Temporal Perceptions and Heuristic Adjustments In Short-term Forecasts

John M. Irvine[1], John Regan[1]

[1] Draper Laboratory, 555 Technology Square, Cambridge, MA, 02139

Abstract

Decision makers in many fields rely on the predictions, including forecasts from subject matter experts. New research is exploring methods for eliciting and combining judgments from multiple experts to arrive at a better overall decision. This study explores the temporal nature of expert forecasts and proposes methods for improving forecast accuracy through temporal adjustments. Consider the forecasting problem of the form: "Will event X occur before the date T?" As the date T approaches, rational forecasters should adjust their predictions, but evidence indicates that most people do not accurately account for these temporal changes. We present a heuristic method called automated updating to adjust individual forecasts. Two related versions of the method are presented. Comparing the performance of the proposed methods to the standard unweighted linear average from the pool of subjects demonstrates the benefits of this approach. Once the outcome of the forecasting problem is known, the Brier score provides an objective measure of performance. Based on the Brier score, these methods outperform the unweighted linear average across a number of forecasting problems.

Key Words: Forecasting, predictions, expert judgments, aggregation methods

1. Introduction

Predictions made through expert judgment are critical to decision making in many fields. Policy makers rely on expert judgment forecasts when formulating strategies for addressing political, economic, and social issues. When multiple experts provide forecast, the merging or aggregation of these judgments presents an interesting challenge. Recent research has shown that combining judgments through averaging leads to poor prediction performance. Experience has shown that individual participants are often reluctant to update their predictions. If new information becomes available, however, updating the prediction to account for the new development should lead to better forecasts. In the absence of new developments, one might envision that, *ceteris paribus*, the probability of the event will decline as the deadline approaches.

In this paper, we discuss a new aggregation process that adjusts the probabilities over time. To illustrate the idea behind out probability adjustment, consider the following hypothetical example. Suppose you have signed up for a four-hour whale watch cruise and suppose the cruise provider says there is a 90 percent chance that you will see a whale. Now suppose that three hours into the four hour cruise, you have not seen a single whale. Do you judge the probability of seeing a whale in the remaining hour to be 90 percent?

This paper presents new methods for making time-based adjustments to the individual forecasts prior to aggregation. We perform a fully automated adjustment and assess the effect on forecast accuracy. An objective measure of prediction performance compiled from a set of forecasting problems quantifies the benefits of this approach. The forecasting problems span a range of topics including politics, economics, and international affairs. The standard for comparison is the simple average of the individual forecasts, also known as the unweighted linear opinion pool (ULinOP). Overall, the proposed method exhibits significantly better performance than the ULinOP.

Researchers have long understood that aggregate estimations built from the individual opinions of a large group of people often outperform the estimations of individual experts (Surowiecki 2004). The use of the Un-weighted Linear Opinion Pool (ULinOP, or group mean) has proven to be a robust method of aggregating forecasts that often outperforms more complex techniques. Draper Laboratory is participating in the Aggregative Contingent Estimation (ACE) Program sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The goal of the ACE Program is to improve the accuracy of forecasts for a broad range of significant event types through the development of advanced techniques that elicit, weight, and combine the judgments of many subject matter experts. Essentially, our aim is to become more accurate in forecasting events of national interest by aggregating predictions from a large number of analysts and experts.

Our research team is tackling two major research challenges under the IARPA ACE Program: How do we best capture the knowledge and understanding that each forecaster has? And, how do we combine this information to produce the best overall forecasts? To answer the first question requires understanding of human perception and sources of bias. Techniques based on cognitive science give the participants multiple ways to view the forecasting problem and convey their estimates. We are conducting a series of experiments to determine which methods are most effective (Miller, Kirlik, and Hendren 2011; Tsai, Miller, and Kirlik 2011; Poore et al. 2011). To solve the second problem of combining the individual forecasts, we are exploring several avenues of research. For example, it would be useful to know who among the forecasters has the real expertise. When collecting forecasts from participants, additional information is elicited that informs the aggregation process and provides indications of individual expertise (Forlines, et al, 2012; Prelec, Seung, and McCoy, 2012).

In this paper, we detail the design of a new aggregation algorithm that meets the goals of

- Being easy to explain to decision makers who have to act upon the aggregate forecast of the group,
- Easy to implement and run on a large collection of forecasts
- Does not require significant effort on the part of the individual forecasters in terms of what information has to be entered.

2. Measuring Forecasting Performance

The measure the accuracy of a probability forecast can be quantified by the Brier score, computed as the average squared deviation between predicted probabilities for a set of events and the (eventual) outcomes (Brier 1950):

$$B = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{r}(f_{ti} - o_{ti})^2$$

where:

f_{ti} is the forecast probability
o_{ti} is the binary indicator of the event outcome
r is the number of possible outcomes
t is the number of forecast instances

The range of the Brier score is [0,2] where 0 indicates a 100% accurate prediction and 2 indicates a completely inaccurate prediction. Applying the Brier scoring rule requires knowledge of the actual resolution for the forecasting problem. Consequently, Brier scoring can only be performed after the forecasting problem has closed and truth is known. To assess performance on a set of forecasting problems, we compute the Brier scores for each individual forecasting problem (IFP) for two competing aggregation methods: the unweighted linear opinion pool (UlinOP) and our new automated updating procedure method.

3. SP♠DE System Overview

To address the ACE Program goals, we have developed the System for Prediction, Aggregation, Display, and Elicitation (SP♠DE), which elicits individual forecasts and related information from a pool of over 1000 participants and generates daily forecasts about a wide variety of world events. The forecasting data collected under the first year of the program forms the basis for the analysis presented here. We performed retrospective analysis on this collection of forecasts with the aim of developing aggregation approaches for use in the next year of the program.

The elicitation methods used in SP♠DE acquire a rich set of information to characterize and model the forecasters and the individual forecast problems (IFPs). A series of experiments have explored the distribution of knowledge among the forecasters, the relationship between knowledge and forecasting accuracy, and the irreducible uncertainty associated with each IFP (Tsai, Miller, and Kirlik 2011; Poore et al. 2014; Miller, Forlines, and Regan 2012). The automated updating procedure relies only on the individual forecasts provided by each participant. An active area of investigation is how to improve performance by incorporating ancillary information into the aggregation process.

Using a web-based interface, the SPADE System elicits forecast and related information from approximately 900 – 1,000 active participants. For each individual forecasting problem (IFP), participants provide judgmental forecasts:

- Will the event occur?
- Probability of the event occurring
- Meta-forecast: What will others predict?
- How would the forecasts improve with access to the knowledge of all participants?

Participants are able to update forecasts, as desired. If news reports indicate a change in conditions related to the forecasting problem, it may be wise to adjust one's predictions based on the emerging story. However, very few participants actually provide updates.

Identifying and recruiting participants with relevant subject matter expertise was a challenge. The participants in this study were recruited through targeted advertisements on numerous, topically relevant announcement boards and academia websites. The team identified and reached out to subject matter experts associated with topical blogs, think tanks, news outlets, and academic institutions. To maximize the effectiveness of these interactions, we employed a three-tiered approach seeking to

1. Stimulate the prospective participant's interest and address their questions about joining the study,
2. Encourage the individual to pass recruitment literature to their colleagues with relevant backgrounds,
3. Invite the individual to share his or her insight about novel venues or mediums which could be used to connect with potential recruits.

Utilizing this three-tiered approach proved successful in achieving the recruiting needed to support the study. All participants are U.S. citizens. The gender balance was approximately two-thirds male. The mean age is 36.5 years and the standard deviation is 13.2. About 88% of participants are college graduates and more than half have advanced degrees.

Table.1: Gender distribution of participants

	Count	Percent
Female	632	37%
Male	1059	63%
Total	1691	100%

To gain a deeper understanding of each forecaster's expertise, we ask participants a variety of addition questions. One question, which we call the meta-forecast, elicits the participant's best estimate of what others in the study are likely to predict. Another question considers their perceptions about the distribution of knowledge among forecasters. In particular, we ask participants how their prediction would change if they had access to all of the knowledge available among the pool of participants. The participants that indicate their forecasts would be unchanged by this additional information are implying that they already have the knowledge and expertise needed to make a good forecast.

Participants were free to return to the SP♠DE UI and update their individual forecasts anytime before the resolution of a forecasting question was known. The frequency with which a participant updated their forecast did seem to have a relationship with the accuracy of these forecasts (Table 2). We hypothesize that this relationship is due to a combination of factors. Firstly, forecasts made toward the end of an individual forecasting problem (IFP) are likely more accurate than those made toward the beginning of an IFP as more information is available and the time horizon is shorter. Frequency of updates is confounded with time of update, as updates are necessarily made closer to the end of an IFP. Secondly, we hypothesize that participants who return to update their forecasts are demonstrating an interest in the forecasting problem itself. In other words, the number of updates is a proxy for the level of engagement of a participant. It stands to reason that more interested, engaged forecasters will produce better forecasts.

Table.2: Relationship between updating and forecast accuracy

Number of Updates	Mean Brier Score
1	0.45
2	0.44
3	0.42
4	0.40

4. Description of the Automated Updating Procedure

A major problem with relying on a group of forecasters over time is that most of the participants rarely update their predictions. As a result, aggregation techniques must rely on forecasts from different time periods even if newer information would have prompted forecasters to change their predictions. Ideally, all forecasts should be made simultaneously so that these time-dependent issues go away. Because this is rarely the case, methods are needed to address temporal nature of the forecasts.

As a remedy to this problem, we have been using an automated updating mechanism on a class of individual forecasting problems (IFPs) that is conducive to updating. These IFPs are characterized by an arbitrary close date and the ability for the IFP to resolve at any time prior to that date. For example, the IFP ***"Will Japan commence parliamentary elections before 1 April 2012?"*** has an arbitrary deadline of April 1 that is not linked to any official deadline and elections could commence at any time prior to that date. Therefore, this forecasting problem is a candidate for automated updating.

The basic premise behind automated updating is that as the arbitrary deadline nears the probability that the event will occur declines. A respondent's most recent forecast represents his best estimate of the probability that the event will occur on or before the deadline. If we then make an assumption about the respondent's complete probability distribution and he/she fails to update the prediction, we can update the prediction based solely on the most recent prediction. In theory, there are numerous assumptions we could make about the respondent's beliefs, but we are focusing on two in particular, linear and exponential. The SP♠DE team explored both mechanisms and we present a retrospective analysis comparing the two methods.

Figure 1: A depiction of the theory behind automated updating

4.1 Linear Updating

Suppose a respondent's most recent forecast, p_n, occurred n days prior to the IFP's arbitrary deadline and that currently there are $k < n$ days remaining. If we assume that each day the respondent would linearly update his probability toward the *status quo* response, then the daily update will depend entirely onp_n. If the default outcome is false, his update slope is $-p_n/n$ and the current updated forecast probability$p_k = kp_n/n$.

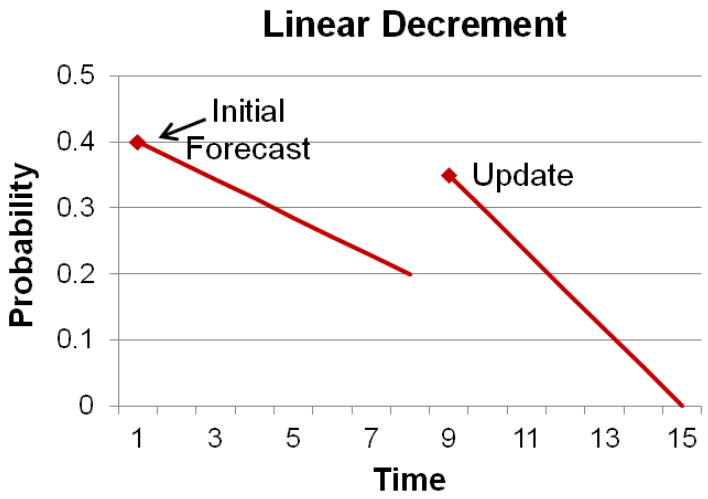

Figure 2: An example of a single forecaster's probability linearly updating as the deadline draws closer

When we take the mean of all automatically-updated forecasts for a given IFP, the converged pattern looks relatively linear as compared with the ULinOP (Figure 3). Figure 3 shows the forecasts for the IFP: ***"Will Myanmar release at least 100 more political prisoners between 21 February 2012 and 1 April 2012?"*** Note how the automated update curve converges towards a zero probability forecast over time while the ULinOP stagnates as respondents stop updating.

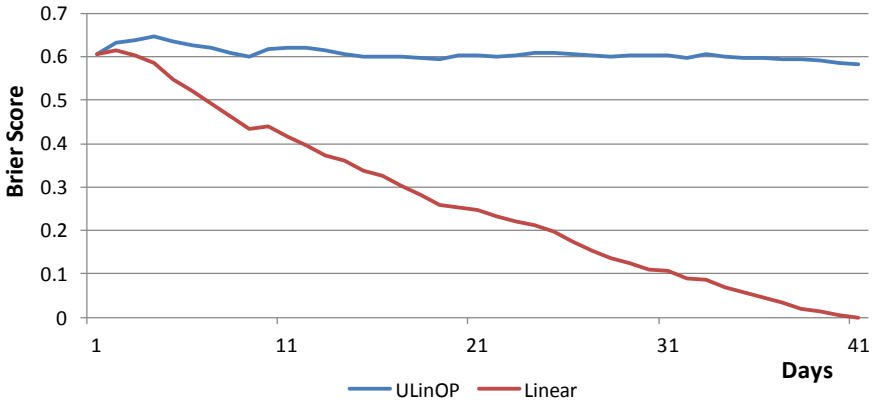

Figure 3: The ULinOP and mean linear auto-update forecast for a given IFP over time

4.2 Exponential Updating

A slightly less intuitive but perhaps more realistic assumption to make is that the event behaves according to an exponential distribution. Under this scheme, the respondent is assumed to believe that waiting time to the event is unchanged over time. The rationale is as follows:

Let X be a random variable representing the number of days it will take for a specified event to occur (e.g. the number of days from today until Palestine is recognized by the UN). Let us assume that respondent bases the forecast probability such that $X \sim \text{Exp}(\lambda)$ once we receive his forecast, we can derive his λ_r parameter and therefore know his overall probability distribution.

To illustrate, suppose the arbitrary deadline for an event to occur is in n days, and suppose respondent r has given the event a probability p_n of occurring on or before that deadline. Assuming the respondent makes his prediction under the assumption that the event is a random variable $X \sim \text{Exp}(\lambda_r)$. In this case, by the CDF of the exponential distribution

$$\mathbf{P}[X \leq n] = 1 - e^{-\lambda_r n} = p_n \quad \rightarrow \quad \lambda_r = \frac{-\ln(1-p_n)}{n}.$$

Based on this single forecast we now have a complete picture of respondent r's exponential distribution and can use this to update his forecast as the event deadline approaches; all we need to do is remember λ_r.

Now suppose that the event did not occur within the next 24 hours. We are left with $n-1$ days left until the deadline so our forecast probability should fall accordingly. Assuming the respondent doesn't update, we use λ_r to do his work for him. Our forecast p_{n-1} becomes

$$\mathbf{P}[X \leq n-1] = 1 - e^{-\lambda_r(n-1)} = 1 - \exp\left\{\frac{(n-1) * \ln(1-\mathrm{p}_n)}{n}\right\} = p_{n-1} < p_n.$$

Clearly, as the number of remaining days approaches zero, the forecast probability will tend towards zero, as desired. If the respondent decides to update the forecast manually, we simply re-compute λ_r in line with his new distribution and carry on as before (Figure 4). In the general case, if the most recent update, p_n, from respondent r came n days prior to the IFP's arbitrary deadline and presently there are $k < n$ days remaining, then his current updated forecast, p_k, will be

$$\mathbf{P}[X \leq k] = 1 - \exp\left\{\frac{k * \ln(1-\mathrm{p}_n)}{n}\right\} = p_k.$$

5. Retrospective Analysis for the Linear and Exponential Methods

Linear automated updating has performed quite well over all appropriate IFPs. Figure 5 shows automated updating performance across IFPs as compared with the ULinOP. Overall, automated updating improves Brier scores when the result is the status quo; however, given the nature of the quadratic scoring rule, performance suffers quite heavily when the event actually transpires. Such was the case for 3 forecasting problems in this study.

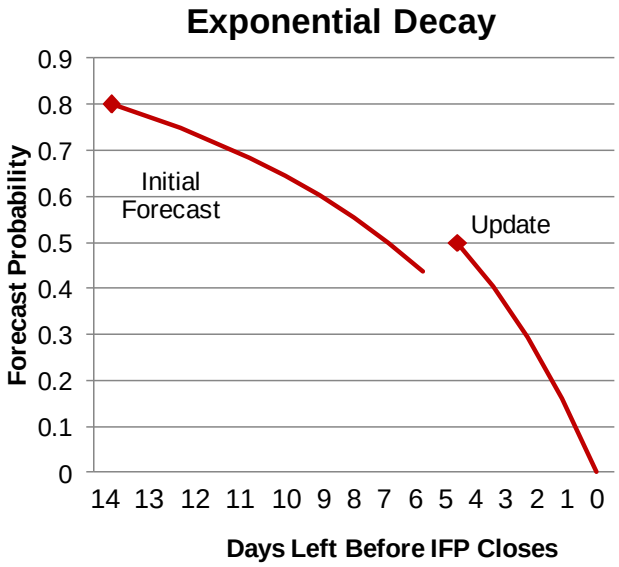

Figure 4: An example of a single forecaster's probability exponentially updating as the deadline draws closer

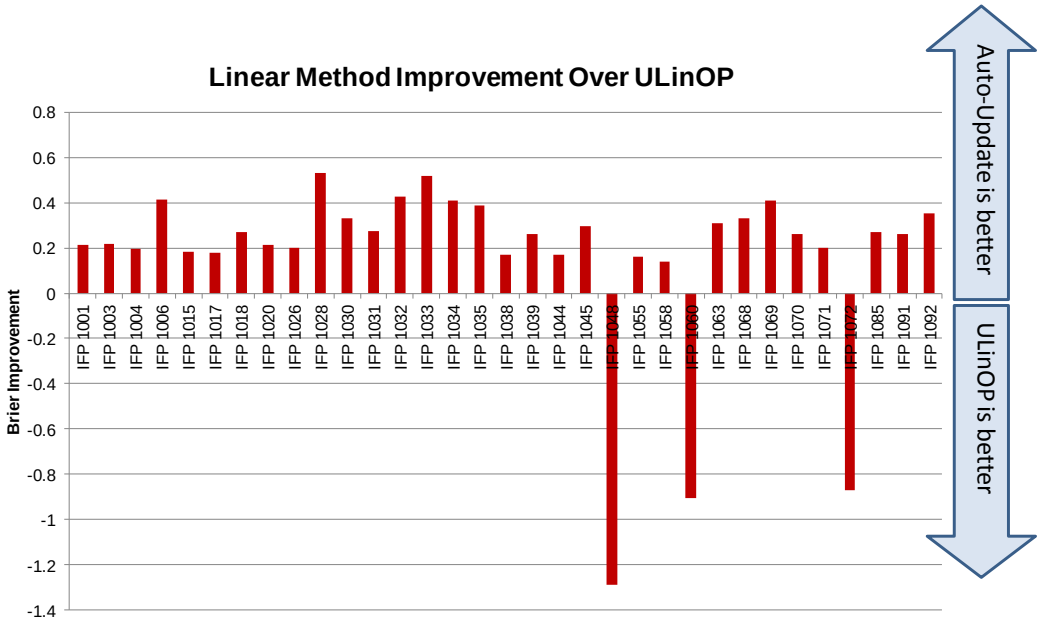

Figure 5: Difference in Brier score between automated updating and the ULinOP for all updatable IFPs

Comparing the linear and exponential auto-update schemes as applied to a given IFP, the main difference is that exponential automated updating convergence pattern is more gradual than that of linear method. While the curvature of the exponential scheme stands out when looking at a single forecast, when averaged across subjects and with the

addition of new forecasts, the resulting exponential auto-update curve resembles its linear counterpart (Figure 6). When *status quo* is less likely, the slight hedge provided by exponential automated updating may well be worth the slight performance loss among IFPs where the result ends up as *status quo*. Looking across all IFPs, we see that the exponential method is slightly more conservative than the linear updating methods (Figure 7). Thus, the exponential method reaps a slightly smaller reward when the *status quo* forecast is correct, but pays a smaller penalty when the truth departs from the *status quo*, as occurred with three IFPs in the study period.

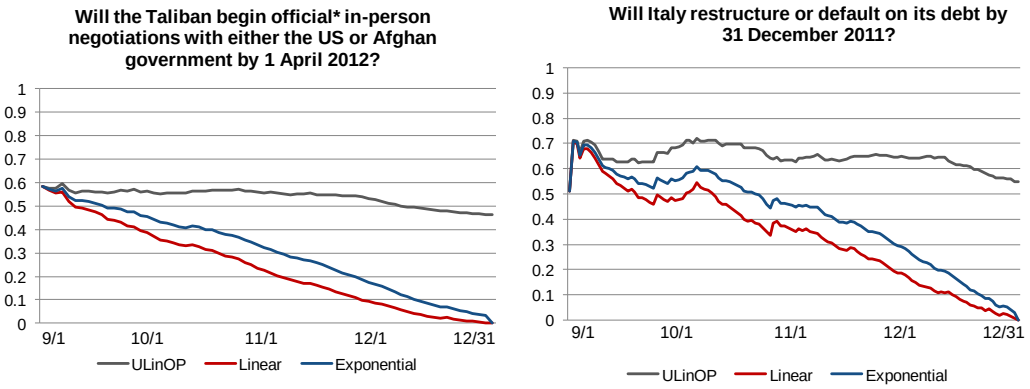

Figure 6: The ULinOP, linear, and exponential automated update forecasts for two IFPs over time

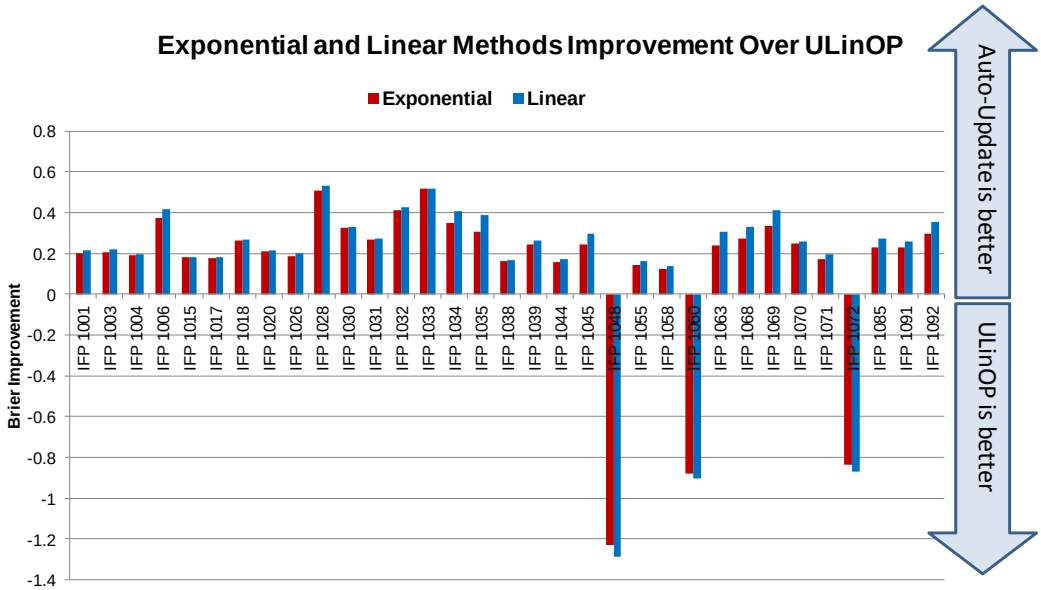

Figure 7: Difference in Brier score between automated updating and the ULinOP for all updatable IFPs

6. Conclusions and Future Research

We have presented a new method for adjusting expert forecasts based on the temporal nature of the forecasting problem. This procedure will drive individual predictions toward the *status quo* outcome, regardless of the actual forecasts. Comparison of this method to the ULinOP, based on retrospective analysis, shows substantial performance benefits as measured by the Brier score. The benefits are realized for both the linear and exponential methods of automated updating of personal predictions.

The current approach has some clear limitations and future research will investigate ways to address these concerns. In particular, the automated updating method will always drive the aggregate forecast towards the *status quo*, regardless of the evidence. A natural refinement would be to assess the personal predictions and associated data to determine if the preponderance of evidence points to a departure from *status quo*. In these cases, early detection of evidence for a departure from *status quo* could prove critical to developing an accurate forecast.

Another concern is that not all forecasters are created equal. Certain forecasters will exhibit better accuracy than their colleagues. In related work, we have explored several methods for identifying the "expert" forecasters and weighting their judgments more heavily (Forlines, *et al*, 2012; Prelec, *et al*, 2012). These papers provide strong evidence that the frequency with which an individual updates the forecasts, the self assessment of knowledge, the meta-forecast, and the recency of the forecast are all indicators of better forecast accuracy, as measured by the Brier scores in retrospective analysis. Combining these results with the current research suggests an aggregation approach that would first perform automated updating of the personal predictions, then weight each forecast by the factors cited above, to produce an aggregate forecast.

Another avenue for exploration is the use of the automated updates as an elicitation tool. When the participant has an opportunity to update a forecast, the automated update could be presented as a guideline for updating. Such an approach poses interesting questions in cognitive analysis. Would the presentation of the automated update induce an anchoring effect? What are the best methods for presenting such a decision aid? And will the presentation of this information cause the respondent to down-weight other information, such as recent news stories, that would otherwise lead to better forecasts? These issues will require additional investigation.

Acknowledgements

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 notwithstanding 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.

References

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

Hammond, K. R. (1996). Human judgment and social policy: Irreducible uncertainty, inevitable error, unavailable injustice. New York: Oxford University Press.

Surowiecki, J. (2004). The wisdom of crowds. New York: Doubleday

Sarah Miller, Alexl Kirlik, and Nathan Hendren (2011) "Applying knowledge and confidence to predict achievement in forecasting" *Human Factors and Ergonomic Society Meetings*, September 19-23, 2011, in Las Vegas, NV

Jennifer Tsai, Sarah Miller, and Alex Kirlik (2011) "Interactive Visualizations to Improve Bayesian Reasoning" *Human Factors and Ergonomic Society Meetings*, September 19-23, 2011, in Las Vegas, NV

Poore, J.C., Forlines, C., Miller, S., Regan, J., Irvine, J. "Personality, Cognitive Style, Motivation and Aptitude Predict Systematic Trends in Analytic Forecasting Behavior and Confidence." *Journal of Cognitive Engineering and Decision Making* (in press).

Sarah Miller, Clifton Forlines, John Regan, "Exploring the Relationship Between Topic Area Knowledge and Forecasting Performance" *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting*, Boston, MA, October 22-26, 2012.

Clifton Forlines, Sarah Miller, Srinivasamurthy Prakash, John Irvine, (2012) "Heuristics for Improving Forecast Aggregation" *Machine Aggregation of Human Judgment: AAAI-12 Fall Symposium*.

Drazen Prelec, Sebastian Seung, John McCoy, "Finding Truth Even if the Crowd is Wrong" *MIT working paper*, 2012.