Exploring the Relationship Between Topic Area Knowledge and Forecasting Performance

Sarah M. Miller, Clifton Forlines, John Regan
C.S. Draper Laboratory
Cambridge, MA

The Intelligence Community (IC) is often asked to make predictions about future world events. One aspect of predicting the quality of forecasts and forecasters is the knowledge that the forecaster has about the question to be forecast. This paper explores the relationship between factual knowledge about a forecast event and eventual performance on a forecast question. The results demonstrated a significant relationship between a forecaster answering a series of factual questions correctly and answering the corresponding forecast question correctly. This relationship is enhanced when controlling for the relative difficulty of the factual question. When controlling for forecaster performance, roughly half of the impact was due to general forecaster performance and half was due to their specific knowledge about a given forecast question. Interestingly, we found that forecasters with more factual knowledge were less calibrated with respect to their probability forecast whereas forecasters who were less knowledgeable were better calibrated in their probability estimates. We discuss the implication of the results related to improving forecast quality.

INTRODUCTION

The Intelligence Community (IC) is often asked to make forecasts, or predictions, about the occurrence of significant world events based on limited quantitative data. For example, an analyst, or group of analysts might be asked to provide an assessment of the probability that a country will withdraw from the European Union or whether Syria will descend into civil war. The accuracy of analysts' forecasts can have large and far-reaching consequences for policy makers, the military, and other government agencies; therefore, even a small improvement in forecast accuracy can have a major impact.

The experiment described in this paper is part of a larger effort to understand and improve forecast accuracy using the collective wisdom of forecasters. Specifically, the program's aim is "to dramatically enhance the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts" (IARPA, 2010).

One of the many challenging aspects of this project is to try to estimate the quality of an individual's forecast on a single, independent, forecast question before the question resolves (i.e. if we don't have any historical responses from an individual forecaster on the current or any prior forecast question). An additional challenge of this research program is that the forecast events cover a wide range of topics (see Table 1 for examples), so it is highly unlikely that an individual will be universally knowledgeable about such a heterogeneous collection of topics. Therefore, if we can begin to understand and assess how much knowledge a forecaster has about a given forecast question, we might be able to estimate how well they might do on that forecast question.

In this paper, we describe a method for 1) identifying individual forecasters with subject matter expertise related to the given forecast question and 2) testing whether individuals with this identified expertise make more accurate forecasts than those that we identified to have less expertise. In other words, we want to determine if we can identify a forecaster with more factual knowledge, or expertise, about a given forecast question and determine if those forecasters actually make better forecasts. If there is a link between the identified factual knowledge and eventual forecast performance, we can use this relationship to weight forecasts from individuals that we identify to have expertise in the forecast question and downgrade the forecasts from individuals without that factual knowledge.

There are two distinct challenges in this project. First, because these forecast questions are complex and multidimensional, it is difficult to determine what factual knowledge might be both relevant and needed to answer any given forecast question. For example, one forecast question is "Will the daily Europe Brent Crude FOB spot price per barrel be greater than or equal to $150 before 3 April 2012? " On the surface, this is a question about markets and oil prices. However, to answer this question, the factual information that might be relevant included (at the time) Iran's threat to close the Strait of Hormuz, changes in demand due to the economic uncertainty in Europe, and a variety of other economic and political factors.

Second, once we determine what information is relevant to the forecast problem, it is unclear whether

DOI 10.1177/1071181312561074

having this information actually helps improve forecast quality. Although it might seem obvious that forecasters who are more knowledgeable about the forecast question will make better forecasts, the results of the research are not clear-cut. Across a variety of domains, expertise, beyond a minimum baseline, does not improve forecast accuracy (Armstrong, 1980). Moreover, in a multi-decade study involving expert forecasters within the IC predicting geo-political events very similar to the questions posed in the current study, experts in their fields performed poorly in the forecasting task (Tetlock, 2006). In fact, these experienced forecasters were not much better than a simple status quo heuristic (i.e. the forecast for the future is the same as what has happened in the past).

Because the relationship between factual knowledge and forecast performance in a geo-political domain is not well understood for forecasters with different levels of skills, we explore this relationship in this paper.

EXPERIMENT

The primary goal of this study was determine if there is a relationship between forecaster's performance on fact-based questions related to a forecast problem and their performance on that forecast problem. To examine this potential relationship, we focus on three specific research questions:

Research Question 1: Is there a relationship between the number of factual questions a participant got correct and the probability that their answer to the forecast question was correct?

Research Question 2: When we adjust for the difficulty of the factual questions, do we gain additional strength in the relationship between forecast performance and knowledge?

Research Question 3: Is there a relationship between the participant's Brier Score on the factual questions and the Brier Score for the forecast questions?

Methods

The experiment and data analysis described in this paper is part of a larger on-line study that began on September 1, 2011 and this portion of the program (Year 1) was completed on April 30, 2012.

Participants: Participants were required to be United States citizens over 18 years of age. Although anyone meeting these criteria could enroll in the study, participants were targeted using primarily online advertising for their interest in world events, politics, and the military. Participants were recruited from a variety of online sources, including political interest blogs, newsgroups and other advertising. Participants were able to join or leave the study at any time throughout the life of the program.

Over 1500 people are currently enrolled in the study and 1060 participants answered at least 1 factual question/forecast question combination Participants were asked approximately 100 forecast questions plus additional demographic, knowledge, and other questions per year. Participants were asked to answer all forecast questions, but were not required to do so. Each participant was paid $575 per year for his or her time.

Forecast Questions: Table 1 lists some example, forecast problems from the study. Forecast problems described in this paper are all True/False questions, and were all resolvable by a predetermined end date that varied from 5 days to 6 months after their initial release. Forecast questions were provided by the sponsor agency and primarily involved geopolitical questions of interest to the IC. For more information on the program seew the website ispade.net.

Table 1. Representative Forecast Problems.

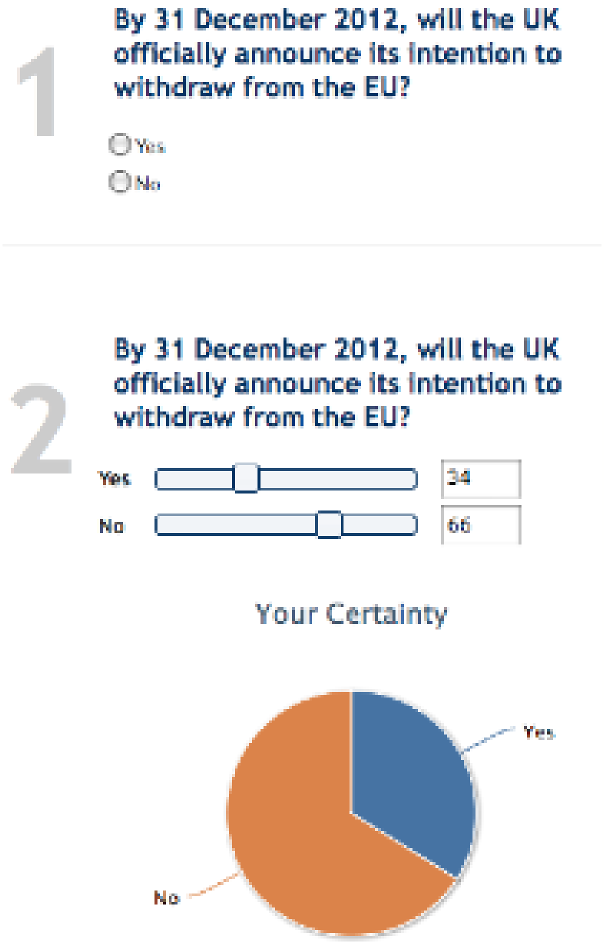
Before April 1, 2012, will Al-Saadi Gaddafi be extradited to Libya?
Will Bashar al-Assad remain President of Syria through January 31st, 2012 (EST)?
Will Mubarak's trial by the SCAF conclude before Feb 1st, 2012 (EST)?
Will a trial for Saif al-Islam Gaddafi begin in any venue by March 31st, 2012?
Will the United Kingdom's Tehran embassy officially reopen by February 29th, 2012?
Will Laurent Lamothe be ratified as Haiti's Prime Minister before April 1, 2012?
Before April 1, 2012, will the Sudan and South Sudan governments officially announce an agreement on oil transit fees?

As part of this study, there were 49 forecast questions with associated factual knowledge questions. We developed three true/false factual questions for each forecast question. Factual questions are described later in this section.

For each forecast question, participants were asked to answer Yes/No if they thought the event would or would not occur by the given date. In addition, participants were asked to provide their probability estimate that the event would, or would not, occur (see

Figure 1). Additional questions were also asked, but they do not pertain to this analysis.

Figure 1. Forecast question elicitation.



1 By 31 December 2012, will the UK officially announce its intention to withdraw from the EU?

○ Yes
○ No

2 By 31 December 2012, will the UK officially announce its intention to withdraw from the EU?

Yes 34
No 66

Factual Questions: For each forecast question, participants were asked three factual True/False questions designed to test a participant's knowledge of the topic area. These factual questions were developed by the research team and were based on research into the forecasting question. Typically, these factual questions were derived from recent news reports related to the forecast question.

An example of a forecast question and its three associated factual questions are included below:

Forecast Question: Before April 1, 2012, will Al-Saadi Gaddafi be extradited to Libya?

Factual Questions:

1. Gaddafi is currently in prison in Algeria.
2. Libya has requested that Gaddafi be extradited back to Libya.
3. Gaddafi is wanted by the International Criminal Court.

The factual questions were each presented as a statement, and the participants were instructed to give their opinion of each statement's truthfulness on an 11-point scale. This scale ranged from "Completely certain the statement is False" through the midpoint "Can't distinguish between True/False" to "Completely certain the statement if True". Through using this scale, we elicited not only their assessment of the truthfulness of the statement, but also their confidence in their answer. Participants were asked to answer all questions without consulting external sources.

Procedure: When a participant logs into our system, they are asked to perform a variety of activities, including answering any open forecasting problems (examples are included later in this section), updating their answers to forecast problems they have already answered, responding to factual questions, or answering other surveys (e.g. demographics, numeracy, personality). They are also able to see their performance on closed forecasting problems.

Participants are asked to complete all open forecast problems, but we do not require completion of all forecast problems to continue to be enrolled in the study. There are typically between 10 – 30 open forecasting problems at any time. The length that the forecast problems are open differs between days and years. Over the last year, approximately 100 forecast problems have been resolved (49 were part of this study because they had factual questions associated with them.

Measuring performance: Performance on both the factual questions and forecast problems were measured in two ways. The first measure was if the forecaster got the question correct or incorrect. The second performance measure used the participant's confidence and probability estimates for the factual and forecast questions respectively. Performance was measured using the Brier Score, a proper scoring function (Brier, 1950). The Brier Score is the mean squared error of the probability forecast over the outcome according to Equation 1.

$$BrierScore = \frac{1}{n}\sum_{1}^{n}(p_i - o_i)^2 \qquad \textbf{Equation 1}$$

In Equation 1, n is the number of observations, p is the probability (or confidence estimate) provided by the participant, and o is the outcome. The outcome is measured as a 0 if the event didn't occur and 1 if it did occur.

RESULTS

The results will be described based on the three research questions outlined earlier in this paper.

Research Question 1: Is an individual's performance on fact-based questions (i.e. did they get the factual questions correct) predictive of eventual forecast performance (i.e. did they get the forecast question correct)?

To evaluate this first research question we compared (via linear regression) the percent factual questions each participant got correct for a given forecast question (0 for 0% correct to 1 for 100% correct) with whether or not the forecast question was correct (0 for incorrect, 1 for correct).

All standard errors are two-way clustered at the participant-by-forecast question level. This accounts for both the fact that participants are the same across forecast questions and forecast questions are the same across participants.

The results are summarized in Table 2. The results showed that a participant who got all of the factual questions correct were 9.8% more likely to get the forecast question correct than a forecaster who did not get any factual questions correct.

Table 2. Results for research question 1. The results are the correlation between the count of factual questions correct and if the forecast problem was correct.

Percent point increase in forecast performance based on the number of forecast questions correct	0.098***
Standard Error	0.029
P-value	0.001

Research Question 2: What factors are driving the relationship between factual question and forecast question correctness?

The first research question evaluated the relationship between performance on the factual question and forecast performance given no additional information about either the factual question or the forecaster. However, because many forecasters answered each factual/forecast question combination and forecasters answered many forecast questions, we have additional information about the relative difficulty of the factual questions as well as how forecasters are performing across factual questions. Therefore, in this analysis we investigate what factors are contributing to the relationship between factual question and forecast performance. The two factors are as follows:

- Factor 1: Controlling for the relative difficulty of factual questions.
- Factor 2: Controlling for individual forecaster factual knowledge.

Factor 1: Controlling for factual question difficulty. Because the research team developed our own factual questions for each forecast question, there was no obvious way to be completely systematic in our development of these questions. Thus, some of the factual questions were likely to be more difficult than others and may provide better discrimination between knowledgeable and unknowledgeable forecasters. Therefore, we wanted to perform an analysis to evaluate if and how the relative difficulty of a factual question affected its relationship to forecast performance.

If the forecast questions have different levels of difficulty, this will bias our estimate of the effect of factual question correctness for a given forecast question. Recall that simply comparing correctness of the knowledge with the forecast question (Research Question 1) yielded a 9.8 percent point improvement. When controlling for forecast question difficulty, the regression coefficient increases to 15.2 percent points (see Table 3; Factor 1). This means that, for a given forecast question, if a forecaster answers 1 percent point more factual questions correct, they are 0.152 percent points more likely to get the forecast question correct. This is almost double the increase without taking question difficulty into account. This suggests there was considerable heterogeneity in the difficulty of the factual questions across forecast questions.

Factor 2: Controlling for differences in factual knowledge Forecaster Performance. Individual forecasters may have different qualities with respect to their performance on factual and forecast questions. An important question is whether the improved forecast performance from those who get more factual questions correct is due to either a) individual performance (fixed across forecast questions) versus b) forecast-specific knowledge held by the individual. To address this question, we add forecaster fixed effect to the regression (so that we now have both forecast and forecaster fixed effects).

The results of this analysis (Table 3; Factor 2) demonstrate that the coefficient on the factual question drops to 0.073. This suggests that roughly half of the impact of correct factual question responses on forecast question performance ((0.15-0.07)/0.15) was due to the inherent individual difference in performance across ALL forecast questions. The remaining half of the impact of knowledge on forecast performance is driven by a respondent's specific knowledge about that particular forecast question. In other words, by asking factual questions, we can determine, in part, who may have ***specific*** knowledge useful for that particular forecast question. The other part is due to ***general*** information across forecast questions.

Table 3. Results for research question 2. The results are the correlation between the count of factual questions correct and if the forecast problem was correct.

	Factor 1: Factual question difficulty	Factor 2: Factual question difficulty AND forecaster performance
Percent point increase in forecast performance	0.152 ***	0.073 ***
Standard Error	0.02	0.016
P-value	0.001	0.001

Research Question 3: Is there a relationship between the participant's Brier Score on the factual questions and the Brier Score for the forecast questions?

Finally, we compared the participant's Brier Score for the factual questions with their Brier Score for the forecast question. Table 4 presents the results of this analysis. Although there was a significant relationship between the number of knowledge questions a forecaster got correct and getting the forecast question correct, there was no significant relationship between Brier Scores (p=0.835). This means that having the additional precision of a forecaster providing probability or confidence estimates provides no predictive power with respect to forecasting.

Table 4. Results for research question 3. The results are the correlation between the Brier Score for the factual questions and if the Brier Score for the forecast question.

Percent point increase in forecast performance based on the number of forecast questions correct	0.003 (ns)
Standard Error	0.016
P-value	0.835

DISCUSSION

This paper evaluated the relationship between participant knowledge on factual questions and eventual performance on forecast questions. The three-part analysis yielded the following results:

- If we know that a forecaster got a factual question correct, we know that they were more likely to get the forecast question correct.
- Controlling for the difficulty of the factual questions almost doubles the
- Almost half of the impact of the relationship between factual and forecast questions were due to differences in knowledge between forecast questions (i.e. specific knowledge). The other half of the impact was due to inherent differences in knowledge for an individual forecaster across forecast questions (i.e. general knowledge).

One of the more surprising findings was the non-significant relationship between Brier Score on the factual and forecast questions, but a significant relationship between the correctness of forecast and factual questions. This is surprising because Brier Score is essentially a more precise measure incorporating confidence and correctness. This finding demonstrates that those who perform better on the factual questions (those with more knowledge) are overconfident on the forecast questions. Perhaps most interestingly, only those with more factual knowledge are overconfident. Those without knowledge are apparently well calibrated to their lack of knowledge when making their forecast.

In the future, we will continue to explore the relationship between confidence in factual questions and performance on the forecast questions to try and determine why there was a non-linear relationship.

ACKNOWLEDGMENTS

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20058. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

The authors also wish to thank Robert Bruzzi, Nathan Hendren, and John Irvine.

REFERENCES

Armstrong, J. S. (1980). The seer-sucker theory: the value of experts in forecasting. *Technology Review.* 82(7):b16-24.

The Intelligence Advanced Research Projects Activity (IARPA) (2010). www.iarpa.gov/solicitations_ace.html

Miller, S., Kirlik, A., Hendren, N. (2011). Applying Knowledge and Confidence Information to Predict Achievement in Forecasting. Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Tetlock (2005). *Expert political judgment: How good is it? How can we know?* Princeton University Press.