International Journal of Forecasting 30 (2014) 313–327

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

The Delphi method in forecasting financial markets—An experimental study

Karlo Kauko[a,*], Peter Palmroos[b,1]

[a] Bank of Finland, PO BOX 160, 00101 Helsinki, Finland

[b] Financial Supervisory Authority, PO BOX 103, 00101 Helsinki, Finland

ARTICLE INFO

Keywords:
Delphi method
Financial markets
Expert judgment
Belief perseverance
Overconfidence
Conditional forecasting

ABSTRACT

Experts were used as Delphi panellists and asked to present forecasts on financial market variables in a controlled experiment. We found that the respondents with the least accurate or least conventional views were particularly likely to modify their answers. Most of these modifications were in the right direction but too small, probably because of belief-perseverance bias. This paper also presents two post-survey adjustment methods for Delphi method based forecasts. First, we present a potential method to correct for the belief perseverance bias. The results seem promising. Secondly, we test a conditional forecasting process, which unexpectedly proves unsuccessful.

1. Introduction

The Delphi method was introduced at the RAND Corporation in the 1950s. It aims to maintain the advantages of an interacting group without potentially counterproductive group dynamics, such as dominant individuals who may not be the best experts.

In short, the traditional version of the method is based on a multi-round survey. Respondents are asked to answer a number of questions in writing. Answering is anonymous; other respondents do not know who answered what. In most cases the answers are numeric estimates, ratings on a scale, or yes/no. Often, the respondents also have the opportunity to write comments on the issues raised in the questionnaire. Statistics on answers and the related comments are subsequently distributed to the respondents, but this information is anonymous and no respondent can identify who answered what. Each respondent is allowed to modify his own answers, and possibly to add more comments.

After a few rounds, some convergence in answers is normally observed due to a group opinion building process, leading to less variance in the answers and more agreement within the panel. The number of rounds can be either predetermined or dependent on the criteria of consensus and stability. In the papers reviewed by Rowe and Wright (1999), the number of respondents in Delphi panels varied from 3 to 98. Ideally, the respondents will all be experts in the same field, but with somewhat different backgrounds.

Linstone and Turoff (2011) emphasised the role of communication in judgemental forecasting, and argued that the Internet will have a major impact on the way in which comparable methods will be used in the future, since the number of potential participants will be much larger than in traditional Delphi panels. In recent years, some studies have used the real-time version of the method (see Gordon & Pease, 2006), which is normally an online application that allows respondents to modify their answers at any time, up until the end of the answering process (for the validation of the method, see Gnatzy, Warth, von der Gracht, & Darkow, 2011).

The final answer of the group is defined as the mean or median of the individual answers. In many cases, even the questions to be answered are proposed and selected by group members themselves before the first answering

* Corresponding author. Tel.: +358 10 8312519; fax: +358 10 8312294.
E-mail addresses: karlo.kauko@bof.fi (K. Kauko), peter.palmroos@bof.fi (P. Palmroos).
1 Tel.: +358 10 8315356.

0169-2070/$ – see front matter © 2013 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijforecast.2013.09.007

round. The forecasting accuracy of the group normally improves over Delphi rounds, and the Delphi method works better than staticized groups, i.e., simple one-round surveys; this finding has been reported by Dalkey (1968), Graefe and Armstrong (2011), Helmer (1964), Parenté et al. (2005) and Song, Gao, and Lin (2013), as well as in various studies reviewed by Rowe and Wright (1999). In some experiments, respondent groups have been asked to provide estimates on almanac events, i.e. issues related to the past and present (see for example Graefe & Armstrong, 2011); whereas in other cases they have been asked to forecast the future (Parenté, Anderson, Myers, & O'Brien, 1984; Parenté et al., 2005). More detailed descriptions are provided by Linstone and Turoff (1975), Parenté and Anderson-Parenté (1987) and Rowe, Wright, and Bolger (1991).

In light of most of the earlier experimental research, the Delphi method seems to be either substantially (Basu & Schroeder, 1977; Riggs, 1983) or somewhat (Graefe & Armstrong, 2011) better than Face-to-Face (FTF) meetings, although some authors have found the differences to be negligible (Brockhoff, Kaerger, & Rehder, 1975). Findings by previous authors have been summarised by Rowe and Wright (1999, Table 4) and Woudenberg (1991, Table 3); according to both surveys, most previous contributions had found that, with some exceptions, the Delphi method had outperformed traditional meetings.

So is it reasonable to use sophisticated, structured processes if there is no unambiguous evidence that they yield significantly better forecasts than a simple FTF meeting? As will be seen in this paper, structured techniques have at least one clear advantage: they can be improved and the forecasts honed.

In our study, the FTF meeting is used as the benchmark case. This is the simplest and probably the most commonly used method; group members sit in the same room and discuss the issues until they reach a consensus, or at least until the majority backs a view. According to Kerr and Tindale (2011), such meetings are good for pooling information, mutual error checking and motivation enhancement. On the other hand, they may be particularly vulnerable to the 'tyranny of the majority', the dominance of powerful individuals, inattention to unshared information, or group overconfidence. Other potential problems of FTF meetings include the bandwagon effect (the tendency of ideas to spread among people like fads), the underdog effect (the tendency of some people to vote for losing candidates or views), and the halo effect (the tendency to weight an opinion according to a general impression of the person who expresses it).

According to Ang and O'Connor (1991, p. 142), the Delphi method combines mathematical and behavioural approaches, with an 'aim to improve behavioural aggregation by substituting the dysfunctional aspects of achieving consensus with a mathematical process of achieving the final group judgement'. In the best case, the method helps to eliminate a number of problems with FTF meetings, such as the influence of dominant individuals and the unwillingness of many people to defend unorthodox views, even well-founded ones. Different biases in the Delphi method have also been studied. For example, Ecken, Gnatzy, and von der Gracht (2011) discuss the desirability bias: the general tendency of respondents to over-estimate the probability of events that they consider to be desirable. Unfortunately, no direct test of the ability of the proposed correction to improve the forecast accuracy was presented, but evidence was given on the existence of the bias.

This paper has two main objectives. First, we address the development of individual answers during the process. Secondly, we try to develop the Delphi method further, using observations on panellists' behaviours and findings from existing research in psychology.

The use of post survey methods to increase the accuracy of forecasts is not a new idea. Armstrong (2006) listed and evaluated evidence on numerous post hoc methods. However, our methods differ from those presented by Armstrong. Instead of assigning different weights to the panellists depending on, for example, previous forecasting performances, we correct the assumed undersized adjustments which the panellists had made between the first and last rounds.

This paper is organised as follows. In the next section we provide a number of hypotheses that are subsequently tested empirically. Section 3 details the experiment. Section 4 focuses on forecasting accuracy, and Section 5 examines the dynamics of individual forecasts in the Delphi context. Section 6 examines post-forecast correction methods, and the last section concludes.

2. Development of hypotheses

Instead of comparing the Delphi and FTF meeting methods, the main aim of this paper is to study whether the post survey methods we used can reduce the forecast errors. This is done by testing a number of hypotheses with data from a controlled laboratory experiment. We present three hypotheses on the modification of individual answers.

Hypothesis 1. Individual answers improve more than average if they originally differed from the group average.

To express the same idea in statistical terminology, there is a negative correlation between changes in the absolute value of a respondent's forecast error $|\psi_{j,p,\text{final}}| - |\psi_{j,p,1}|$ and the original deviation from the group mean $|X_{j,p,1} - \overline{X}_{j,1}|$, where $\psi_{j,p,n}$ is a measure of respondent p's answering inaccuracy in round n, question j, $\overline{X}_{j,1}$ is the first round group answer (mean of individual answers) to question j, and $X_{j,p,1}$ is respondent p's first round answer to question j.

Hypothesis 2. Individual answers improve more than average if they were originally more inaccurate than average.

Using statistical terminology, there is a negative correlation between changes in the absolute value of the error $|\psi_{j,p,\text{final}}| - |\psi_{j,p,1}|$ and the size of the original error $|X_{j,p,1} - Y_j|$, Y_j = correct answer (observed later).

Hypothesis 3. Being an outlier and being inaccurate have an interaction effect on improvements in answers—simultaneously being an outlier and inaccurate predicts an improvement in accuracy that exceeds the sum of the two separate effects.