Use of expert knowledge to anticipate the future: Issues, analysis and directions[1]

Fergus Bolger

Minerva Consulting, Nottingham
62, Packman drive
Ruddington
Nottingham NG11 6GE
UK
E-mail: fbolger42@gmail.com

George Wright

Strathclyde Business School
University of Strathclyde
199 Cathedral Street
Glasgow G4 0QU
E-mail: george.wright@strath.ac.uk

[1] The authors would like to thank Paul Goodwin and Nigel Harvey for their suggestions following review of this manuscript, and Gene Rowe for his comments on an earlier draft.

Use of expert knowledge to anticipate the future: Issues, analysis and directions

Fergus Bolger and George Wright

Abstract

Unless the anticipation problem is routine and short-term, and objective data are plentiful, expert judgment will be needed. Risk assessment is analogous to anticipation of the future in that models need to be developed and applied to data. Since objective data are often scanty, expert knowledge elicitation (EKE) techniques have been developed for risk assessment that allow model development and parametrization using expert judgments with minimal cognitive and social biases. Here, we conceptualize how EKE can be developed and applied to support anticipation of the future. Accordingly, we first define EKE as an entire process, that involves considering experts as a source of data, and that comprises various methods for ensuring the quality of this data, including – selecting the best experts, training experts in normative aspects of anticipation, and combining judgments of several experts – as well as eliciting unbiased estimates and constructs from experts. We detail aspects of the papers that constitute the Special Issue and analyse these in terms of the stages within the EKE future-anticipation process that they address. We identify the remaining gaps in our knowledge. Our conceptualization of EKE to support anticipation of the future is compared and contrasted with the extant research effort into judgmental forecasting.

Introduction

Broadly speaking, anticipating the future is about applying some model of the world (that connects the past and the present to the future) to a set of data to produce predictions regarding the future state of the world; these data can be either quantitative or qualitative, similarly models too can be quantitative or qualitative (e.g. statistical time-series versus causal models). An important part of anticipation is the assessment of uncertainty regarding predictions because decision makers and planners need to know, for instance, how much resource to allocate to particular eventualities, or to reducing uncertainty by collecting more data. Models, data and uncertainty all lie on continua between subjective and objective.

Psychologists and decision scientists have catalogued a number of potential weaknesses in human judgment and decision making that apply to experts and laypeople alike. These weaknesses include the use of mental shortcuts or heuristics that can result in biases in both predictions and assessments of uncertainty surrounding those predictions (e.g. Tversky & Kahneman, 1974; Kahneman & Tversky, 1984). All stages of the anticipation process require some input from humans – to a greater or lesser extent – preferably from those with some relevant expertise. For example, experts must recognize the need to make predictions, describe the problem, formulate a model, identify and search for data, chose an appropriate method to make a model operational (e.g., a particular time-series method or a particular scenario development method), apply the method to make predictions (involving integration of information from different sources), assess uncertainty, and evaluate the whole anticipation process and its outcomes. There is therefore ample opportunity for error and bias to affect the quality of anticipation of the future at each stage (see e.g. Bolger and Harvey, 1998).

One area that has seen quite a large amount of attention in recent years is the development of methods to elicit estimates of parameters of risk assessment models from experts as applied, for instance, to hazards from earthquakes, volcanoes, climate change and threats to the food supply (Aspinall, 2010; Bolger et al., 2014; Budnitz et al., 1997; US EPA, 2009; Reilly et al., 2010). These elicitation methods applied to risk analysis have come to be known as "expert knowledge elicitation" (EKE) techniques.

What is EKE?

EKE is an emerging field and, as such is not yet well defined. As we have just indicated, the term "EKE" has so far usually been applied in quite a narrow sense to elicitation methods applied to risk analysis: these methods are chiefly concerned with eliciting estimates of uncertain **quantities**[1], usually in the form of **probability distributions (see, e.g., O'Hagan et al., 2006)**, from **groups** of **experts**: several of the papers in this Special Issue define EKE in these terms. It is important to stress, however, that EKE is *not* a single method, or even a methodology, but an **approach** that encompasses, but is not restricted to, several extant modelling approaches and their linked methods, some of which we will describe later. As an approach, EKE has some defining characteristics that we will now discuss in turn: this analysis permits a broadening of the definition of EKE so that it is more applicable to the range of uses of expert knowledge for anticipation.

Foremost of these general characteristics, EKE is a **practical** enterprise that applies the findings of social science research to the problem of extracting the best possible estimates from people in the face of lack of hard evidence, and with the presence of uncertainty, to be used for specific purposes (e.g. risk analysis, decision and policy making, and, indeed, anticipation of the future). Related to this, several presentations of EKE (e.g., Bolger et al, 2014; Budnitz et al, 1997; Cooke & Goossens, 2008; Knol et al., 2010) take the form of guidelines or protocols that embrace the entire **process** of eliciting judgments for the given purpose, even if some parts of this process have received more attention (in terms of both fundamental research and practical implementations) than others. We organize the following discussion of common principles and features of EKE by where they fit into this overall process.

[1] EKE, as defined here, should not be confused with knowledge elicitation for expert systems. The latter requires expert knowledge and/or judgment processes to be verbalizable, while the former simply requires verbalization of the end-product of such knowledge and processes, which psychological research suggests that experts often do not have access to (e.g. Berry & Broadbent, 1984; Nisbett & Wilson, 1977), perhaps because their expertise derives from very many exemplars or instances acquired through experience (e.g. Shanks, 1997).

The EKE process as applied to anticipation

Since our current concern is to extend EKE to problems of anticipation we have decided to use Armstrong's (1985) stages of the forecasting process – which can be applied to anticipating the future more generally – as a template, rather than any of the characterizations of the EKE process created for other purposes. Armstrong's first stage is what he refers to as "Implementation", where the anticipation problem is formulated and defined (i.e. a model is created and sources of relevant data are identified). Second, a particular method is chosen to apply the identified model to the data in order to produce predictions, and/or to gather judgment data relevant to important parameters of the model. The third stage is the application of the chosen method to anticipate the future: if more than one anticipator (person or machine) is involved then there may be comparison, combination and adjustment of predictions in the light of other predictions at this stage. Also it is at this stage where uncertainty is usually assessed, although it has been argued that it should be assessed from the outset as part of problem formulation (Knol et al. 2010). Finally, at the fourth stage, the success of the anticipation process and outcomes may be evaluated and documented. Next, we expand discussion of these stages and document the connections between our Special Issue papers and particular stages.

Stage 1: Implementation

This initial stage, where the need for foresight is recognised (i.e., there is recognition that the future may not be an exact replica of the past), a model formulated, relevant variables identified, and data search initiated has received relatively little attention in EKE applied to risk analysis, with most work in effect starting with a defined problem and the specific need for a particular type of (expert) judgement pre-identified. In contrast, model-building is integral to Scenario Planning and there has also been some work concerned with monitoring and detection, see below.

a). Monitoring and detection (of indicators of an emerging future)

Following from a couple of papers by Paul Schoemaker (Schoemaker & Day, 2009; Schoemaker et al., 2013), one of the founding fathers of Scenario Planning, there has been an interest in looking for so-called "weak signals" which are pieces of information that, in themselves, appear as just noise but are indicators of significant future events or trends when seen in the context of other information, or looked at differently. This idea has been taken-up by both the private and public sectors in the UK who have begun long-term monitoring or "horizon-scanning" defined as:

"A systematic examination of information to identify potential threats, risks, emerging issues and opportunities…allowing for better preparedness and the incorporation of mitigation and exploitation into the policy making process." (Day, 2013, p. 2).

Meissner et al. (this issue [1]) propose a new method for horizon scanning that they call "360^{o} Stakeholder Feedback" that attempts to reduce bias arising from an excessive internal focus, which can lead to "blindspots" in foresight, by eliciting the judgments of external stakeholders as well as organization insiders. Derbyshire and Wright (this issue [2] also criticize horizon-scanning, as it is currently practiced, as being over-simplistic - since it is based on the identification of a single type of cause and there is therefore, implicitly, a simple causal chain of unfolding events leading to a particular outcome. They propose, as a frame-broadening solution, a more thorough description of the present, plus consideration of additional types of causes.

b). Problem perceptions, and the development and documentation of mental models

There seems potential here to use methods for capturing more qualitative aspects of expert knowledge (e.g. arguments, classifications, and perceived causal relations) for model development. Tools available for this include influence diagrams (e.g. Howard, 1989; Oliver & Smith, 1990); cognitive maps (e.g. Eden, 1988); card sorts and repertory grids (e.g. Bolger et al., 1989), or simply “brainstorming” (e.g. Rowe & Bolger, in press used brainstorming as the first round of a Delphi process, in order to identify those precursor factors seen as fundamental to the prediction problem). All these approaches are concerned with ways in which to elicit and document individual and group-based perceptions. The Delphi approach, especially, facilitates challenge and subsequent change in perceptions or viewpoints – allowing the in-group evaluation of individual frames (see Bolger and Wright, 2013)[2].

Scenario Planning facilitates the modelling of perceptions and viewpoints and provides a documentation of this. For example, Derbyshire and Wright (this issue [2]) show how the Intuitive Logics (IL) approach can be used to identify an issue of concern - around which anticipations of the future are subsequently developed using predetermined elements, critical uncertainties, and perceived causal relationships between both. These authors propose to improve on IL with a more thorough description of the present, plus consideration of additional causes and a new focus on transformational change and causal loops. Further, Meissner et al. (this issue [1]) use brainstorming within a Delphi-like process which is nested within a Scenario Planning framework to uncover potential influences on the future development of an organization’s business environment.

Rich qualitative expert knowledge can also be elicited in a traditional Delphi process in the form of supporting arguments or rationales for quantitative judgments: we (Bolger & Wright, 2011) have argued that feedback of rationales – in other words, the particular models underpinning anticipations of the future – is key to opinion change for the better in Delphi. In the IDEA protocol outlined by Hanea et al. (this issue [3]), although there is no explicit model development, experts can research a given problem for a couple of weeks individually before submitting their predictions – and uncertainty assessments – and justifications thereof. The problem is then reassessed on receipt of feedback (aggregated by

[2] With respect to the distinction made in the previous footnote, these methods require knowledge (and sometimes reasoning processes) to be verbalizable; thus there is a potential limit to their effectiveness. However, there is some reason to believe that by no means all expert knowledge and processes are exemplar-based and thus inaccessible to them (see, e.g., Karlsson, Juslin & Olsson, 2008). Further research is needed to determine the conditions under which eliciting this more qualitative expertise is useful for improving judgment quality, and when it is not.

a facilitator) regarding others' judgments and rationales: these rationales can potentially allow participants to build a joint (implicit) model of the problem.

c). Identify data and experts

Once the model has been developed it should be clearer what sort of data are needed (e.g. what parameters in the model need instantiating): initial data search and collection can begin. It should also become apparent at this stage whether input from additional experts[3] will be required and thus a need for EKE identified: note that **experts can be considered as data sources** as well as potential contributors to modelling and uncertainty assessment. This brings us to the second defining feature of EKE which is its concern with **identifying and measuring expertise**. In some cases it will be clear at this stage who the relevant experts are or, at least, what the potential pool of experts is. In most cases, though, the selection of experts is bound up with further analysis of the problem, and the choice of methods. For this reason we will defer further consideration of expert selection until later.

Stage 2: Choice of Method

This Stage can be seen as a further refinement of Stage 1 on the basis of data requirements identified there. In particular, characteristics of tasks, experts, and methods must be matched to each other. Although we will treat this here in a linear fashion, it is really an iterative process of examining characteristics of task, experts and method to establish the best fit.

a). Task analysis.

Perhaps the most important feature of foresight tasks relevant to both the selection of data and methods is how far ahead we want to anticipate. It is clear that the longer the anticipation **horizon,** the harder it is to predict what will happen. The longer the horizon for anticipation, then, the less we can rely on data sets of previously-collected data on the forecast variable and the more we will need to use expert judgment to synthesise qualitative and quantitative information to aid prediction or anticipation.

Uncertainty will also increase with time horizon, hence the need to explicitly represent this uncertainty will also increase. In short-range forecasting with good historical data and stable environments there is neither need for expert judgment – beyond searching for data and choosing the method – nor judgment of uncertainty (estimates of uncertainty can be made from statistical analysis of the data e.g. its variability). However, such conditions will rarely pertain in practice – the horizon will be long, or data will be scanty, or the environment will be changing, or any combination of these things – expert judgement will be required and some of which will be judgement of uncertainty. Thus **how best to represent uncertainty and elicit it from experts is** a central concern of EKE, reflected in several of the papers in this Special Issue.

[3] We presume that there are already an expert or experts in the foresight problem involved.

Once expert judgement is involved it is not simply the case that all tasks are equal with regard to uncertainty assessment. Bolger and Wright (1994) proposed that the quality of uncertainty judgement (i.e. its reliability and validity) depends to a large degree on the "**learnability**" of the judgement task, that is, the extent to which experts can learn to make reliable and valid assessments of the uncertainty surrounding judgments of target variables. Bolger and Wright also propose that tasks must be "**ecologically valid**", meaning that they correspond to the professional ecology of the experts -- thus the experts' acquired experience is valid for that task – and can lead to elevated performance. In many studies of expert judgment the tasks are not ecologically valid.

While the foresight horizon influences the type of model that we might use – and thence the kind of data – the **availability of good quality objective data** may also independently impact upon the type of model. Thus if there is little or no relevant historic data (e.g. when forecasting the demand for a new technological product) one may be forced to rely on models that have a predominantly judgemental character, even if the horizon is short.

In Meissner et al. (this issue [1]) experts generate then rate factors potentially affecting the construction industry in the future (i.e. quantitative and qualitative data are elicited). The uncertainty analysis proposed here is not an input to the anticipation model, as is usually the case, but applied to the model's output: weak signals and foresight "blind spots" are identified through examination of the factors generated, and their ratings. Meissner et al's approach is not typical of Scenario Planning, where uncertainty is not usually quantified as a numerical assessment. In Derbyshire and Wright (this issue [2]) uncertainty is represented by the complexity of the causal model(s) developed. In their approach, greater complexity is permitted relative to IL by virtue of additional types of cause (formal, material and final) and causal loops. However, Derbyshire and Wright claim that uncertainty may be reduced by more detailed analysis of predetermined elements of the future.

Hanea et al. (this issue [3]) do not analyse the characteristics of their task (geopolitical forecasting) or data in detail, however, it is noted that the method requires relatively short-term forecasts because actual realizations are needed in order to measure performance for expert selection and weighting. Further, certain specifics of the method – for example, the elicitation technique, and performance measures – require judgments of likelihood of dichotomous events. Wilson (this issue [7]) analyses judgments for tasks drawn from the Delft database of studies using Cooke's method (Cooke, 1991), which,, like Hanea et al's IDEA, needs realizations of earlier predictions to select and weight experts: in this case for earlier forecasts that are similar to the target (known as "**seed questions**"). Wilson uses the experts' performance on the seeds to investigate characteristics of experts and tasks (in particular dependencies between experts and/or items). The quantity and quality of seeds (in particular the closeness of the epistemic relationship of seeds with the target – see e.g. Bolger & Rowe, 2015a,b) have implications for choice of both experts and aggregation method (see below). Note that Wilson suggests that short-term forecasts could be used as seed for long(er)-term forecasting

Uncertainty is present in the time-series extrapolation task used by Onkal et al. (this issue [4]) as 90% confidence intervals in the advice. However, the advisor's track record for forecasting was available to some: the advisors were artificial experts with either high or low forecast error), which reduces uncertainty. In practice, though, we will usually have little or no objective data regarding the performance of experts – with the exception of short-term forecasting tasks, such as weather forecasting, stock market forecasting and weekly demand forecasting. Meanwhile, Alvarado et al. (this issue [5]) examine the effects of manipulating some task characteristics: "need for correction" (i.e. when there is room for improvement on statistical forecasts), and credibility of statistical forecasts (i.e. whether they are fair, unbiased, complete, accurate, and trustworthy). The authors examine the effects of their manipulations on judgmental extrapolation performance for what is a relatively ecologically valid task. Petropoulous et al. (this issue [6]) also examine judgmental extrapolation of time series. Forecasters received rolling feedback about their performance (bias or accuracy) and values of realizations: the effects of this feedback on performance were examined. In all three studies the authors used real series containing features such as trend and seasonality, which added to ecological validity, but in Onkal et al. and Petropoulous et al., but not Alvarado et al. which was a field study, the forecasters had no contextual information, so could not use their experience to the full. Note, however, that in the latter paper the emphasis was on how people evaluate expertise when presented with forecasts – not on how they use their own inherent expertise to make these forecasts.

b). Expert selection.

At the centre of the EKE approach is the **expert**. As we already indicated, the expert is a source of data in EKE and the raison d'etre of EKE procedures is to maximize the reliability and validity of judgments elicited from experts. As a first step in this goal attention should be paid to **selecting** experts with the "best" knowledge of the domain in question: some methods for accomplishing this have been suggested (e.g. Bolger & Wentholt, 2014; Meyer & Booker, 2001).

Expertise can be regarded as a property of individuals due to extensive practice and/or innate characteristics. Some of these might be manifest on an expert's CV as academic and professional qualifications; years of professional experience; number of publications, patents and citations; prizes and so on. Another source is peer opinion in the form of references or, alternatively, by their responses to a questionnaire such as the Generalized Expertise Measure (GEM: Germain & Tejeda, 2012). Although the GEM's 16-item scale contains objective indicators of expertise (e.g. education, training and qualifications) it also contains some more subjective items (e.g. self-assurance, potential for self-improvement and intuition). It should be noted that the reliability and validity of these peer assessments is not yet well-established, and that some characteristics that may be associated with expertise, such as confidence or self-assuredness, are not necessarily desirable (i.e. it can bias both personal judgements and result in the wielding of undue influence in groups).

An important distinction in EKE is between **substantive expertise**, which is domain or content knowledge (usually associated with the notion of expertise in common parlance), and **normative expertise**, which are agreed methods (e.g. data collection techniques), benchmarks (e.g. professional standards) and measures (e.g. expressing uncertainty as probabilities). Although in EKE we are usually primarily interested in substantive expertise, possession of appropriate normative expertise can also be important, for instance, in aiding communication between experts, and between experts and elicitors. Possession of normative expertise with respect to probabilities is particularly useful when eliciting uncertainty. It is not particularly difficult, however, to train experts in this regard so long as they are willing to put in the time (although they often are not): the effectiveness of such training for increasing the quality of uncertainty assessment has yet to be established, though.

Another way that expert performance is commonly defined is by consensus within a particular group. This is sometimes referred to as "**social expertise**" and may be manifest through status symbols (e.g. titles, honorifics, job role) and a high media profile. Bolger and Wentholt (2014) contend that social expertise is a poor proxy for true expertise and propose that indicators of social expertise not be used as the sole means for identifying experts: in lieu of a bespoke expertise assessment instrument, CV's should be preferred to peer-assessment, which in turn should be preferred to self-assessment, which is better than social expertise.

Meissner et al. (this issue [1]) utilize a role-based search to find experts internal to the organization (managers and others involved in strategic decision-making) and external to it (senior personnel in stakeholder organizations). Hanea et al. (this issue [3]) recruit an initial convenience sample of volunteers via the internet but in subsequent surveys they performed some selection on the basis of performance in the first year (a "Supergroup" was formed and its performance relative to the other groups was assessed). In Onkal et al. (this issue [4]) expertise (as advice) could be sought by forecasters on basis of either the relevant experience or status of the advisor (i.e. social expertise): effects of the basis of expertise on advice utilization was investigated as a function of expertise of advisees (novices: undergraduate students taking a forecasting course *versus* professionals who "regularly" give or receive relevant advice with 7-12 years' experience).

Alvarado et al. (this issue [5]) Distinguish between a priori indicators of expertise (e.g. things that can be gleaned from a CV or peer assessment i.e. relating to the main objective characteristics of experts: specialized domain knowledge, and outstanding and consistent performance) and on-task measures (i.e. performance data). In this case the former (role and peer assessment via interviews with a "key contact" in the organization and GEM also completed by the key contact, respectively) were used to categorize experts, while measures of forecasting ability (e.g. APE) were dependent variables. The authors suggest that expert knowledge that adds value to statistical forecasts tends to come from people in "job positions that are in permanent contact with… unmodeled environmental information". Years' experience in company and with products were also measured. Finally, while Wilson (this issue [7]) does not explicitly address the issue of expert selection, it appears that one could potentially pick the experts on basis of patterns of dependency revealed by answers to seed questions (e.g. those who show greatest consistency across answers could be included, or given greater weights).

c). Choosing the method.

Here we regard both means of collecting data from experts, and ways of using data from any source to anticipate the future, as methods. Thus methods include all EKE and Scenario Planning protocols (e.g. Cooke's method, Delphi, IL– to be described shortly), statistical forecasting techniques (e.g. time-series decomposition, Box-Jenkins) and pure judgment methods (e.g. Charting, time-series extrapolation). Mixtures of judgmental and statistical approaches, such as judgmental adjustment of statistical time-series forecasts, we also consider as methods[4].

Choice of method will depend on the analyses performed above. For example, if plentiful good quality data are available for many years in the past, and the prediction horizon is short to medium term, then statistical forecasting methods would be indicated as the primary method, perhaps with some judgmental adjustment. Alternatively, at the other extreme, if predictions are to be made long into the future about events for which there is little relevant historical precedent then a more judgmental approach (and thus greater role for experts) would be indicated.

Another central characteristic of EKE, that has so far been implicit in our discussion, is that knowledge will usually be elicited **from more than one expert**. Many EKE concerns arise from this feature, in particular, there is a preoccupation about how best to **aggregate** several experts' judgments into a single one to be used in decision and policy making (as is common practice even though there are arguments for not doing so, especially if there is disagreement amongst experts, (e.g., Morgan, 2014).

There are three basic approaches to aggregation: **behavioural**, **mathematical**, or **mixed** (a combination of the other two).

In behavioural aggregation, experts interact (freely or under the guidance of a facilitator), and (hopefully) some consensus will be finally achieved: this consensus forecast or opinion is what is then used for policy making. In mathematical aggregation experts make their judgements individually and then these are combined into a single forecast by averaging[5]: whether this averaging should be performed using differential or equal **weights** is hotly debated (see e.g. Bolger & Rowe, 2015 a, b and commentaries). Briefly, to differentially weight opinion requires a sound basis for weighting: "sound" because equal weighting has been shown to be generally a better bet than using weak or noisy criteria to determine the

[4] Methods are not the same as models. In some cases this is obvious, for instance, judgmental extrapolation is a method of forecasting that, unless random, is based upon the judge's mental model relating past and present to future (e.g. "tomorrow's weather is most likely the same as today's"). In other cases we might call the mental model the same name as the formal method (e.g. "a regression model") but this is misleading: although some formal methods specify the form of the relationship between the past/present and future, it is not an operational method that makes operational a mental model until it has been parameterized for the particular problem under consideration (e.g. predictors identified and betas fitted to them in multiple regression).

[5] Mathematical aggregation can also be used to combine the judgments of experts with statistical forecasts.

weights (e.g., Bolger & Rowe, 2015a, and commentaries; Clemen & Winkler,1999). Some criteria have not been shown to be sound, such as using citations of published work or self-assessed ability (e.g., Burgman et al., 2011; Cooke et al., 2008) However, proponents of Cooke's method, which uses as the weights ability to answer a set of seed questions – where the true answer is known, and which are related to the target variable to be assessed – provide evidence that this procedure outperforms equal weighting as an aggregation procedure (Cooke et al. 2014, Eggstaff et al., 2014). Bolger and Rowe (2015b) dispute this evidence and argue that the jury must remain out until further research is conducted.

Statistical aggregation is valuable because it eliminates random noise in judgments but its value decreases with each additional expert that is added and with increasing lack of independence between the knowledge of different experts: an issue that is discussed by Wilson ([7] this issue). Further to this, when experts who have overlapping knowledge get together to have a discussion in order to make a decision, they tend to discuss the knowledge that they have in common rather than the individual knowledge that each one of them can contribute to the group. This may be one reason why group decision making is not as effective as it should be (see e.g., Larson, Christensen, Franz & Abbot, 1998; Stasser & Titus; 1985, 2006).

Another difference between methods is that in some (e.g. in a Scenario Planning workshop) experts meet face to face, whereas in others (e.g. Delphi or Cooke's method) they usually do not: whether and how experts **interact** with each other are thus further important concerns of EKE. The goal of EKE is to elicit expert knowledge in an unbiased manner as possible, but freely interacting groups have been shown to be subject to bias such as Groupthink leading to "process loss" (i.e. the interaction process produces sub-optimal outputs relative to non-interacting, nominal groups (e.g., Rowe & Wright, 1999). On the other hand, the restricted information exchange in non-interacting groups means that "process gain" (i.e. the advantage to be had by different experts debating and pooling their knowledge) may not be as great as it could be.

Behavioural aggregation methods, such as the Sheffield method (Oakley & O'Hagan, 2010), attempt to avoid process loss through careful facilitation following a well-researched **protocol**[6]. Alternatively, the Delphi method (Linstone & Turoff, 1975) utilizes non-interacting groups where experts make their judgments individually, these are summarized and/or aggregated by facilitators (usually with equal weight given to each expert opinion) and fed back to the experts who are invited to revise their original opinions. This procedure continues until no significant change in opinions is observed, at which point there is generally sufficient consensus to justify putting forward the final aggregated judgment. In classic Delphi applied to forecasting, usually only quantitative feedback is given (e.g. aggregate point forecasts and confidence in these) but some recommend that reasons for judgments are also exchanged in order to facilitate process gain (e.g. Bolger & Wright, 2011).

[6] The Sheffield method is designed for eliciting knowledge from a group of experts in a face-to-face, *facilitated* workshop within which there is a two-stage process beginning with elicitation of individual judgements followed by a group discussion. The end result is normally an agreed single probability distribution representing the aggregated judgements of the experts. A detailed protocol is given that permits an untrained person to act as a facilitator see: http://www.tonyohagan.co.uk/shelf/

Dependency between the knowledge of experts is another issue of interest in the EKE literature that is relevant when multiple experts' judgments are considered. If there is high homogeneity in expertise then there is little to be gained by sampling multiple experts as they will tend to agree, further, many methods of mathematical aggregation assume independence between expert judgments (as discussed in this issue [7]) so excessive homogeneity can impact on the accuracy of judgment, unless the dependencies are accounted for. Consequently methods for introducing heterogeneity into groups of experts such as "Devil's advocacy" and "dialectical inquiry" have been proposed (e.g. Bolger & Wright, 2014), but if heterogeneity is too great then it can be difficult to reach consensus and aggregation may not make sense.

The Special Issue paper that most directly addresses the issue of how to choose the method is that of Alvarado et al. (this issue [5]). Here, three integration methods are compared (judgmental adjustment, 50-50 and divide-and-conquer) which constitute different ways of combining statistical and expert judgment. Results of the study show that judgmental adjustment is the best method if both expertise and need-for-correction are high and credibility of statistical forecasts is low. However, if situations can be modelled well statistically (e.g. where events impacting on the variable to be forecasted are under the company's control, such as planned promotions) the authors suggest that perhaps no EKE is needed.

The rest of the papers in the Special Issue focus on only one method, however, several make comparisons to other methods thus speak to the issue of method choice. Further, some of these methods address aggregation of expert judgment either with other experts, or with statistical forecasts. For example, Derbyshire and Wright (this issue [2]) describe the IL protocol for Scenario Planning, which is best described as a behavioural aggregation method. Experts interact in a workshop and thus exchange rich ideas about causes, trends and so forth: this should permit experts to self-weight their contribution to the final model. The authors compare this process to a Delphi where experts' rationales are given as feedback. The IDEA EKE method proposed by Hanea et al. (this issue [2]) is something of a hybrid between Delphi (iteration and anonymity of final judgments), the Sheffield method (structured interaction between experts in the first round), and Cooke's method (mathematical aggregation of final-round judgments using performance weights): the authors discuss the pros and cons of each approach (and also prediction markets – described below – which have also been used successfully for the geopolitical forecasting discussed by Hanea et al.).

Three of the Special Issue contributions (this issue [1], [4] and [6]) focus on forecasting by individuals so it may appear that the issues of aggregation and weighting do not arise. However, in the "360 degrees Stakeholder Feedback" method (this issue [1]) ratings of impact and uncertainty are averaged across experts (for clustering analysis which is then fed back to the experts), implying equal weighting of expert opinion. Further, behavioural aggregation presumably occurs in the final discussions between experts, when the results of the exercise are applied to strategy making. Also in Onkal et al. (this issue [4]) advice may be integrated with an individual forecaster's own judgment before making a final forecast. The authors of both papers consider how informal aggregation may be biased: in Meissner et al. (this issue [1]), their proposed method is designed to reduce the weight placed on the judgments of experts internal to the organization by explicitly eliciting the opinions of external stakeholders; in Onkal et al. (this issue [4]) experiments show that weights placed on advice (relative to the forecaster's own judgment) are influenced by a number of factors that often do not lead to the best forecasting outcomes, implying that the combination of forecasts and advice should not be left to the discretion of individual forecasters. These two papers discuss their approaches in relation to others such as the Delphi technique.

The analysis by Wilson (this issue [7]) suggests that dependencies – in particular positive ones – often occur between experts, probably due to shared knowledge and/or heuristic use (and thus experts may also have similar biases, such as overconfidence). Given these dependencies Wilson suggests that it is better to use Bayesian aggregation methods, which take them into account, than opinion pooling methods (such as Cooke's method)[7], which do not. However, if sufficient seed-variable judgments are available for a number of experts then one could potentially measure between- and within-expert dependencies and match experts to aggregation measures accordingly (e.g. if there are strong within and weak between dependencies then use Cooke's method, if the opposite pattern is found then use a Bayesian method). We suggest that an analysis of dependency could also be used to see if sufficient diversity of opinion exists amongst experts and, if not, then try to recruit more or generate diversity using the techniques like Devil's advocacy mentioned above.

Stage 3: Application of Method

This is where the method chosen in the previous stage is applied in order to anticipate the future. Since this is EKE, experts are central to any method that is applied, but expert involvement may either occur directly (e.g. expert judgement is used to extrapolate a time-series, or adjust a statistical forecast) or indirectly (e.g. experts judge values for parameters in a statistical model that is then used to make a forecast). Following from this it should be clear that an initial step in applying an EKE method must be finding quality expertise.

a). Screening and training experts

The goal of EKE is to reach a final set of judgments – in the current context, judgments regarding the future state of the world – that are as close to their eventual "realisations" as possible. Treating EKE as a piece of empirical research, one way to try and achieve this goal is, as with all empirical research, to collect as much data as possible: for EKE this means eliciting the knowledge of a large number of experts. However, it often takes many years of training and practice to reach the highest levels of skill and knowledge in a domain, so usually those with the highest expertise are in short supply. This leaves those wishing to conduct EKE with two basic strategies: sample a larger number of less skilled experts or

[7] The authors suggest that too much positive dependency can lead to overconfident aggregated forecasts – by opinion pooling – because they overestimate unique information each expert brings to bear.

sample a smaller number of highly skilled experts, perhaps eliciting more knowledge and/or to a greater depth.

The first strategy has been applied to anticipating the future, and makes use of the "wisdom of crowds" phenomenon, where the average of a crowd's judgment is close to the truth due to averaging out the noise due to idiosyncracy in individual judgments (2004). One example of this so-called "crowdsourcing" that has been used for forecasting is prediction markets. Here a great number of educated but non-expert people are involved and are paid on the basis of forecast accuracy. Prediction markets have been found to produce good forecasts in certain domains, such as prediction of geo-political events (Mellers et al., 2014).

The second strategy – in-depth elicitation with a few top experts – is the most usual approach. This is partly due to the aforementioned scarcity of experts, but also due to practicalities of applying some of the methods. For instance, the Sheffield method and Scenario Planning require experts to be brought to a particular place at a particular time for a workshop[8], and the larger the group the harder it is to facilitate the process effectively. For Cooke's method experts can be seen individually at different times and locations, but since it is preferred to interview each in person at some length, having large numbers of experts would be very time consuming. It has also been noted that there are diminishing returns to having more than about ten to twelve experts per group in Cooke's method, although a lower limit of around six is recommended (Aspinall, 2010; Cooke & Probst, 2006).While the Delphi method can be applied remotely, with considerable flexibility regarding timing (e.g Gordon & Pease, 2006), it can be difficult to achieve consensus with large groups. Further, providing feedback after each round from and to numerous experts without overwhelming the participants can be laborious and difficult for the facilitator[9].

In the unusual situation where there are more experts who are willing to take part in the elicitation exercise than you need, those experts with the most, and most relevant, expertise could be put onto a short-list. **Screening** for the short-list might be accomplished by asking candidate experts to answer some questions designed to test either their substantive or normative expertise, or both. For example, the seed questions asked of experts in Cooke's method can be considered as tests of both domain knowledge (i.e. knowledge related to the anticipation problem) and meta-knowledge concerning the uncertainty surrounding judgments (i.e. how realistic their probability judgments are): both of these assessments are combined to produce the weighting for an expert's opinion[10]. Often a threshold is set such that experts whose performance on the seed questions is below this level have their weights set to zero, meaning essentially that they are screened out of the EKE exercise. As an alternative to testing, potential experts can be asked to fill out a questionnaire regarding their normative and substantive expertise thus permitting both screening and assessment of training needs (see e.g. this issue [5] and Bolger & Wentholt's (2014) expert-skills questionnaire).

[8] As technology improves, bringing people physically together is becoming less of an issue, however, there are still technical (and temporal) constraints on the size of groups that can be managed.

[9] However, more experts can mean greater variation in opinion thus helping to reduce overconfidence: a compromise might be to induce variability into a smaller group.

[10] However, Bolger and Rowe (2015) argue that the weights in Cooke's method favour normative abilities (i.e realism of probability judgment) over substantive (i.e. domain knowledge).

Training is most usually related to expression of uncertainty (i.e. normative rather than substantive aspect of expert judgment). Another approach is to give all experts normative training as part of induction. Training could also be given in substantive expertise, though. For example, if anticipation would benefit from expertise from several different specialisms, then some training, for instance in terminology and basic concepts, could be given across specialisms so as to assist communication (i.e. knowledge exchange between experts). Such training in substantive issues might most easily be accomplished face-to-face in a facilitated workshop, but could also conceivably be achieved online, for instance in a Delphi process.

In Alvarado et al. (this issue [5]) the experts were sorted into relatively high and low expertise groups on basis of GEM (utilizing the knowledge sub-scale – which had high reliability – and overall GEM score, where high-scoring experts made more valid adjustments and forecasts than low). The authors comment that selection might be further improved if personality characteristics associated with good forecasting can be identified and a questionnaire tailored to specific domain knowledge requirements developed. The authors also suggest that there is room for training experts with outcome feedback and inducing healthy scepticism with regard to statistical forecasts. Meanwhile, Petropolous et al. (this issue [6]) explicitly focus on the issue of training experts (in substantive rather than normative aspects of the forecasting task). They provide outcome and performance feedback after each non-probabilistic judgmental-extrapolation forecast in an attempt to reduce documented biases. They conclude that training was successful in this respect, with feedback that revealed judgmental bias being particularly effective.

Hanea et al. (this issue [3]) use a crowdsourcing approach: thus they used a large number of relatively inexpert judges rather than a few judges high in expertise. There was no initial screening but participants were evaluated in the first year and sorted and weighted on the basis of this for the second year. Although there was no formal training (e.g. in probability, since probability distributions were elicited), participants were given an initial briefing by e-mail and/or phone regarding the study's purpose, question format, and possible biases: this included some example practice questions.

b). Application of method

As mentioned at the outset, anticipation involves applying a model to data to produce a prediction: the method determines how the model is applied to data and/or it produces data that can then become input to a model (e.g. probabilities and values of other parameters).

Research into judgmental forecasting has thrown up a number of psychological biases affecting forecast performance, however, it speaks little to what should be done about them to improve the use of judgment in foresight exercises (see e.g. Lawrence et al., 2006). In contrast, **a central concern of EKE is how to elicit judgments from experts that are free from bias**. These biases may arise from the way that data and models are presented to experts (so-called "framing effects" see, e.g. Kahneman & Tversky, 1984), from the use of heuristics by individual experts (e.g. anchoring and adjustment, availability, and representativeness), or from the social dynamics of groups (e.g. Groupthink, risky shift, and influence by dominant individuals), or any combination of the three. To this end EKE protocols are specifically designed to minimize such problems. For instance, the Sheffield protocol provides an ordering for eliciting estimates that reduces anchoring effects, Cooke's method uses a scoring rule designed to encourage the expression of true beliefs and thus reduces the tendency towards overconfidence, and the Delphi method's lack of direct interaction minimizes social biases (see e.g., Bolger et al., 2014).

As we have already indicated, **how best to elicit uncertainty** is of particular interest in EKE. One issue of concern here is a tension between how experts usually express uncertainty and how we would ideally wish them to, from the viewpoint of modelling uncertainty for use in foresight exercises. Many experts are reluctant to quantify their uncertainty, preferring instead to use natural language terms such as "highly probable", "little chance" and "quite likely". Unfortunately these terms are not used consistently even by a single expert on different occasions, so mapping the verbal terms onto numeric probabilities required as inputs to particular methods is problematic (see e.g., Dhami & Wallsten, 2005; Wallsten and Budescu, 1995).

Even if numeric probabilities are elicited they can take different forms. If the occurrence of an event is being predicted then likelihood might be given as a percentage between 50 to 100, where 50% means the expert thinks that the event is as likely to occur as not and 100% means the expert is sure it will happen. If uncertain quantities are being judged then the simplest method is for experts to be asked to give an interval around their best estimate within which the true value will fall with a given probability (e.g. 90%). Proponents of the Cooke and Sheffield methods prefer to go farther than simply eliciting a single interval, instead they ask for further intervals of different probabilities so that they can build up experts' probability distributions for each uncertain quantity (see this issue [3] and [7], and Bolger et al., 2014).

Overconfidence is a persistent bias found in both event probability judgments and interval judgments, including distributions (e.g. Glaser et al., 2013 ; Lichtenstein et al., 1980; Lin & Bier, 2008) The extent of bias depends to some extent on the way that probabilities are elicited. For example, overconfidence has been found to be less for smaller probability intervals (e.g. 75%) than bigger (e.g. 95%), or when experts are free to choose their own intervals (e.g., Teigen & Jorgensen, 2005). EKE is therefore concerned with how best to minimize overconfidence bias in expert probability assessment.

The generation of accurate probability judgments may be an individual trait. An assumption of Cooke's method is that some experts are better than others with regard to the realism of their judgments: measures of this ability on the seed variables are then used to weight experts for aggregation of judgments regarding the target variable. But there is little evidence for stability of probability-assessment skill over time or task (see e.g. Bolger & Rowe, 2015b) although Tetlock and others (Mellers et al., 2015; Tetlock & Gardner, 2015) found some geo-political forecasters seemed to consistently outperform others. Interestingly, these individuals were laypeople with no formally-acquired designation of differential expertise.

Domain of expertise may also influence the ability to assess probability. For instance, horse bettors, bridge players, and weather forecasters have been found to be quite realistic in their probability assessments (Hoerl & Fallin, 1974; Keren, 1987; Murphy & Brown, 1985). Bolger & Wright (1994) attribute this to the fact that experts in these domains have rapid feedback about their performance (i.e. learnability is high) and also have regular experience at expressing their uncertainty in a formal manner (i.e. ecological validity is high). More commonly, though, feedback will be much delayed or essentially unusable for probability assessment (e.g. each event is unique, or policies affect outcomes): we argue that probability judgment is unlikely to be reliable in such cases.

On the topic of **feedback**, another concern in EKE is what feedback to provide to experts during elicitation. In the Delphi method, feedback of assessments is central to the process of "virtuous opinion change" (Bolger and Wright, 2011) whereby quality of judgment improves over rounds: the richer the feedback the better, so feedback of rationales for judgment is encouraged. In the Sheffield method and Scenario Planning, feedback from other experts and the facilitator can be instantaneous due to the face-to-face nature of these methods. In Cooke's method provision of feedback *after* an elicitation exercise is recommended but not an integral part of the process (see Bolger & Rowe, 2014b).

Uncertainty is elicited in some of the Special Issue papers. For example, in Meissner et al (this issue [1]) experts first generated factors impacting on the future development of the organization, then rated the degree of impact and the uncertainty of this impact on 10-point scales. Hanea et al (this issue [3]) elicit uncertainty regarding future events in an order designed to reduce bias (first, lowest probability; second, highest probability; and third best probability). Wilson (this issue [7]) makes use of the Delft database of studies utilizing Cooke's method which elicits probability distributions most commonly using three quantiles (the 5th, 50th and 95th).Cooke's method attempts to debias probability estimates in two ways: first it uses a "proper scoring rule" (see e.g. Brier, 1950) to ensure that judges provide estimates that truly reflect their beliefs; and second it weights experts' performance on seed-variable evaluations so that the best calibrated (most realistic and, potentially, least biased) get the most weight. However, Bolger and Rowe (2014a) question the effectiveness of both these strategies. How should suitable "seed problems" be identified? They have to be similar to the substantive problem in some way but how should this similarity be measured? Further, as we have already noted, Wilson suggests that the choice of aggregation method itself has implications for de-biasing, with choice of a Bayesian method potentially producing less overconfidence than opinion-pooling methods such as Cooke's.

Although other papers in the Special Issue do not elicit uncertainty, they all speak to the issue of de-biasing. For example, Alvarado et al. (this issue [5]) describe the de-biasing effects of 50-50 and divide-and-conquer (in particular to reduce effects of anchoring and adjustment) but comment that they did not compensate for the restricted information relative to judgmental adjustment. Further, the authors comment that it should be possible to combine judgmental adjustment with Delphi to harness the insights of the group without increasing bias. Onkal et al. (this issue [4]) identify a "truth bias", whereby forecasters place too much faith in the veracity of advice, which leads to over-adjustment. They also find that informal weighting which is suboptimal being subject to biasing influences such as the status of the advisor. Onkal et al. suggest concealing status by means of anonymous advisors and similarly restricting other indicators of the veracity of advice to those that reliably indicate its quality. Finally, Petropoulous et al. (this issue [6]) show that the provision of performance feedback, especially that which highlights bias on the part of forecasters, is effective for debiasing, at least in those situations where feedback is readily available (i.e. short-to-medium-term forecasting).

Stage 4: Evaluation and documentation

Although this is offered here as the final stage, in some cases, particularly for rolling foresight projects and/or short- to medium-term forecasts, evaluation and documentation might be conducted at intermediate intervals in order to fine-tune the process.

a). Evaluation of performance

Performance should be evaluated wherever possible (i.e. in any case other than very long-term horizons) and the results fed-back to relevant personnel (e.g. managers, participating experts) in order to try and improve the quality of prediction in future. The reliability and validity of individual input judgments and uncertainty assessments should also be evaluated and fed-back as appropriate.

There is an issue as to how best to evaluate performance. If judgments are quantitative in nature, similar to each other (i.e. are based on essentially the same kind of information), are made by the same experts several times, and the outcomes are available within a useable time-frame for the organization, then error measures of the judgments such as mean squared error (for judgments of quantities) or Brier score (for probability judgments) are appropriate for both evaluation and feedback, and for use in training. However, as we have already mentioned above in relation to Cooke's method, caution should be exercised in using such performance measures to select or weight experts; at least until further research has demonstrated that such procedures provide practically significant advantages over equal treatment of experts (Bolger & Rowe, 2015 a, b). If not all of these conditions pertain, then the reliability and validity of these performance measures can be questioned, and coherence measures might be better used (e.g. consistency with axioms of logic and probability). For qualitative judgments subjective assessment of validity and coherence is all that is really possible.

In Meissner et al. (this issue [1]) judgments of uncertainty and impact by internal and external stakeholders are compared but not formally (e.g. statistically) evaluated. However, descriptive (graphical) summaries of results for impact and uncertainty are fed-back to experts, discussed, and potentially influence subsequent decision making. In Scenario Planning, as described by Derbyshire and Wright (this issue [2]), strategy can be tested against scenarios generated by experts through simulation and thought experiment. However, given long horizons, and awareness that the process itself influences the outcomes (since actions may be taken by workshop participants to both facilitate the occurrence of favorable futures and to avoid unfavorable ones), more formal evaluation is difficult-to-impossible. Hanea et al. (this issue [3]) evaluate probabilistic forecasts for binary events using the Brier score (which they criticize). Further, within IDEA's Delphi-like structure, judgments and rationales are fed back to judges (but not performance, as this can only be assessed after forecast events occur). In Onkal et al. (this issue [4]) realizations of forecasts are given as feedback to evaluate advice (but no rationales). In Alvarado et al. (this issue [5]) forecast error is computed as average percent error but no outcome or performance feedback is given to experts in the study (although the authors make a recommendation to provide this as training in real applications). Further, Petroplous et al. (this issue [6]) find that such feedback is useful for debiasing, particularly when given after each forecast rather than averaged over several. Finally, Wilson (this issue [7]) measures performance (on seed variables) using error scores like MAPE, but does not discuss providing such measures as feedback (e.g. for training) or the potential effects of doing this on dependencies and aggregation: this is something that may warrant future research attention.

b). Documentation of Process

In addition to reporting on the quality of the forecasts (and inputs) the entire forecasting process should be documented as fully as possible in order to again improve the outcomes both locally (i.e. for the organization commissioning the forecasts). and for the wider community. If forecasts are to be used for policy making then documentation is essential to maintain transparency and create a consultable audit trail. In addition, provision of feedback and documentation to experts is important to reward participation, and retain expert participation in future exercises (or ongoing exercises). These issues are not explicitly addressed in any of the papers in the Special Issue so we will not discuss them beyond noting that documentation is an important part of the EKE process, particularly in practical settings.

Discussion

As a piece of applied research, with many characteristics of a methodology, EKE is to be contrasted with most previous work concerned with the role of judgment in forecasting, which has largely been basic research investigating how judgment compares to statistical forecasting under various conditions – mostly focussing on the "application of method" stage in the forecasting process – with rather limited practical recommendations arising (see, e.g., Lawrence et al., 2006, for a review). Conversely, EKE as we have defined it above, rather than in its narrow interpretation in relation to risk analysis, has much in common with Scenario Planning, in that it is a practical enterprise, concerned with the entire process of eliciting and modelling expert knowledge: for this reason we have regarded Scenario Planning as another kind of EKE in this editorial and Special Issue.

Similarities and differences between EKE, judgmental forecasting research (henceforth JF), and Scenario Planning are shown in Table 1. Note that the analysis in this table is something of an over-simplification, for instance, sometimes experiments are used in EKE, and real experts studied in JF, but we maintain that the characterizations in the table are the most typical. Nevertheless, we note that recent examples to the contrary are now emerging –

see Fildes et al's (2009) field study and the studies by Franses et al (e.g., 2011) on both macro-economic and company forecasts. Despite the differences between the JF approach and the EKE approach there is also much overlap, and the papers in this Special Issue that come from a judgmental forecasting tradition (this issue [4],[5] and [6]) provide valuable insight into particular issues that we have discussed above as being important to EKE such as: the selection, weighting and training of experts; the choice of method on the basis of features of tasks and experts; and the evaluation of performance and its use as feedback.

Table 1: Typical focus of the three approaches

	Expert Knowledge Elicitation	Judgmental Forecasting	Scenario Planning
Participants	Groups of professionals/experts	Individual laypeople/novices	Groups of professionals/experts
Methodology	Case-based	Experimental	Case-based
Focus	Whole process	Specific stage	Whole process
Rationale	Applied: Inform strategy/decision making	Pure: Test hypothesis or answer specific research question	Applied: Inform strategy/decision making
Models	Causal-explicit or Statistical/deterministic	Causal-implicit* or Statistical	Causal-explicit
Usual Time Horizon	Medium-Long	Short-Medium	Long

*e.g. judgmental extrapolation or adjustment without underlying models being made explicit

Perhaps the biggest issue with EKE is assessing its effectiveness. This is partly because the extant studies are case-based and do not obviously generalize. Also, the in-practice use of medium to long time-horizons precludes outcome validation. Further, as Green and Armstrong (2015) argue, simpler forecasting methods tend to be more effective than more complex, and the EKE procedures described in the papers in this Special Issue are relatively complex. Although, as we have stressed above, EKE is usually applied when simpler methods are inadequate, nonetheless further research is needed in order to evaluate whether the additional resources needed for EKE are worth it. Scenario planning is not a forecasting method, however, similar issues regarding effectiveness are inherent in its use.

Since many of the readers of this Special Issue will be familiar with JF but not EKE we think it is worth exploring the similarities and differences between the two approaches further in the form of a short "Q and A" that follows next.

Q1) Expert judgment has been studied in the JF literature for many years: just what does the new EKE conceptualisation add to this?

As we can see from Table 1 there are several differences between the two approaches.

In EKE the expert is crucial whereas in JF it is just another variable that is manipulated in some studies, many JF research studies making use of non-professionals (e.g. students) or novices. This focus on the expert means that characteristics of expertise (e.g. normative versus substantive characteristics) are measured and analyzed in detail in EKE but are often ignored or not treated in detail in JF. Further, JF usually examines individual judgment whereas mostly EKE elicits the judgments of several experts, thus raising the issue of how best to combine the judgments that is rarely addressed in JF research, beyond how to integrate the judgment of one person with statistical output, or combine the judgments of an advisor and advisee.

Another difference between EKE and JF is that the former is concerned with the whole process, whereas the latter generally focuses on specific stages of the process. This means that the focus on experts in EKE feeds into several aspects of the process such as selection, screening, and training, as well as having implications for the choice of method, type of aggregation and weighting.

The different methodologies typically used by the two approaches – case-based versus experimental: the former with explicit applied goals, the latter focused more on theory testing – also has implications for the theoretical and applied contributions of this research. On the plus side for JF the greater internal validity offered by the experimental approach means that the factors that influence the quality of anticipation involving judgment can be isolated and examined in detail, thus permitting improvement in foresight methods on the one hand, and testing of (social) psychological theories, on the other. On the plus side for EKE the case-based approach has greater ecological validity than the experimental one, and generally produces richer data (e.g. in the form of protocols or causal models). However, both approaches have their downsides, with the applicability of JF findings to real-world problems often being in question, while it can be difficult to generalize lessons of EKE research to different task domains. We plead wider use of experimental methods to investigate expert judgment for anticipation, but with true experts performing real, or realistic, tasks. This is may not appear easy, since it is practically difficult to perform experiments with experts (by definition they are busy and scarce) but, in fact, most people are expert at something (or can be trained to be[11]).

Two other differences between EKE and JF intersect with the methodological concerns we discuss in the previous paragraph: the models and the anticipation horizon. EKE tends to elicit or integrate expert knowledge into rich, often causal, models, then use those for long-term anticipation, whereas JF usually does not make explicit the causal knowledge of the judges (or characterizes them in rather simple terms e.g. as heuristics); as is the case for judgmental extrapolation and for judgmental adjustment of statistical output, for short- to medium-term forecasting.

[11] For example student participants could be trained to make diagnoses on the basis of cues within an experimental session thus accelerating a learning process that might otherwise take years.

Q2) You describe developments and advances in EKE that have been made within the field of risk estimation: does this mean that some areas of EKE have been under-studied in the forecasting literature? If so, what should be the new topics of research focus in judgmental forecasting?

This question has been partially answered in the previous section. JF would benefit from more frequently studying real experts performing on tasks that are clearly related to their expertise (and are framed in a familiar way with regard both to presentation of data and mode of response): this would entail more in-depth analysis of both tasks and the capabilities of the experts. Also it would be good to see more work in JF on the characteristics of expert knowledge (e.g. mapping it and measuring its coherence, reliability over time, generalizability etc.). Further, there is a need for more studies of forecasting in groups (nominal or real) and how best to elicit and combine group knowledge and forecasts.

Both from an EKE and a JF perspective, problem detection and the elicitation and construction of expert domain models have received relatively little attention: both could be informed by work in Scenario Planning in these areas. Documentation of the process has also received relatively little attention, particularly from JF.

Q3) You have introduced several new approaches such as expert selection, IDEA, IL, GEM, and Sheffield and Cooke's methods. What are the take-home implications for the practical improvement of judgmental forecasting practice?

The main message is that expert judgment should be considered as data and, as such, the methods used to obtain and use this data (i.e. EKE) should be such that they maximize the reliability and validity of this data (as is true of any empirical method). There are a number of different ways of achieving reliability and validity of expert judgment, which are mostly not mutually exclusive, and thus can be combined:

- measuring the normative and substantive expertise of potential experts through tests (e.g. of answers to seed questions), self-assessment questionnaires (e.g. answers to the expert-skills questionnaire), assessment by others (e.g. answers to GEM), experience indicators (e.g. CV: years in role, qualifications, publications, citations, patents etc.), and social indicators (e.g. job title, media presence etc.);
- using such measures of expertise to select, screen or weight experts;
- using such measures to identify training needs and train accordingly;
- removing noise by careful use of well-researched elicitation protocols that include the use of: proper scoring rules, rich feedback of judgments, and opportunities to reflect on judgments and revise them;
- also using well-researched and administered protocols that avoid effects of framing, availability, representativeness and anchoring to debias individual judgments, and that also provide both anonymity and facilitated information exchange to de-bias group judgments;

- collecting as much data as is possible while balancing costs and benefits of increasing sample size (noting that there may be a trade-off between sample size and degree of participant expertise).

In summary, focus on the above will underpin the development of a “meta-protocol” guiding the whole process of EKE to aid anticipation of the future: from problem recognition through to documentation.

Finally, our discussion and analysis here is not comprehensive: some important aspects of the EKE process have not been addressed (or are not addressed very thoroughly) as they are not discussed explicitly in any of the papers in the Special Issue These include: motivation and incentives; decomposition (and re-composition of judgment); reliability and validity of data/expert judgment; anonymity; transparency of the process; and ethical issues. These are also possible subjects for future research – the initial and final stages (i.e. initiation and documentation) are also areas which, we believe, will particularly benefit from focussed work.

References

Alvarado-Valencia, J., Barrero, L.H., Onkal, D. and Dennerlein, J.T. (this volume [5]). Expertise, credibility of systems forecasts and integration of methods in judgmental demand forecasting. *International Journal of Forecasting.*

Armstrong, J.S (1985). *Long-range forecasting: From crystal ball to computer*, 2nd ed., New York: Wiley.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature*, *463*, 294-295.

Berry, D.C. & Broadbent, D.E. (1984). On the relationship between task performance and associated verbalizable knowledge, *Quarterly Journal of Experimental Psychology, 36A,* 209-231.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3.

Budnitz, R.J., Boore, D.M., Apostolakis, G., Cluff, L.S., Coppersmith, K.J., Cornell, C.A., & Moms, P.A. (1997*). Recommendations for probabilistic seismic hazard analysis: guidance on uncertainty and use of experts*, Vol. 1, Washington, DC: US Nuclear Regulatory Commission.

Burgman, M.A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fidler, F., Rumpff, L. & Twardy, C. (2011). Expert status and performance. *PLoS ONE,* 6, e22998.

Bolger, F and Harvey, N. (1998). Heuristics and biases in judgmental forecasting. In: G. Wright, G. & P. Goodwin (Eds.) *Forecasting with judgment,* New York: Wiley.

Bolger, F., Hanea, A., Mosbach-Schulz, O., Oakley, J., O'Hagan, A., Rowe, G. & Wentholt, M. (2014). *Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment*. Parma, Italy. European Food Safety Authority (EFSA). http://www.efsa.europa.eu/en/efsajournal/pub/3734. Accessed on 22 June 2016.

Bolger, F**.** & Rowe, G. (2015a). The aggregation of expert judgment: Do good things come to those who weight? *Risk Analysis*, *35*(1)**,** 5-11.

Bolger, F. & Rowe, G. (2015b). There is data, and then there is *data*: Only experimental evidence will determine the utility of differential weighting of expert judgment. *Risk Analysis*, *35*(1), 21-26.

Bolger, F. & Wentholt, M. (2014). Principles and practice of selecting and motivating experts. In: Bolger et al. *EFSA Journal 2014: Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment*. Parma, Italy: European Food Safety Authority (EFSA), pp. 138- 162. http://www.efsa.europa.eu/en/efsajournal/pub/3734. Accessed on 22 June 2016.

Bolger, F. & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems*, *11*, 1-24.

Bolger, F**.** & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change*, *78*, 1500-1513.

Bolger, F., Wright, G., Rowe, G., Gammack, J. & Wood, R. (1989). LUST for life: Developing expert systems for life assurance underwriting. In: N. Shadbolt (Ed.), *Research and Development in Expert Systems VI*, Cambridge: Cambridge University Press, pp. 128-139.

Brier, G. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review*, *78*, 1-3.

Clemen, R.T. & Winkler, R.L. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis*, *19*, 187–203.

Cooke, R.M. (1991). *Experts in uncertainty: Opinion and subjective probability in science.* Oxford: Oxford University Press

Cooke, R.M., ElSaadany, S. & Huanga X. (2008). On the performance of social network and likelihood-based expert weighting schemes. *Reliability Engineering and System Safety, 93*, 745–756.

Cooke, R.M. & Goossens, L.H.J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90*, 303–9.

Cooke, R.M and Probst, K.N. (2006). *Highlights of the expert judgment policy symposium and technical workshop*. Washington, D.C.: Resource for the Future.

Day, J. (2013). *Review of cross-government horizon scanning*. London: Cabinet Office.

Dhami, M.K. & Wallsten, T.S. (2005). Interpersonal comparison of subjective probabilities. *Memory & Cognition*, *33*, 1057–1068.

Derbyshire, J. & Wright, G. (this volume [2]). Augmenting the Intuitive Logics scenario planning method for a more comprehensive analysis of causation. *International Journal of Forecasting*.
Eden, C. (1988). Cognitive mapping*. European Journal of Operational Research*, *36*(1), 1-13.

Eggstaff, J.W., Mazzuchi, T.A. & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering and System Safety*, *121*, 72–82.

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, 3-23.

Franses, P.H., Kranendonk, H. C. & Lanser, D. (2011). One model and various experts: evaluating Dutch macroeconomic forecasts. *International Journal of Forecasting*, 27, 482-495.

US EPA. (2009). Expert elicitation task force white paper. Washington, DC: Science and Technology Policy Council. http://www.epa.gov/osa/pdfs/elicitation/ Expert_Elicitation_White_Paper-January_06_2009.pdf. Accessed 22 June 2016.

Germain, M.L., & Tejeda, M J. (2012). A preliminary exploration on the measurement of expertise: an initial development of a psychometric scale. *Human Resource Development Quarterly*, *23*(2), 203–232.

Glaser, M., Langer, T. & Weber, M. (2013) True overconfidence in interval estimates: Evidence based on a new measure of miscalibration. *Journal of Behavioral Decision Making*, *26*, 405–417.

Gordon, T. & Pease, A. (2006). RT Delphi: An efficient, "round-less" almost real time Delphi method. *Technological Forecasting and Social Change*, *73*, 321-333.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, *68*(8), 1678-1685.

Hanea, A.M., McBride, M.F., Burgman, B.C., Wintle, F., Flander, L., Twardy, C.R., Manning, B. & Mascaro, S. (this volume [3]). InvestigateDiscussEstimateAggregate for structured expert judgment, *International Journal of Forecasting.*

Hoerl, A.E. & Fallin, H.K. (1974). Reliability of subjective evaluations in a high incentive situation. *Journal of the Royal Statistical Society. Series A (General)*, *137,* 227-230.

Howard,R.A. & Matheson, J.E. (2005) Influence diagrams, *Decision Analysis*, *3*, 127-143.

Kahneman, D. & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

Karlsson, L., Juslin, P. & Olsson, H. (2007). Adaptive changes between cue abstraction and exemplar memory in a multiple-cue judgment task with continuous cues. *Psychonomic Bulletin & Review, 14 (6),* 1140-1146.

Keren, G. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*, 98-114.

Knol, A.B., Slottje, P., van der Sluijs, J.P. & Lebret, E. (2010). The use of expert elicitation in environmental health impact assessment: a seven step procedure. *Environmental Health*, *9*, 19-35.

Larson, J.R. Jr., Christensen, C., Franz, T.M. & Abbott, A.S. (1998). Diagnosing groups: The pooling, management, and impact of shared and unshared case information in team-based medical decision making. *Journal of Personality and Social Psychology*, *75*, 93-108.

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

Lichtenstein, S., Fischhoff, B. & Phillips L. (1982) Calibration of probabilities: The state of the art to 1980. In: D. Kahneman, P. Slovic & A. Tversky (Eds.) *Judgment under uncertainty: Heuristics and biases.* Cambridge, England: Cambridge University Press, pp. 306-334.

Lin, S-W. & Bier, V.M. (2008) A study of expert overconfidence. *Reliability Engineering and System Safety*, *93*, 711–721.

Linstone, H.A. & Turoff, M. (1975) *The Delphi Method: techniques and applications,* London: Addison-Wesley.

Meissner, P., Brands, C. & Wulf, T. (this volume [1]) Quantifying blind spots and weak signals in executive judgment: a structured integration of expert judgment into the scenario development process, *International Journal of Forecasting.*

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D. Atanasov, P., Swift, S.A.; Murray, T., Stone, E. & Tetlock, P.E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science,* 25, 1106–1115.

Mellers, B.A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N.; Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L. & Tetlock, P.E. (2015). Identifying and cultivating "superforecasters" as a method of improving probabilistic predictions. *Perspectives in Psychological Science, 10*, 267–281.

Meyer, M.A. & Booker J.M. (2001). *Eliciting and analyzing expert judgment: A practical guide*. Alexandria, Virginia, USA: American Statistical Association and the Society for Industrial and Applied Mathematics.

Morgan, M.G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*, 7176-7184.

Murphy, A. H., & Brown, B. G. (1985). A comparative evaluation of objective and subjective weather forecasts in the United States. In G. Wright (Ed.) *Behavioral decision making.* New York: Plenum, pp. 329-359.

Nisbett, R.E. & Wilson. T.D. (1977). Telling more than we can known: Verbal reports on mental processes, *Psychological Review, 84*, 231-259.

Oakley, J.E. & O'Hagan, A. (2010). SHELF: the Sheffield Elicitation Framework. Version 2.0. Sheffield (UK): University of Sheffield, School of Mathematics and Statistics. http://tonyohagan.co.uk/shelf. Accessed 22 June 2016.

O'Hagan, A,, Buck, C.E., Daneshkhah, A., Eiser,, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. & Rakow, T. (2006). Uncertain judgements: Eliciting experts' probabilities. Chichester, U.K.: Wiley.

Oliver, R.M. & Smith, J.Q. (1990). *Influence diagrams, belief nets and decision analysis*. New York: Wiley.

Onkal, D., Gonul, M.S., Goodwin, P., Thomson, M. & Oz, E. (this volume [4]). Evaluating expert advice in forecasting: Users' reactions to presumed versus experienced credibility. *International Journal of Forecasting.*

Petropoulos, P., Goodwin, P & Fildes, R. (this volume [6]). Using a rolling training approach to improve judgmental extrapolations elicited from forecasters with technical knowledge. *International Journal of Forecasting.*

Reilly, J., Stone P.H., Forest, C.E., Webster, M.D., Jacoby, H.D., et al. (2001). Uncertainty and climate change assessments. *Science*, *293*, 430-433.

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

Schoemaker, P.J. & Day, G. S. (2009). How to make sense of weak signals. *MIT Sloan Management Review, 50*, 81–89.

Schoemaker, P.J., Day, G.S. & Snyder, S.A. (2013). Integrating organizational networks, weak signals, strategic radars and scenario planning. *Technological Forecasting and Social Change, 80*, 815–824.

Shanks, D. R. (1997). Representation of categories and concepts in memory. In M.A. Conway (Ed.), *Cognitive Models of Memory*, (pp. 111-146). Cambridge, MA: MIT Press.

Stasser, G. & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48, 1467-1578.

Stasser, G. & Titus, W. (2006). Pooling of unshared information in group decision making: Biased information sampling during discussion. In J.M. Levine & R.L. Moreland (Eds.), *Small Groups* (pp.227-239). New York: Psychology Press.

Teigen K.H. & Jorgensen, M. (2005). When 90% confidence intervals are only 50% certain: On credibility of credible intervals. *Applied Cognitive Psychology*, *19*, 455–475.

Tetlock, P.E. & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Wallsten, T.S. & Budescu, D.V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review*, *10*, 43–62.

Wilson, K.J. (this volume [7]). An investigation of dependence in expert judgment studies with multiple experts. *International Journal of Forecasting.*