Relative performance of methods for forecasting Special Events

Konstantinos Nikolopoulos[a,*], Akrivi Litsa[b], Fotios Petropoulos[c],

Vasileios Bougioukos[a] and Marwan Khammash[d]

June 2014

Abstract

Forecasting Special Events such as conflicts and epidemics is challenging because of their very nature and the limited amount of historical information from which a reference base can be built. This study evaluates the performances of Structured Analogies, the Delphi method and Interaction Groups in forecasting the impact of such events. The empirical evidence reveals that the use of Structured Analogy leads to an average accuracy improvement of 8.4% in forecasting Special Events compared to Unaided Judgment. This improvement in accuracy is greater when the use of Structured Analogies is accompanied by an increase in the level of expertise, the use of more analogies, the relevance of these analogies, and the introduction of

[a]Bangor Business School, Bangor University, Bangor, UK

[b]Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece

[c]Lancaster Centre for Forecasting, Lancaster University, Lancaster, UK

[d]School of Business, Management and Economics, University of Sussex, UK

*Corresponding author e-mail address: k.nikolopoulos@bangor.ac.uk;

Co-authors' e-mail addresses: alitsa@fsu.gr; f.petropoulos@lancaster.ac.uk; bbougiou@gmail.com; m.khammash@sussex.ac.uk

pooling analogies through interaction with experts. Furthermore, the results from group Judgmental Forecasting approaches showed promising results; the Delphi method and Interaction Groups improved accuracy by 27% and 54.4%, respectively.

Key words: Judgmental Forecasting, Structured Analogies, Delphi, Interaction Groups, Governmental Forecasting

Acknowledgements

The authors would like to thank Dr. Kesten Green (University of South Australia), Professor Scott Armstrong (Wharton), Professor Spyros Makridakis (INSEAD), Professor Vasileios Assimakopoulos (NTUA), Professor Aris A. Syntetos (Cardiff), Professor Paul Goodwin (Bath) and three anonymous referees for their very useful and constructive comments that led to significant improvements of this article. The authors would also like to thank for their feedback the attendants of sessions where earlier versions of this work have been presented in the: University of Sussex in October 2012, OR54 in September 2012 (Edinburgh, UK), and International Symposium of Forecasting in June 2012 (Boston, USA), June 2010 (San Diego, USA), June 2009 (Hong Kong) and June 2008 (Nice, France). The authors would also like to state that: (a) at least one of the authors has read each of the original studies cited, and (b) most of the authors cited in this work have been contacted directly so as to ensure that their work has been properly summarized and that no other relevant research has been overlooked.

1. Introduction

Forecasting the timing and impact of Special Events such as natural catastrophes, conflicts, and even economic meltdowns like the 2008–2009 steep recession can be challenging because of the limited amount of historical information from which a reference base can be built. In this study, the performance of different methods of forecasting Special Events is evaluated by presenting empirical results for two real examples of policy implementation that pose typical Special Events because similar policies have rarely been implemented in the past. To forecast the impact of new policies, governments use Impact Assessments (IAs) and Cost-Benefit Analyses (CBAs). Both techniques are lengthy and costly processes that are typically outsourced and rarely contain any type of quantitative forecast of the impact of the introduced policy. Savio and Nikolopoulos (2013) propose a solution to this problem and suggest that forecasts should be prepared with simple Judgmental Forecasting methods before the employment of IAs or CBAs. Thus, although forecasting methods are not an alternative to IAs and CBAs, they might be used as a simple screening tool to indicate which policy implementations should be tested further with the complex and more expensive IA and/or CBA methods.

Although the empirical evidence in this study was derived from a governmental decision-making context, the results may be generalized and applied to a variety of business situations in which the proposed forecasting methods might be used to successfully forecast projects, investments or even more regular events, such as marketing communications. In essence, the literature that favors the use of simple methods to forecast with information (Nikolopoulos, Goodwin, Patelis, & Assimakopoulos, 2007) was corroborated.

The rest of the paper is structured as follows. Section 2 surveys the relevant literature on policy implementation and forecasting. Section 3 explains the methodological approach

employed in selecting the cases, methods, and evaluation metrics, as well as in choosing the experts and deciding their level of expertise. Section 4 presents the results, and section 5 discusses the findings. Finally, the last section offers concluding remarks and roadmaps for future research.

2. Background Literature

The application of the simplicity principle to theories is sometimes defended as an application of Occam's Razor, that is, “accept the simplest theory that works” (Simon, 1979). Zellner (2007), a leading economist, believed that complicated problems could be solved by the application of a few powerful, simplifying concepts, which he called "sophisticated simplicity". These powerful and simplifying concepts have been implemented in a myriad of industries and services. Simplicity also plays an integral role in shaping decision-making heuristics. Gigerenzer (1996) argues that biases that stem from heuristics can be eliminated by utilizing particular methods in a suitable context.

2.1 Policy Implementation

In governmental decision-making, finding simple tools that generate the same quality results as more complex tools is sometimes difficult. Additionally, the fact that public expenditures are involved makes decision makers less inclined to use methods that might seem simplistic in the eyes of the watchdog.

Impact Assessment (IA) is an aid to political decision making that aims to identify and assess the effectiveness of policies and objectives pursued (European Commission, 2009). IA consists of a set of logical steps leading toward the formulation and preparation of proposals through a

balanced appraisal of political impact. Moreover, policies concerned with new technologies and innovations are often assessed with their adoption and diffusion rates, which are typically measured in terms of the proportion of agents using the new technique compared to those using older techniques (Askarany, 2006).

The main goal of modern paradigms of public administration, such as Public Value Management, is to enhance public values through forces that do not rely solely on traditional reformative norms (Stoker, 2006). Thus, Public Value Management emphasizes the feasibility and value creation of individual actions. The core idea of adding value to the public domain by ensuring that policy objectives are met while improving the efficiency of the public policy process is consistent with the fundamental notion of this research (Pitts, 2007; Talbot, 2009). Public Value Management would effectively require any government to base its decisions on a priori forecasts of policy effectiveness, which is defined as the extent of change in the current situation in the direction of the policy target. Ex-ante evaluations of policy effectiveness typically involve a mixture of Impact Assessment and Cost Benefit Analysis.

IA may be performed by using a variety of different models (European Commission, 2009). The selection of a particular model is dependent on the availability of data in each particular case (De Gooijer & Hyndman, 2006; Savio & Nikolopoulos, 2009); IA is considered a rather costly and resource-extensive tool (Savio & Nikolopoulos, 2010, 2013).

Although CBA is a useful tool, it is limited because it only evaluates policies in terms of economic efficiency (Maas, 1966; Simpson & Walker, 1987). Both IA and CBA are tools that can be used after a specific policy implementation has been decided upon (Savio & Nikolopoulos, 2013). As a result, they are not used in the preliminary screening of alternative policy implementations, which leads to the space for simple and fast forecasting approaches that

estimate the effectiveness of policies that may be implemented. Consequently, those forecasts might be used to select which alternative to implement, and then IA or CBA would be employed.

2.2 Forecasting

The standard benchmark of the Judgmental Forecasting approach is Unaided Judgment (Green & Armstrong, 2007a) in which individuals are not given guidance as to proper forecasting procedures. The unstructured employment of panels of experts (Savio & Nikolopoulos, 2010) has several limitations (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007), such as the inability of forecasters to recall analogous cases and the recollection of unusual or inappropriate past cases. Thus, the adoption of structured approaches is seen as a better way to overcome these limitations and fully capitalize on expert judgment (Green & Armstrong, 2007b).

The Delphi method (Rowe & Wright, 2001) is a multiple-round survey in which experts participate anonymously and provide their forecasts and feedback. At the end of each round, participants receive a report, including descriptive statistics of the forecasts provided. The Delphi method is completed after a predefined number of rounds or whenever a desired consensus level is reached. Generally, four key features tend to define a 'Delphi' a group procedure – anonymity, iteration, controlled feedback, and the statistical aggregation and presentation of group responses. Conversely, the Interaction Groups method suggests active interaction with a group of experts until a consensus forecast is reached through debate and discussion. A key driver in this method's success is the pooling of information. However, potential problems arise from group biases introduced by the face–to-face contact of the experts, such as the 'central tendency' and the 'dominant personalities' effects (Van de Ven & Delbecq, 1971). Evidence of the forecasting potential of Interaction Groups is not consistent (Armstrong, 2006; Boje & Murnighan, 1982;

Graefe & Armstrong, 2011). Moreover, group-based approaches incur extra costs resulting from multiple rounds in the Delphi setup or the need for meetings in the formulation of Interaction Groups. This fact renders these methods relatively more costly than other methods that group-based approaches are competing against.

3. Methodology

The Special Events examined in this study are two policy implementations (PIS – a term introduced in Savio & Nikolopoulos, 2013) provided by an EU country's Special Secretariat for Digital Planning, a governmental body that focuses on controlling budgets that aims to accelerate the use of IT.

The first policy (PIS A) was entitled **"See Your Life Digitally"** and aimed to promote the laptop purchases among undergraduate students in universities. The government was willing to provide a subsidy of up to 400€ for the purchase of a laptop computer. With the €400 incentive and the overall policy budget that was to be allocated, decision makers were interested to forecast the following:

[PIS A - Q1]: *What percentage of eligible students will buy a laptop?*

[PIS A - Q2]: *How many weeks will it take for 50% of eligible students to participate in the scheme?*

The second policy (PIS B) was entitled "**Parents.eu**" and aimed to train and certify parents of high school pupils in 'Internet safety'. Parents had free online access to a distance-learning platform and free home tuition from instructors. Moreover, the policy budget funded a two-month subscription to a broadband service chosen by the parents and covered the expenses and

fees for their certification exam. The policy makers in this instance were interested in obtaining forecasts for the following:

[PIS B - Q1]: *What percentage of eligible parents will receive training?*

[PIS B - Q2]: *What percentage of eligible parents will receive certification?*

[PIS B - Q3]: *What percentage of eligible parents will obtain broadband Internet access (using the funding provided by the policy scheme)?*

Table 1 presents the actual results from the implementation of the two policies, and these outcomes will be used to evaluate the accuracy of the forecasts.

Table 1 here.

It must be emphasized that the Special Secretariat did not produce advance forecasts for the effectiveness of the proposed policies. In fact, the incentives (400€ in the first policy and the funded services in the second) were ad-hoc decisions. The lack of forecasts might sound irrational prima-facie, but it is common practice in such projects for the following reasons:

- Lack of in-house resources to produce forecasts, so forecasting would have to have been outsourced to a consultancy firm, which would have meant costs and delays for the secretariat. Furthermore, the cost of obtaining external forecasts for such a small-scale EU project would be inappropriate.
- In such projects, delays create problems because of the way EU funds are allocated. Although the funds are secure in principle and provided from the EU to member states to run investment and development projects, delays in completing projects can lead to the cessation of funding.

3.1 Experts

As no similar policy schemes had been implemented in the past, no quantitative data were available. Thus, experts were used to provide judgmental forecasts. The participants were chosen based on their expertise in one or more of the following areas:

- Digital Planning policies;
- Forecasting; and
- Information Technology.

A list of 300 experts was formed with the assistance of the Special Secretariat. All experts were initially contacted via email, in which a brief description of the forecasting exercise and a formal invitation to take part in the study were provided. The invitation came directly from the Special Secretariat of Digital Planning of the Ministry of Economics and Finance to assure potential participants of the importance of this operation. In total, 55 experts responded positively to the call and participated in the research. These experts were sourced from a wide variety of sectors, including academia, industry, financial services and consultancy firms.

No monetary or in-kind incentives were provided to any participant for taking part in the experiment except for the personal invitation from the Secretariat. All participants were provided with full descriptions of the two policies and the entire procedure was administered remotely through email.

3.2 Methods

Four methods have been evaluated in this study; the first – Unaided Judgment – is the benchmark. Experts were instructed to provide point forecasts and 90% prediction intervals. A

full calendar week was given to return their forecasts via email. The methods that were deployed included the following:

Group A - (20 experts), Unaided Judgment (UJ): This method is a simple and quite popular Judgmental Forecasting approach. Experts are given no guidance except for a general description of the intended policies.

Group B - (20 experts), semi-Structured Analogies (s-SA): The Structured Analogies approach was proposed by Green and Armstrong (2007b) and is based on forecasting by analogy by exploiting the similarities of past events or experiences. These past events/situations have the same or similar characteristics as the problem to be forecasted and can be used as templates. These types of mental templates are the analogies. The experts are first asked to recall as many analogies as possible. Subsequently, they produce a quantitative similarity rating between each analogy and the problem to be forecasted and state the outcome of that analogy. The administrator uses the experts' data to produce a final forecast. In this study, a slightly simpler version of the method, called semi-Structured Analogies (s-SA, Savio & Nikolopoulos, 2013) was implemented. In this approach, similarity ratings and outcomes are not used by the administrator to generate forecasts because the final forecasts are produced by the experts.

Group C - (10 experts), Delphi (D): This approach is a popular group Judgmental Forecasting method that includes multiple rounds of questionnaires administered to a group of experts. Although several variations of the method exist (Rowe & Wright, 1999, 2001), only two rounds were run in the current implementation to limit the process to two weeks (and to avoid having experts drop out). In the first round, the experts forecasted with Unaided Judgment. Once the forecasts were collected, feedback was provided to the group in the form of an average forecast for the group, in addition to the maximum and minimum forecasts and the justifications

for those extreme forecasts (in a short memo). In the second round, the participants could revise their forecasts in light of the initial feedback. The average of the second round of forecasts was used as the group forecast.

Group D - (5 experts), Interaction Group (IG): This group met in a restaurant, and the entire process was supervised by an experienced facilitator. The meeting lasted three hours and was recorded. The first hour was spent with introductions and a light dinner. In the next two hours, the group forecasting exercise occurred, in which the experts were first given the questionnaires, then encouraged to recall analogies and their corresponding outcomes, and then to rate those analogies in terms of similarity. Finally, the experts were asked to select the most appropriate analogies to produce point forecasts as well as 90% prediction intervals. This process was first performed individually and was then followed by the group interaction in which experts repeated the process aloud and exchanged their information until a consensus group forecast was reached.

3.3 Participants' Expertise

The participants' expertise was rated based on the following three specific questions:

[EXP - Q1]: *How would you rate your expertise (out of 10) in policy forecasting?*

[EXP - Q2]: *How many years of experience do you have in the ICT market?*

[EXP - Q3]: *How many years of experience do you have in policy making for the ICT market?*

Based on the normalized responses to the above three questions, an index was created. Subsequently, according to the scores from that index, experts were assigned to two groups (low

expertise and high expertise). Table 2 demonstrates how the experts were allocated to each group.

Table 2 here.

3.4 Measuring Performance

Forecasting accuracy was measured through an absolute metric (Mean Absolute Error, or MAE) and a relative metric (Relative Absolute Error, or RAE) as the ratio of the MAE of interest to the MAE of UJ-low-expertise. Whenever RAE<1, the treatment is an improvement over the benchmark.

Finally, for the overall comparison between all methods and all questions, Mean Absolute Percentage Errors (MAPE) was also calculated.

4. Results

For the five questions presented in 3.1 (with the realized outcomes listed in Table 1), all errors for the experts' forecasts were calculated. For each of the methods and questions, the MAE, the RAE, the percentage frequency with which the forecasts fall within the experts' prediction intervals, and the standard deviation of the errors were calculated. Moreover, all results were broken down according to the level of participants' expertise. Four forecasts are given in the form of *percentages* (%) because these are essentially 'take-up rates', whereas one is given in *weeks* (PIS A - Q2). The results for each method will be presented separately; at the end of the section, a cross-comparison will be conducted and the results will be discussed in the subsequent section.

4.1 Unaided Judgment

The results for UJ (Group A) are presented in Table 3. First, experts with high level of expertise produced the best forecasts in almost all cases. Second, this group seemed to be rather uncertain about the submitted forecasts as evidenced by the low accuracy for the prediction intervals (38%, on average).

Table 3 here.

4.2 Semi-Structured Analogies

The results for s-SA are presented in Table 4. Participants with low expertise produced worse forecasts in three out of the five questions. Low-expertise participants were more accurate in four out of the five questions, as far as prediction intervals are concerned. Notably, the standard deviation is lower for all responses provided by high-expertise participants.

Table 4 here.

Many experts recalled one to two analogies per policy, whereas others provided no analogies at all. Table 5 presents the average number of analogies recalled and the respective mean similarity rating. The latter rating measures the similarity between the recalled analogies and the case under consideration; the score is judgmentally assigned by participants for each of the recalled analogies. Overall, one analogy was produced on average and the mean similarity rating was medium to high (7/10). Notably, the number of analogies recalled and the similarity ratings increased for PIS B as the level of expertise increases. Figures 1 and 2 present the accuracy (measured by MAPE) with respect to the number of analogies recalled and the similarity ratings. When experts provided two or three analogies, the accuracy seemed to increase. Furthermore, experts who provided analogies with high similarity ratings had better accuracy on average than

those who provided no analogies at all or who produced analogies with average similarity ratings.

Table 5 here.

Figure 1 here.

Figure 2 here.

4.3 The Delphi Method(D)

Table 6 presents the performance of each round of the Delphi method (Group C). The first notable result is that both forecasting errors and standard deviation are lower in the second round. As a result, the Delphi method achieved a higher level of consensus across the group through its iterative nature that was coupled with improved accuracy. The first round of Delphi forecasts were produced with UJ; comparing the first round of Delphi (Table 6) and UJ (Table 3) shows that Group C produces better forecasts for four questions and higher accuracy rates of prediction intervals (54%, on average).

Table 6 here.

4.4 Interaction Group (IG)

Table 7 presents the results for the IG. The group forecast is compared to the average of the individual forecasts of the same experts. The face-to-face interaction of IG appears to have led the experts to produce more accurate forecasts for four out of the five questions. The accuracy improvement for certain questions was remarkable (e.g., PIS B – Q1). The prediction intervals were also more accurate. The experts used s-SA to produce their individual forecasts; when they used the same method to produce a group forecast, the number and similarity of the recalled

analogies increased, as shown in Table 8 because a larger pool of analogies was used when the experts were sharing information. In addition, it must be noted that all five participants have high levels of expertise in the field, as shown in Table 2, which may be responsible for some of the success of this group.

Table 7 here.

Table 8 here.

4.5 Methods Comparison

The group forecasting techniques provided the most accurate forecasts. Quantifying the achieved improvement, a relative improvement of 8.4% is observed when using Structured Analogies, 27% when using Delphi and 54.4% for the Interaction Group.

Table 9 here.

IG is the most accurate method in four out of the five questions, whereas the D performs best in the remaining one question. Across all five questions in total (the last two rows in Table 9), IG is the best method, followed by the D and then s-SA.

5. Discussion

The empirical findings of this study illustrate that group-forecasting techniques provide the most accurate forecasts. To elaborate more on this, groups A and B (in which experts worked by themselves) were pooled together and compared to groups C and D (in which experts worked within *groups*). The results are shown in Table 10, which illustrates an improvement across the five questions (both in terms of the average and the geometric mean) when experts work in

groups (Mean RAE 0.60 vs. 0.86) and even more improvement if they are actually interacting with each other (Mean RAE 0.36).

Table 10 here.

The next question was to investigate if *structuring* the way experts elicited their forecasts provided better results. Thus, groups A and C (the first round only where experts used UJ) were pooled together and their results were compared to groups B and D. The results are shown in Table 11, which reveals an improvement across the five questions (both in terms of the average and the geometric mean) if a structuring approach is used (Mean RAE 0.75 vs. 0.81).

Table 11 here.

As far as the level of expertise is concerned, the empirical results indicate that an increased level of expertise had a positive impact on accuracy. Moreover, the number of analogies provided and the greater similarity rates seem to have a smaller impact on individuals' accuracy when the s-SA approach is used.

One of the study's contributions comes from the performance of the IG. Pooling analogies through open-dialogue in which experts exchange opinions and experiences provided very good forecasts; conversely, in the Delphi method, only descriptive statistics and limited reasoning are disseminated at the end of each round, as no exchange of analogies has taken place. The IG included experts with diverse professional backgrounds and this may have contributed in the success of this group because experts from a wide range of fields have been able to better identify good analogies by drawing on a wider pool of experience and knowledge.

Social facilitation theory claims that people tend to achieve better results in a task they master when they work with other people (Zanjonc, 1965). Zanjonc argues that even the mere presence of others can make people work harder and achieve better outcomes. Cottrell (1972) posits that

agents tend to increase their efforts because of competition or in the presence of evaluation procedures. Williams, Harkins, and Letane (1981) argue that anonymity has an important influence in an individual's output in a group; when anonymity is eliminated, social loafing is reduced. Harkins (1987) addresses that identifying and evaluating an agent's output can drive motivation. The Interaction Groups enhanced social facilitation, reduced free riding (social loafing) and enabled the group to achieve good results. The escalation rate of accuracy is greater when high-expertise groups tend to work under stimulating conditions compared to low-expertise groups. This reason was one of the drivers to use only high-expertise experts in this judgmentally sampled group. However, the process may take longer because of the 'overconfidence' and 'glossing over' effects that affect the cognitive energy (Alexander, 2003; Chi, 1978, 2006) of experts with higher expertise who must mobilize their systems to solve problems that might require extra effort, according to Kahneman (2011).

It is notable that this study's results are based on small samples (ten experts for the D and five experts for the IG), with only one group forecast for each method per question; as only five questions in the study existed, hence, only five group forecasts and five errors for each method were measured. It is difficult to gather a sufficient number of experts in reality experiments, thus results that are presented without any statistical significance tests should not be discounted prima-facie. In addition, recent discussions in the forecasting literature advocate presenting results based on small samples even in the absence of any rigorous statistical testing (Armstrong, 2007a,b; Goodwin, 2007).

To be more confident that empirical findings from the study can be generalized, sensitivity analysis of the results was conducted and presented in Tables 10 and 11 by employing a standard cross-validation approach. 10% of the samples were randomly excluded and the analysis was re-

ran in Tables 10 and 11; if the ‘same’ results were found (‘same’ in the sense that the same method continues to perform better), then it can be claimed that the results have some level of robustness. In fact, this cross-validation experiment was repeated 10 times; in all cases, the same methods continued to prevail, and the confidence in the findings became stronger.

Finally, it is notable that the proposed simple Judgmental Forecasting methods cost very little; certainly, the cost is much less than the actual policy cost. It also costs much less than any alternative forecasting approach, in which those policies could only be parts of advanced forecasting methods within costly and time-consuming CBA or IA reports; and of course much less than the actual cost of CBA or IA. However, it must be stressed that the proposed methods are not alternatives to CBA and/or IA but occur at an earlier step in the lengthy policy formulation process. Nonetheless, this step is crucial because money can be saved by investing in CBA/IA only for the most promising policy implementations. To find the most promising policies, simple, reliable, and relatively inexpensive forecasts were required. It is better to make informed decisions based on forecasts instead of making ad-hoc decisions without forecasts, which was the actual case for the specific policies under investigation in this study (see section 3.2).

6. Conclusions and Further Research

Forecasting Special Events is challenging. This study utilizes policy implementation, one of the most challenging Special Events to forecast, in which available historical information is limited and the forecasting horizon is long. The results presented here might be generalized and applied to any business or management situations. The results also corroborate the stream of forecasting research in the presence of information cues. Decision makers are expected to benefit by

adopting these simple Judgmental Forecasting methods. Nevertheless, further experiments should be conducted to improve estimates of effect size and knowledge of how conditions and variations in the methods affect their relative accuracy in different contexts.

The relative performances of four forecasting methods (Unaided Judgment, Structured Analogies, the Delphi method, and Interaction Groups) were evaluated for two different levels of expertise. Simple group techniques such as the Delphi method perform very well. Structured Analogies and Interaction Groups outperform the benchmark (Unaided Judgment) when:

- the level of *expertise* rises;
- *more analogies* are used;
- more '*relevant'* analogies are used;
- *pooling* of analogies is facilitated (through interactions of experts); and
- experts from a more *diverse background* are employed.

Indeed the findings suggest that overall actual forecasting improvement might be as high as 54% when most of these conditions are met. These results are consistent with the previous body of literature; however, the exact effect size varies depending on the context of each study.

With the aforementioned results, it can be claimed that this study corroborates the existing body of evidence that supports the forecasting principles as maintained by J.S. Armstrong (2001a) at www.forprin.com and his respective book. In further detail, empirical evidence is provided in favor of the following source: www.forprin.com, "Armstrong_2001_Checklist – form.doc" or "Standardshort.pdf"), Armstrong, J. S. (2001b).

Principle 3.5: *Obtain information from similar (analogous) series or cases.*
Principle 6.3: *Use structured forecasting methods rather than unstructured.*
Principle 7.1: *Keep methods simple.*

Principle 8.3: *Ask experts to justify their forecasts.*

Principle 8.5: *Obtain forecasts from heterogeneous experts.*

Principle 12.2: *Use many approaches (or forecasters), preferably at least five.*

Principle 13.25: *Use multiple measures of accuracy.*

Principle 13.26: *Use out-of-sample (ex ante) error measures.*

Principle 14.1: *Estimate prediction intervals (PI).*

The results presented herein are based on small-sized samples of experts, a fact that might be an impediment for generalizing the findings. However, a sensitivity analysis and a fair amount of argumentation were provided which gave more confidence that the findings can be generalized. Furthermore, if the context of this case study was take into account, and how public administration is actually organized in real life conditions, these results might provide valid insights into the performance of each method. Repetition in other case studies might help to prove the validity of the findings and provide a generalized output for the superiority of some these methods, especially the simpler ones, such as Structured Analogies.

Overall, satisfactory empirical evidence was presented to support the view that it is more beneficial for policy-makers to discontinue (or delay for a later stage) the use of more complex forecasting methods such as IA or CBA by taking into account the trade-offs of using forecasting techniques that are simple compared to those that are complex. By keeping things simple, policy makers are able to reduce the time and costs necessary for choosing policy strategies without major sacrifices in accuracy.

As far as the future of such studies is concerned, the proposed approaches could also be tested in different contexts to gather further evidence that would allow for the full generalization of the results. The application in different contexts, such as forecasting the sales of a new product or the

success of promotions in retail campaigns, would be an interesting facet of the relative performance of the methods to back up the findings and assign them extra validity and applicability, especially for the readership of this specific journal.

Moreover, an evaluation of other judgmental approaches, such as the Nominal Group Technique (Van de Ven & Delbecq, 1971), might be explored (Graefe & Armstrong, 2011). In addition, sampling more experts would offer the opportunity to test more treatments, such as IGs with UJ versus IGs with s-SA or to test SA as it was originally designed by Green and Armstrong (2007b).

Finally, the option to offer incentives to the experts has not yet been tested, and this feature has provided strong insights into similar studies in the past. Certainly more avenues could be pursued in this research domain, and it is hoped that this study will provide interest for future investigations.

References

Alexander, P. A. (2003). Can we get there from here? *Educational Research, 32,* 3–4.

Armstrong, J. S. (2001a). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–440). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). *Standards and practices for forecasting*. Available from http://www.forecastingprinciples.com/

Armstrong, J. S. (2006). How to make better forecasts and decisions: Avoid face-to-face meetings. *Foresight-The International Journal of Applied Forecasting*, *5,* 3–8.

Armstrong, J. S. (2007a). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*, 321–327.

Armstrong, J. S. (2007b). Statistical significance tests are unnecessary even when properly done and properly interpreted: Reply to commentaries. *International Journal of Forecasting*, *23*, 335–336.

Askarany, D. (2006). Characteristics of adopters and organizational changes. *Thunderbird International Business Review*, *48*, 705–725.

Boje, D. M., & Murnighan, J. K. (1982). Group confidence pressures in iterative decisions. *Management Science, 28,* 1187–1196.

Chi, M. T. H. (1978). Knowledge structure and memory development. In L.B. Resnick (Ed.), *Children's thinking: What develops?* (pp. 73–98). Hillsdale, NJ: Erlbaum.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K.A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Cambridge handbook of expertise and expert performance*. (pp. 121–130). New York, NY: Cambridge University Press.

Cottrell, N. (1972). Social facilitation. In C. McClintock (Ed.), *Experimental social psychology*, (pp. 185–236). New York, NY: Holt, Rinehart & Winston

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*, 443–473.

European Commission. (2009). *Impact assessment guidelines*. Retrieved from http://ec.europa.eu/governance/impact/commission_guidelines/docs/iag_2009_en.pdf

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592–596.

Goodwin, P. (2007). Should we be using significance tests in forecasting research? *International Journal of Forecasting*, *23*, 333–334.

Graefe, A., & Armstrong J. S. (2011). Comparing face-to-face meetings, nominal groups, Delphi and prediction markets on an estimation task. *International Journal of Forecasting, 27*, 183–195.

Green, K. C., & Armstrong, J. S. (2007a). Value of expertise for forecasting decisions in conflicts. *Interfaces*, *37*, 287–299.

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

Harkins, G. S., (1987) Social loafing and social facilitation. *Journal of Experimental Social Psychology*, *23*, 1–18.

Kahneman, D. (2011). *Thinking fast and slow*. London, UK: Penguin Books.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, *23*, 377–390.

Maas, A. (1966). Benefit-cost analysis: Its relevance to public investment decisions. *The Quarterly Journal of Economics, 80,* 208–226.

Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: a comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research, 180,* 354–368.

Pitts, D. W. (2007). Implementation of diversity management programs in public organizations: Lessons from policy implementation research. *International Journal of Public Administration*, *30,* 1573–1590.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting, 15,* 353–375.

Rowe, G., & Wright, G. (2001). Expert opinions in forecasting. Role of the Delphi technique. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 125–144). Norwell, MA: Kluwer Academic Publishers.

Savio, N. D., & Nikolopoulos, K. (2009). Forecasting the economic impact of new policies. *Foresight, 11*, 7–18.

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

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

Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review*, *69,* 493–513.

Simpson, D., & Walker, J. (1987). Extending cost-benefit analysis for energy investment choices. *Energy Policy, 15*, 217–227.

Stoker, G. (2006). Public management: A new narrative for networked governance? *The American Review of Public Administration*, *36*, 41–57.

Talbot, C. (2009). Public value – The next "big thing" in public management? *International Journal of Public Administration*, *32*, 167–170.

Van de Ven, A., & Delbecq, A. L. (1971). Nominal versus interacting group processes for committee decision-making effectiveness. *The Academy of Management Journal*, *14*, 203–212.

Williams, K., Harkins, S., & Letane, B., (1981). Identifiability as a deterrent to social loafing: Two cheering experiments. *Journal of Personality and Social Psychology*, *40*, 303–311.

Zanjonc, R. B. (1965). Social facilitation. *Science*, *149*, 269–274.

Zellner, A. (2007). Philosophy and objectives of econometrics. *Journal of Econometrics*, *136*, 331–339.

Table 1. Realized outcomes for the two Special Events.

PIS Case	Question	Outcome
PIS A - Q1	Percentage of eligible students that eventually bought a portable computer.	87.0%
PIS A - Q2	Number of weeks it took 50% of the eligible students to participate in the scheme.	3.5 weeks
PIS B - Q1	Percentage of eligible parents that have eventually been trained.	52.2%
PIS B - Q2	Percentage of eligible parents that eventually received certification.	31.3%
PIS B - Q3	Percentage of eligible parents that obtained broadband Internet access using the funding provided by the scheme.	15.0%

Table 2. Group of experts.

Group	**A**	**B**	**C**	**D**
Level of expertise	**UJ**	**s-SA**	**D**	**IG**
Low	11	12	4	0
High	9	8	6	5
ALL	**20**	**20**	**10**	**5**

Table 3. Unaided Judgment (Group A).

Question	Expertise	MAE	RAE	Intervals Accuracy Rate	Standard Deviation
PIS A - Q1	Low	11.5%	1.00	54.5%	0.12
	High	8.3%	0.73	77.8%	0.09
	ALL	10.1%	0.88	65.0%	0.11
PIS A - Q2	Low	12.05	1.00	27.3%	12.46
	High	1.94	0.16	55.6%	2.57
	ALL	7.50	0.62	40.0%	10.87
PIS B - Q1	Low	23.7%	1.00	27.3%	0.26
	High	25.5%	1.07	33.3%	0.24
	ALL	24.5%	1.03	30.0%	0.25
PIS B - Q2	Low	20.8%	1.00	27.3%	0.23
	High	13.8%	0.66	33.3%	0.10
	ALL	17.6%	0.85	30.0%	0.20
PIS B - Q3	Low	23.2%	1.00	18.2%	0.24
	High	20.4%	0.88	33.3%	0.20
	ALL	21.9%	0.94	25.0%	0.23

Table 4. Semi-Structured Analogies (Group B).

Question	Expertise	MAE	RAE	Intervals Accuracy Rate	Standard Deviation
PIS A - Q1	Low	13.1%	1.14	66.7%	0.19
	High	10.1%	0.88	50.0%	0.12
	ALL	11.9%	1.04	60.0%	0.17
PIS A - Q2	Low	5.08	0.42	25.0%	6.68
	High	5.13	0.43	12.5%	4.92
	ALL	5.10	0.42	20.0%	6.04
PIS B - Q1	Low	22.8%	0.96	41.7%	0.23
	High	19.1%	0.80	37.5%	0.11
	ALL	21.3%	0.90	40.0%	0.19
PIS B - Q2	Low	15.5%	0.74	25.0%	0.18
	High	11.3%	0.54	75.0%	0.13
	ALL	13.8%	0.66	45.0%	0.17
PIS B - Q3	Low	26.4%	1.14	25.0%	0.28
	High	31.9%	1.37	0.0%	0.24
	ALL	28.6%	1.23	15.0%	0.27

Table 5. Analogies produced in s-SA per level of expertise.

		s-SA	
PIS	**Level of Expertise**	**Number of Analogies**	**Mean Similarity Rate of Analogies**
PIS A	Low	0.75	7.2
	High	0.88	6.6
PIS B	Low	0.50	6.5
	High	1.13	7.0

Table 6. The Delphi method (Group C).

		1st Round				2nd Round		
Question	**Expertise**	**MAE**	**RAE**	**Prediction Interval Accuracy Rate (%)**	**Standard Deviation**	**MAE**	**RAE**	**Standard Deviation**
PIS A - Q1	Low	2.5%	0.22	100.0%	0.03	2.5%	0.22	0.03
	High	8.3%	0.73	66.7%	0.10	6.7%	0.58	0.07
	ALL	6.0%	0.52	80.0%	0.08	5.0%	0.44	0.06
PIS A - Q2	Low	1.25	0.10	50.0%	1.12	1.25	0.10	1.12
	High	6.00	0.50	50.0%	7.65	5.67	0.47	5.72
	ALL	4.10	0.34	50.0%	6.42	3.90	0.32	5.00

PIS B - Q1	Low	14.9%	0.63	75.0%	0.17	14.9%	0.63	0.17
	High	14.0%	0.59	66.7%	0.12	14.0%	0.59	0.12
	ALL	14.3%	0.60	70.0%	0.14	14.3%	0.60	0.14
PIS B - Q2	Low	16.3%	0.78	25.0%	0.17	16.3%	0.78	0.17
	High	18.4%	0.88	33.3%	0.27	7.9%	0.38	0.13
	ALL	17.5%	0.84	30.0%	0.24	11.2%	0.54	0.15
PIS B - Q3	Low	22.5%	0.97	25.0%	0.21	22.5%	0.97	0.21
	High	28.4%	1.22	50.0%	0.31	26.7%	1.15	0.19
	ALL	26.0%	1.12	40.0%	0.28	25.0%	1.08	0.20

Table 7. IG group forecast vs. the average error of the individuals participating in this subgroup.

	IG			Individuals participating in IG		
Question	**AE**	**RAE**	**Intervals Accuracy Rate**	**MAE**	**RAE**	**Intervals Accuracy Rate**
PIS A - Q1	3.0%	0.26	100.0%	7.5%	0.66	40.0%
PIS A - Q2	2.50	0.21	100.0%	4.00	0.33	0.0%
PIS B - Q1	2.8%	0.12	100.0%	14.9%	0.63	60.0%
PIS B - Q2	11.3%	0.54	100.0%	10.1%	0.48	60.0%
PIS B – Q3	15.1%	0.65	0.0%	20.8%	0.90	0.0%

Table 8. Analogies recalled in IG.

	IG	
PIS	**Number of Analogies**	**Mean Similarity Rate of Analogies**
PIS A	3	6.3
PIS B	4	7.2

Table 9. Methods comparison (APE%).

	UJ (%)	**s-SA (%)**	**D (%)**	**IG (%)**
PIS A - Q1	11.6	13.7	*5.7*	**3.4***
PIS A - Q2	214.3	145.7	111.4	**71.4***
PIS B - Q1	46.9	40.8	27.4	**5.4***
PIS B - Q2	56.2	44.1	**35.8***	36.1
PIS B - Q3	146.0	190.7	166.7	**100.3***
Mean MAPE	95.0	87.0	69.4	**43.3***
Relative improvement (to UJ)	Benchmark	8.4	27.0	54.4

** These numbers indicate the most accurate forecasts for the relevant questions*

Table 10. Statistical accuracy of different treatments (RAE of individuals).

	No Grouping (40 experts)	**Delphi Group** (10 experts)	**Interaction Group** (5 experts)
PIS A - Q1	0.96	0.44	0.26
PIS A - Q2	0.52	0.32	0.21
PIS B - Q1	0.96	0.60	0.12
PIS B - Q2	0.76	0.54	0.54
PIS B - Q3	1.09	1.08	0.65
Mean RAE	0.86	0.60	0.36
GM RAE	0.83	0.55	0.30

Table 11. Statistical accuracy of the different treatments (RAE of individuals).

	No Structured Approach (30 experts)	**Semi Structured Analogies (25 experts)**
PIS A - Q1	0.76	0.88
PIS A - Q2	0.53	0.38
PIS B - Q1	0.89	0.74
PIS B - Q2	0.85	0.64
PIS B - Q3	1.00	1.11
Mean RAE	0.81	0.75
GM RAE	0.79	0.71

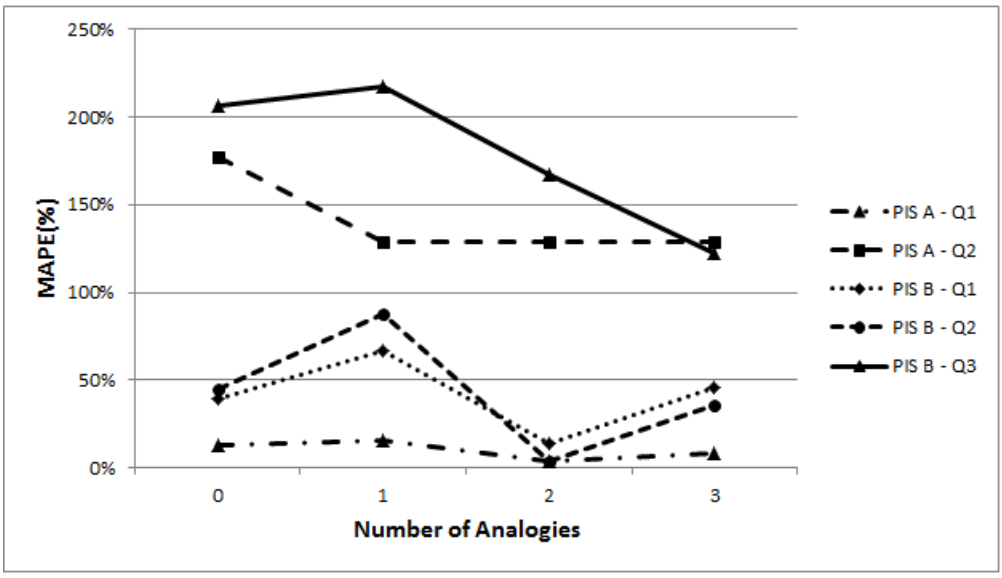

Figure 1. Forecasting accuracy of s-SA vs. the number of analogies recalled.

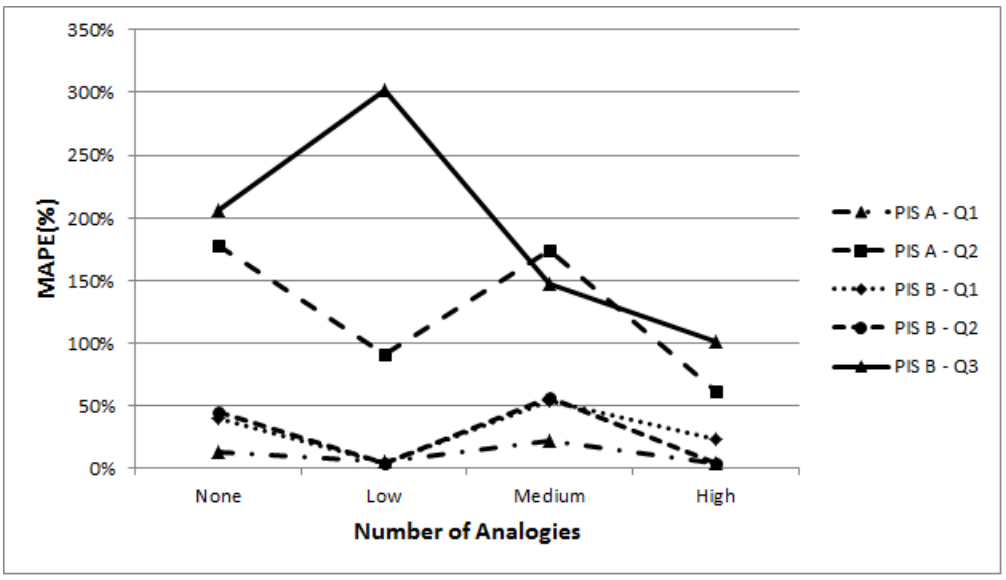

Figure 2. Forecasting accuracy of s-SA vs. similarity of analogies.