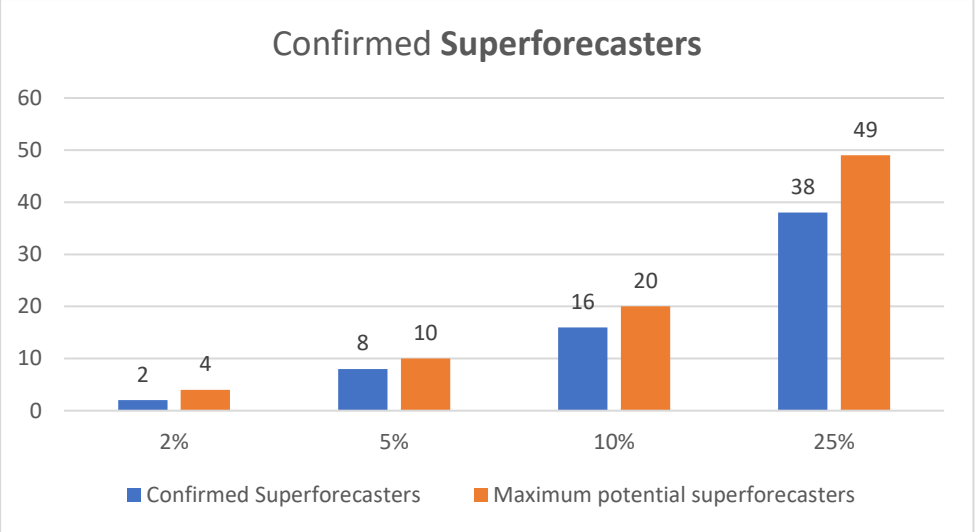


Figure 2. Confirmed **Superforecasters** per percentage bin
(ranking via the Standardized over the mean Average Brier Scores)

This first bar in figure 2, the identification[8] and confirmation of **two** (2) **superforecasters** in our study is by and large the answer to RQ1, and our empirical **evidence in support of the superforecasting hypothesis**. One could argue that in such a small sample (sample size n =195), and with a stricter than GJP selection rule, no superforecaster would be found; and given the aforementioned discussed (small) probabilities to find superforecasters, as well as anecdotal discussions of the corresponding author with members of the GJP team, the prospect was that with such a small initial pool of experts (n=195) and the expedited identification (of less than a year), there will be no evidence of the superforecasting hypothesis from our experimental setup.

The 8, 16, and 38 experts confirmed at 5%, 10%, and 25% are not considered superforecasters; these are the experts that satisfied a similar criterion as the superforecasting group (but for a different %). These experts could be used instead of the superforecasters in a real organization, in case no superforecasters were confirmed, as these are still the best forecasters in the house: and this is a very important implication for practice, a way forward for small organizations to get good forecasts via utilizing their own resources.

4.2. Characteristics and skills of superforecasters

We wanted to identify some characteristics and skills of the better-performing forecasters that would help organizations in the future to identify more easily their top-forecasters and superforecasters. Given the very small samples for the superforecasting group of 2%, as well as the small ones for 5% and 10% bins, we did perform this analysis on the 25% bin, analyzing both Group a and Group B as one large group; thus the 38 top-forecasters visualized as the penultimate bar in figure 1. The key profile characteristics that differentiate them from the remaining participants are presented below in Table 3.

[8] There were three potential superforecasters identified at 2% (out of the maximum four that is 2% of n=195), two of which have been confirmed in the second phase of the experiment as superforecasters.

These have been confirmed as being statistically significant with Pearson's Chi-Squared independence tests.

Gender	The test identified a significant difference in performance ($X2(1)=4.808$, $p<0.05$) between males (24.2% in the 25% bin) and females (11.3% in the 25% bin).
Working experience	The test identified a significant difference in performance ($X2(6)=19.04$, $p<0.05$) between the various levels of experience. Specifically, of 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) comes from those with no experience (mainly in academia).
English language proficiency	The test identified a significant difference in performance ($X2(2)=7.31$, $p<0.05$) between the three levels of English language knowledge. Specifically, the greatest contributing percentage (25% in the 25% bin) comes from those with advanced proficiency in the English language. Those with the intermediate knowledge contributed with 12.3%, whereas those with basic knowledge had no contribution at all.
Frequency of information gathering	The test identified a significant difference in performance ($X2(4)=9.98$, $p<0.05$) between the various levels of information frequency. Specifically, the greatest contributing percentage (31.8% in the 25% bin) comes from those that stay informed on a daily basis. The contribution declined in an almost linear fashion: 20.9% (weekly), 14.3% (monthly), 3.6% (more scarce), 0% (never).
Type of information sources	The participants were requested to denote their principal sources of information. The choices provided were: (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. The test identified a significant difference in performance ($X2(1)=5.32$, $p<0.05$) between those who selected choice (2) (24.2% in the 25% bin) and those who did not (11.3% in the 25% bin). The results for the 3rd source (independent websites) were also similar ($X2(1)=9.84$, $p<0.05$), with the corresponding percentages: 26.4% (yes), 8.1% (no). The remaining sources (1, 4 & 5) did not contribute significantly to forecasting accuracy.
Language of information sources	The participants were requested to denote the language of their sources of information (more than one could be selected). The choices provided where: (1) Gr, (2) En, (3) Fr, (4) De, (5) Ru, (6) Ar, (7) Other. The test only identified a significant difference in performance ($X2(1)=6.87$, $p<0.05$) between those who selected choice (2) (24.4% in the 25% bin) and those who did not (8.3% in the 25% bin). Thus sources in English are of essential importance.

Table 3. Key characteristics of top-forecasters.

We believe that the gender-related finding bears some further discussion. Similar findings were provided by Frederick (2005), where men were receiving consistently higher scores than women. Although Frederick's tests were not evaluating pure forecasting skills, rather pure cognitive abilities, it has also been verified by Mellers et al. (2015) that there exists a positive correlation between the two. The difference in performance during Frederick's experiments was attributed neither to biases nor to lack of attention. It was the difference in the mathematical reasoning skills that did actually helping men perform better. Although our experiment's questions did not necessarily require advanced mathematic literacy, the approach that was proposed to them in order to help them derive to a more accurate forecast, required some relevant skills, and probabilistic reasoning as the experts may have to resort to tactics like define base rates, aggregate probabilities, and Bayesian analysis.

4.3. The effect of additional training

RQ2 questions the benefits of an extra short-training; essentially the impact of specialized training. To provide some evidence to that end, we analyzed two groups that were imposed on different treatments. Group A was smaller than group B as from the initial pool of 314 experts, and the respective equal initial split, an uneven number of experts did not engage till the end of the experiment, resulting in 89 experts in group A and 109 in group B. The number of superforecasters identified for each group is illustrated in figure 3. The same number (1) of one superforecaster for each group is identified; while more top-forecasters are identified in group B that had more training, in the respective higher percentage bins: 8 versus 7 for 10%, and the larger difference (8 experts) of 22 to 14 for the 25% bin. Nevertheless and given the uneven samples, these differences are not statistically significant. This result holds for all four performance metrics we apply in our analysis.

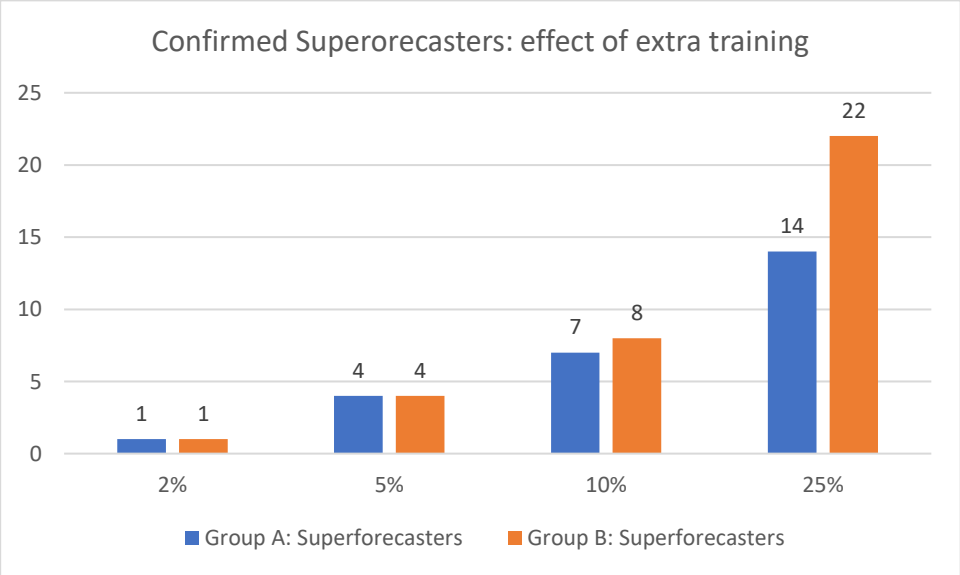


Figure 3. Number of confirmed superforecasters per bin and per group (sAvg_mean score). Group A (n_A= 86) had standard GJP training, while Group B(n_B= 109) had an extra 20-minute training structured judgmental forecasting techniques.

Following on the previous analysis, our first insight is that more training does not necessarily lead to identifying more top-forecasters. But even if these are not more, they may well be performing slightly better, at least as a whole group? We can assess that by looking more carefully at the performance scores. In figure 4 we can see that the median[9] for sAvg_mean is slightly lower for Group B than Group A (-0.437 vs -0.252), as is the case for the entire boxplot. So there is some preliminary evidence that more training may lead to slightly better performance overall, but this does not translate directly to more superforecasters or top-forecasters identified; at least for the (small) scale of our experiment, we do not find statistically significant evidence to that end.

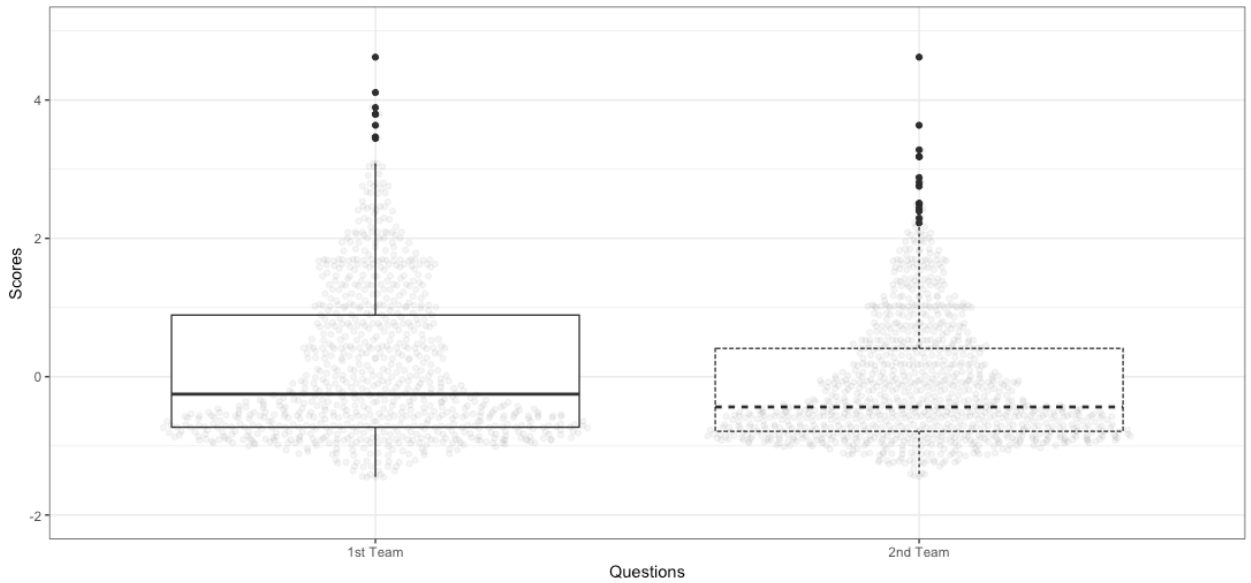


Figure 4. Boxplot sAvg_mean scores for Group A (n_A= 86) and Group B(n_B= 109) had an extra 20 minute training structured judgmental forecasting techniques.

We further employ stochastic dominance tests (SD) in order to evaluate the comparative performance of our two groups (Hadar & Russell, 1969) 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 rather to minimize them, given that the lower the Brier Scores the better the forecasts. We perform two consecutive bootstrap re-samplings of our samples as follows:

- Block Bootstraps (100 iterations per group) to create new samples with an equal number of entries.
- Bootstraps to estimate p-values for each of the blocks produced above (20 iterations per block).

[9] Group B has also a lower mean at -0.120 versus 0.143 for group A. A value in the range [-2,0) indicates that the forecaster is performing better than average.

Across all metrics employed in our study, and for all the p-values calculated following the aforementioned analysis, group B outperforms group A. Figure 5 depicts the Empirical Cumulative Distribution Functions (ECDF) which derive from the above analysis. Thus, by aiming for the lowest scores which mean smaller forecasting errors, we can conclude the performance of group B stochastically dominates that of group A.

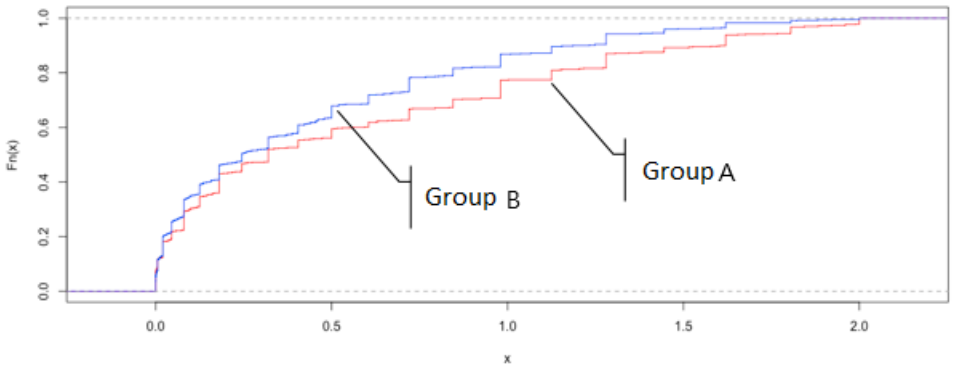


Figure 5. ECDFs for stochastic dominance tests for Group A and B.

Concluding our empirical analysis, we do believe we provided sufficient evidence for RQ1 in favor of the superforecasting hypothesis in the face of scarcity of resources: the number of experts and the time to identify them. For RQ2 the evidence was much less and was only sufficient if the analysis was focusing on the entire distribution of our experts.

5. Conclusion, limitations, implications for theory and practice, and the future

This research sought to do a reality check of the well-celebrated GJP results and the hype around superforecasting. Our argument and the main motivation is, the GJP proposition would be much stronger and relevant to the business world if it could be applied with fewer resources: we examined the influence of small sample sizes and limited time to identify the superforecasters. The evidence was supportive and this is a very promising result, which can inspire further research towards this direction. We further examined if more training can lead to even more evidence and more superforecasters: there our analysis was inconclusive, although we provided some preliminary evidence – consistent with GJP – that more training helps.

We did also try to replicate the conditions of a corporate environment by avoiding a strict experimental setup. The only limitations we imposed were as follows:

- Fully anonymized participation to avoid interactions and influence.
- Password-protected training sessions to ensure that only the designated participants had access to the relevant training modules.

The principal difficulty faced throughout our study was keeping the engagement of our participants. Given the limited visibility and lack of sponsorship for our experiment (particularly when compared to the GJP and the support from IARPA), we had to counterbalance the participants' desire to drop out with a major incentive of a nominal value of 650€. We, therefore, believe that a formal reward scheme should be established if similar outcomes are to be achieved in a real-life corporate environment: this could come in the form of a vacation voucher, or more annual leave, or an end-year bonus for example.

As far as implications for theory are concerned, our study clearly corroborates to the judgmental forecasting and superforecasting literature: for the latter, we do provide evidence for an unexplored context: the scarcity of resources, of the two most important resources, that of time and availability of experts. We do further contribute to the body of literature in training and learning via exploring the impact of additional short-training in the aforementioned context.

The aforementioned theoretical and methodological contribution leads naturally to what is the main takeaway for practitioners: the main implication for practice. An SME or even a larger organization will have difficulty accessing the superforecasters from the GJP: supply, demand, and cost would be detrimental. It would make more sense for example for a national bank to rely on their own experts in-house for the most important forecasts they need to produce. How that can be done? Within 6 months or a year, a company can run a forecasting tournament in sequential phases (identification first and confirmation next, probably equally split time-wise), with a series of weekly and monthly questions, among all employees. At the end of the tournament, the top-forecasters would be confirmed. You may not necessarily find confirmed superforecasters (at 2%) but the top-forecasters will still do a decent job for future forecasting tasks. And if you train them further, they will do better over time. And a further striking result - from GJP directly, if you team them up (rather than ex-post averaging) you will get even better forecasts. So a clear way forward for any organization.

We also came across a very interesting finding: that relating to the English language skills. The language used when retrieving information appears to have a big impact. The majority of information on the World Wide Web is provided in English, and forecasters are, in a way, forced to adapt. Consequently, English language proficiency can be considered a key asset when it comes to information collection. We believe that further research should be conducted in this direction, to identify the contribution of language skills to the various types of forecasting questions - primarily in terms of the context.

GJP results on and on do emphasize when it comes to finding global superforecasters: size and time matters. Be that as it may, this predicament tends to make the whole superforecasting proposition impractical for real-life conditions in today's business environment. As such, we do believe this study shows the way forward for a practical seize of the main GJP results, and also highlights the roadmap for future research. More and more studies with fewer experts, less time in hand, smaller levels of expertise, different and more diverse types of questions – focusing on economics and financial aspects too, more extreme events – even black swans like COVID-19 nowadays, and the relative impact in the future of mankind, healthcare, economy, and society.

Supplementary Material

- In this study we have used only freeware software tools, in order to make our proposition even more appealing to companies with limited resources: Google forms, R code, and raw data, can be made available upon request to the authoring team.

References

Antonacopoulou, E. P. (1999). Training does not imply Learning. The Individual's Perspective. *International Journal of Training and Development, 3*(1), 14–33.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin*, *110*(3), 486–498.

Armstrong, J. S. (2001). *Principles of Forecasting.* Norwell, MA: Kluwer.

Arvan, M., Fahimnia, B., Reisi, M., & Siemsen, E. (2019). Integrating human judgement into quantitative forecasting methods: A review. *Omega*, 2019, 86, 237-252

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, *78*(1), 1–3.

Camp, D. (2012). Here, There, and Anywhere: Transfer of Learning. *Critical Questions in Education*, *3*(1), 35-42.

Carbone, R., & Gorr, W. L. (1985). Accuracy of Judgmental Forecasting of Time Series. *Decision Sciences*, *16*(2), 153–160.

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.

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.

De Baets, S., & Harvey, N. (2020). Using judgment to select and adjust forecasts from statistical models. *European Journal of Operational Research* 284(3), 882-895.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving Intelligence Analysis with Decision Science. *Perspectives on Psychological Science, 10*(6)*,* 753–757.

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.

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

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

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.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42.

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. https://www.semanticscholar.org/paper/Why-Quantitative-Probability-Assessments-Are-in-Friedman-Baker/3d11ba65cc4874358c2e2ff9e66643040d23a574

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, 62(2),* 410–422,

Gardner, D. (2011). *Future Babble: Why Expert Predictions Are Next to Worthless, and You Can Do Better*. New York: Penguin Group (USA) Inc.

Gentner, D., & Smith, L. (2012). *Analogical Reasoning*. *Encyclopedia of Human Behavior: Second Edition*

(2nd ed., Vol. 1). Elsevier Inc.

Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind : advances in the study of language and thought*. MIT Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gentner, D. (2002). Analogy in Scientific Discovery: The Case of Johannes Kepler. In *Model-Based Reasoning* (pp. 21–39). Boston, MA: Springer US.

Green, K. C., & Armstrong, J. S. (2007). Structured analogies for forecasting. *International Journal of Forecasting*, *23*(3), 365–376.

Goodwin, P., Gönül, M. S., & Önkal, D. (2019). When providing optimistic and pessimistic scenarios can be detrimental to judgmental demand forecasts and production decisions. *European Journal of Operational Research* 273(3), 992-1004.

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.

Hanoch, G., & Levy, H. (1969). The Efficiency Analysis of Choices Involving Risk. *Review of Economic Studies*, *36*(3), 107–335.

Hernandez, D. (2017). How Our Company Learned to Make Better Predictions About Everything. *Harvard Business Review*, (May 15, 2017).

Horowitz, M., Brandon, M., Stewart, B. M., Tingley, D., Bishop, M., Resnick Samotin, L., Roberts, M., Chang, W., Mellers, B., & Tetlock, P. (2019). What Makes Foreign Policy Teams Tick: Explaining Variation in Group Performance at Geopolitical Forecasting, *The Journal of Politics,* 81(4), 1388-1404.

IARPA. (n.d.). The Good Judgment Project. Retrieved January 12, 2020, from https://www.iarpa.gov/index.php/newsroom/iarpa-in-the-news/2015/439-the-good-judgment-project

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

Khong, Y. F. (1992). *Analogies at war : Korea, Munich, Dien Bien Phu, and the Vietnam decisions of 1965*. New Jersey: Princeton University Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

Litsiou, K., Polychronakis, Y., Karami, A., & Nikolopoulos, K. (2020). Relative performance of judgmental methods for forecasting the success of megaprojects, International Journal of Forecasting, in press, https://doi.org/10.1016/j.ijforecast.2019.05.018

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518.

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.

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.

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.

Makridakis, S. G., Hogarth, R. M., & Gaba, A. (2010). Why forecasts fail. What to do instead. *MIT Sloan*

Management Review, *51*(2), 83–90.

Makridakis, S. G., Wheelwright, S. C., & Hyndman, R. J. (1998). *Forecasting: Methods and Applications*. New York: Wiley.

Malkiel, G. (1973). *A Random Walk Down Wall Stree*t. New York: W. W. Norton & Company, Inc.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou,

Y.,, Horowitz, M., Ungar, L. & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a

Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science, 10*(3), 267–281.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. & Horowitz, M. (2015). The psychology of intelligence analysis: drivers of prediction accuracy in world politics. Journal of Experimental Psychology. Applied, 21(1), 1–14.

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.

Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and Motivational Biases in Decision and Risk Analysis. *Risk Analysis*, *35*(7), 1230–1251.

Nikolopoulos, K. (2020). We need to talk about intermittent demand forecasting, European Journal of Operational Research, in press, https://doi.org/10.1016/j.ejor.2019.12.046

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.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgemental forecasting in times of change. *International Journal of Forecasting*, *9*(2), 163–172.

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.

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

Petropoulos, F., Makridakis, S., Assimakopoulos, V. & Nikolopoulos, K. (2014). 'Horses for Courses' in demand forecasting. European Journal of Operational Research 237 (1), 152-163

Plutarch. (1936). *Plutarch Moralia V*. (G. P. Goold, Ed.). London: Loeb Classical Library. https://doi.org/10.4159/DLCL.plutarch-moralia_e_delphi.1936

Reisberg, D., Gentner, D., & Smith, L. (2013). *Analogical Learning and Reasoning*. *The Oxford Handbook of Cognitive Psychology*. Oxford University Press.

Rekik, Y., Glock, C. H., & Syntetos, A. A. (2017). Enriching demand forecasts with managerial information to improve inventory replenishment decisions: Exploiting judgment and fostering learning. *European Journal of Operational Research* 261(1), 182-194.

Sanders, N. (1992). Accuracy of judgmental forecasts: A comparison. *Omega*, *20*(3), 353–364.

Savio, N. D., & Nikolopoulos, K. (2013). A strategic forecasting framework for governmental decision-making and planning. *International Journal of Forecasting*, *29*(2), 311–321.

Savio, N., & Nikolopoulos, K. (2010). Forecasting the effectiveness of policy implementation strategies. *International Journal of Public Administration*, *33*(2), 88–97.

Schoemaker, P. J. H., & Tetlock, P. E. (2016). Superforecasting: How to upgrade your company???s judgment. *Harvard Business Review*, *2016*(May).

Syntetos, A. A., Kholidasari, I., & Naim, M. M. (2016). The effects of integrating management judgement into OUT levels: In or out of context? *European Journal of Operational Research* 249(3), 853-863.

Syntetos, A., Babai, Z., Boylan, J., Kolassa, S. & Nikolopoulos, K. (2016) Supply chain forecasting:theory, practice, their gap and the future. European Journal of Operational Research, 252 (1), 1-26

Teodoridis, F., Bikard, M, & Vakili, K. (2018). When Generalists Are Better Than Specialists, and Vice Versa, Harvard Business Review

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.

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 E., & Gardner, D. (2015). *Superforecasting : the art and science of prediction*. Crown Publishing Group

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.

Ungar, L., Mellors, B., Satopaä̈, 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.

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