Measuring the Forecast Accuracy of Intelligence Products

Paul Lehner and Avra Michelson
The MITRE Corporation

Leonard Adelman
George Mason University

December 2010

Abstract

Our experience has been that many in the Intelligence Community are resistant to the idea of rigorous, scientific measurement of the accuracy of analytic forecasts, preferring instead to evaluate analyses through a critical review process. Unfortunately, research and experience in other complex domains show that expert self-assessments based only on critical reviews frequently result in measurably incorrect lessons learned. In this paper we argue that the Intelligence Community should adopt a program of rigorous, scientific measurement of forecast accuracy, because such a program is essential to improving accuracy. The paper also describes a new method for measuring the accuracy of analytic forecasts expressed with verbal imprecision. The method was used to evaluate the accuracy of ten open source intelligence products, including the declassified key judgments in two National Intelligence Estimates. Results show that forecasts in these products were reasonably calibrated, with a strong positive correlation between the strength of the language used to express forecast certainty and the frequency with which forecast events actually occurred. These results demonstrate that the forecast accuracy of analytic products can be measured rigorously.

Introduction

Our experience has been that many in the Intelligence Community (IC) are resistant to the idea of rigorous, scientific measurement of the accuracy of analytic forecasts. Anecdotally, when we mention "measuring forecasting accuracy" to analysts, they typically voice strong objections. These objections fall into two broad categories: measuring accuracy is either *impossible* or *undesirable*.

Regarding the claim that it is impossible to measure accuracy, we often hear arguments such as those paraphrased below:

> "Intelligence forecasts are usually expressed with varying levels of certainty. How can accuracy be judged when no definitive forecasts were actually made?"
>
> "Ground truth cannot be known with certainty. Even in hindsight there will be many instances where we will never actually know whether the forecast event occurred."
>
> "Intelligence products impact policy decisions which in turn impact the evolution of events. A forecast event may not occur precisely because the event was forecast and policy makers took action to prevent it. Should this be counted as an inaccurate forecast?
>
> Furthermore, it's rarely known whether or not policy decisions were actually influenced by an analysis. So the nonoccurrence of a forecast event could be evidence of either an accurate or an inaccurate forecast. There is no way to tell, so there is no way to accurately measure accuracy."

Other analysts argue that whether or not accuracy can be measured, it should not be:

> "Forecasting is a fool's errand. The world is too complex and dynamic to expect accurate forecasts. Black Swans dominate. Accuracy is the wrong standard."
>
> "A high-quality intelligence analysis will provide policy makers with understanding and options. Accuracy of forecasts is secondary. A focus on accuracy distracts attention from the true value of analysis."
>
> "Many intelligence products warn policy makers of future possibilities. If analysts wait until they are confident in their warning, then they will have waited too long because policy makers will not have time to prepare. Less accurate early warnings are much more useful to policy makers."
>
> "Excellent analyses can still produce inaccurate forecasts. Poor analyses can produce accurate forecasts. Evaluation of analysis should focus on the analytic processes that were employed and whether those processes resulted in good judgments given the facts available at the time. Evaluation should not focus on accuracy."

The few open source publications that do address forecast accuracy in intelligence analysis (e.g., Wheaton and Chido, 2007; Smith 1969) express similar objections. Rather than focus on outcome accuracy, most analysts prefer an approach to evaluating intelligence analysis products that employs a critical review and lessons-learned process (e.g., Nolte, 2004). This line of reasoning has considerable merit. However, we suggest an alternative perspective: that a program of rigorous and scientific evaluation of forecast accuracy is necessary precisely because it is so difficult.

In this paper we endeavor to achieve two objectives. First, we hope to convince the reader that forecast accuracy can and absolutely should be measured with scientific rigor. Such scientific assessments of accuracy are most appropriately applied *across* collections of products so that one can draw general conclusions about overall and relative accuracy, for example, the relative accuracy of products developed with different analytic methods. Consequently, we do not argue that critical reviews of individual analyses or products are inappropriate, but simply that by themselves they are an insufficient basis for determining how to improve analysis. Second, by describing our own studies we demonstrate that forecast accuracy can be measured rigorously and that doing so will obtain useful results.

Terminology

We use the word *forecast* to signify any statement about a future occurrence, whether a discrete event (e.g. "The peace process will likely break down.") or a quantity (e.g. "GDP will probably increase more than 3% next year."). *Judgment-based forecasts* are forecasts made by experts; they are sometimes called 'estimates' or 'judgment calls.' *Probabilistic forecasts* are generally defined as forecast statements expressed with a degree of certainty (e.g., "There is a good chance the recession will end in the next quarter"), where the degree of certainty is often stated quantitatively (e.g. "There is a 70% chance that the recession will end in the next quarter.").

Furthermore we use the word *accuracy* to refer to both whether a forecasted event occured and whether the forecast was expressed with an appropriate degree of certainty. Specific measures of accuracy are presented below.

Rationale

Intelligence analysis requires knowledgeable subject matter experts (SMEs), often working collaboratively, to make judgment calls about current and future events. These judgment calls must be made in a context of sparse, unreliable and deceptive information, where the evolution of events themselves is often determined by adversaries who intend surprise and where the cost of bad judgments can be immense. It is hard to imagine a more difficult context for making good judgments.

Judgment calls, and the process of making judgments, are the subject of considerable scientific inquiry. Judgment and decision making (JDM) research has consistently shown that experts, like everyone else, are prone to a number of *judgment biases*. For example, experts often exhibit a *confirmation bias*, meaning that they seek and overvalue information confirming their current

hypothesis while simultaneously undervaluing disconfirming information (see Nickerson, 1998, for a review of confirmation bias research). In our own work we have observed this phenomenon in experiments with professional intelligence analysts (e.g., Lehner et al., 2008, 2009).

From the perspective of measuring forecast accuracy, the biases that truly matter are biases that affect *self-assessments* of performance: more specifically, biases that impact the ability of analysts, either individually or as a community, to correctly assess the accuracy of their own forecasts. If self-assessments are reasonably accurate, then there is every reason to expect that the IC's preferred approach of critical reviews and lessons-learned processes will eventually lead to better and more accurate analyses. On the other hand, if self-assessments are suspect, then any conclusions or lessons learned from such assessments are also suspect.

Are self-assessments of expert practices reliably accurate? Here the science of judgment diverges from common practice and intuition: for reasons explained below, it appears that the answer is, quite simply, "No."

To begin with, consider the *better-than-average effect*. In general people are predisposed to assess themselves as "better than average" in a great many areas. In one classic study 93% of US drivers surveyed believed that their driving skills were above average, while 88% believed themselves to be above average in safety (Swenson, 1981). In a study of university faculty, 68% of those surveyed believed they were in the top 25% in teaching performance (Cross, 1977). In yet another study, 25% percent of students believed themselves to be in the top 1% in leadership ability. Statistics such as these are commonplace, and hold whether people are assessing their own intelligence, memory ability, problem-solving skills, job performance, social skills, or almost any other trait (see Alicke and Govorun, 2005, for a recent review).

When assessing expert judgment, this better-than-average effect seems to manifest itself as a ubiquitous *expert overconfidence*. In stock market trading for example, overconfidence among professional traders is often used to explain high trading volumes (e.g., Ogden 1998). In one study with medical doctors, Christensen-Szalanski and Bushyhead (1981) found that doctors asserting 80% certainty in a diagnosis of pneumonia were correct only about 18% of the time. (Other medical studies show overconfidence, but less extreme.) Another study examining the history of science found that the true values of physical constants (speed of light, electron mass, Avogadro's number, etc.) were consistently outside the 98% confidence interval (Henrion and Fischhoff, 1986) in published studies. Similar results are seen in the twenty-five studies with engineers summarized in Lin and Bier (2008). These are just a few of many similar examples. The disciplines vary, but experts' overconfidence in their judgments appears ubiquitous.

Related to the general phenomenon of expert overconfidence is the empirical relationship among judgment accuracy, confidence and amount of information. Research suggests that increasing amounts of information increases a *confidence-accuracy disparity* (e.g., Tsai, et al., 2008). More information increases confidence more than it does accuracy, leading to greater overconfidence. Indeed, confidence often increases even when there is no change in accuracy.

Given the above results, it should come as no surprise that across many disciplines expert practitioners' assessments of their own practices often turn out to be very inaccurate. For example, the history of medicine shows that well established and accepted medical procedures are frequently proven ineffective or counterproductive.[1] Research examining verbal psychotherapy shows that the technique is effective, but that effectiveness is not correlated with any of the factors that therapists value – therapeutic method, level of training or years of experience (Wampold, 2001). Research examining the extent to which experienced law enforcement investigators and interrogators can detect whether or not a suspect is lying consistently shows that these experts are confident in their abilities but that their conclusions are actually no more accurate than those of untrained college students (Kassin, et al., 2005). Furthermore the behavioral cues they use to detect deception are not correlated with lying (Hazlett, 2006). Research in forecasting typically shows that forecasts resulting from traditional face-to-face meetings are less accurate than those obtained by averaging the pre-meeting forecasts of the meeting participants (Armstrong, 2006).

Similar results are found across many disciplines. We selected the above examples because they illustrate expert communities that routinely perform critical reviews of difficult judgments, and because the practitioners are sufficiently close to the facts of each case that one would expect them to draw more accurate conclusions about their practices. Medical doctors can observe the effects of the treatments they and others provide, psychotherapists should have deep personal knowledge of their patients and treatment outcomes, law enforcement investigators can observe the eventual resolution of their cases and one would think that forecasters would eventually notice that their post-meeting forecasts were no better than their pre-meeting forecasts.

So why do expert practitioners frequently learn measurably incorrect lessons? Certainly the above-mentioned confirmation bias contributes to this. When evaluating their own practices they are likely to look for and overweight evidence confirming the efficacy of the practices they employed. We have little doubt that these practitioners, as dedicated professionals, fully intend unbiased assessments of their own performance and believe they have succeeded, but, as summarized in Nickerson (1998), an explicit intent to avoid confirmation bias does little to mitigate it.

Even more problematic from the perspective of forecasting is *hindsight memory bias*: a tendency to incorrectly recall having made accurate forecasts. For example, in his extensive research with political analysts, Tetlock (2005, see page 149) found that the analysts' memory of their probabilistic forecasts drifted around 10% to 15% in the direction of events as they occurred. If the analyst asserted with 60% certainty that event X would occur, and the event did occur, then the analyst will probably remember having asserted 70% certainty or greater. If the event did not occur, the analyst might well recall having asserted 50%. In general, the strength of the hindsight memory bias increases with time (see Schacter, 2001, for an introduction to memory biases).

Thus, overall there is considerable evidence to support a claim that expert practitioners, and communities of practitioners, tend to overrate the accuracy of their judgments and consequently

[1] For examples we recommend that readers look at news archives for reports on the use of steroids for head injuries circa 2004, post-heart attack angioplasty circa 2006, or vertebroplasty circa 2009.

tend to draw correspondingly poor lessons about the effectiveness of different practices. This seems to occur despite every intention of drawing unbiased lessons learned from experience.

The situation with intelligence analysts and the evaluation of analytic forecasts may be even worse. In addition to the predispositions described above, we hypothesize that analysts may also be prone to a strong *hindsight interpretation bias*. A review of analytic products quickly reveals that forecasts are often expressed with verbal imprecision. Expressions such as "fair chance" and "might happen" or "could occur" are quite common. This leaves open the opportunity to claim success whether or not the forecast event occurs. We suspect that analysts often remember their "may occur" forecasts as having meant "probable" or "greater than 50% chance" when the event occurs and as having meant "improbable" or "less than 50% chance" when the event does not occur. We would not brand such selective interpretations as disingenuous; because of memory biases, they probably represent what analysts honestly remember about "what they were thinking at the time."

Because of these biases, analysts may have a natural tendency to overestimate the historical accuracy of their forecasts. Further, repeated biased assessments of their forecasting success may lead analysts to develop a robust illusion of forecasting accuracy.

Unfortunately, such an illusion would have two serious consequences. First, analysts will see little reason to change their analytic practices. If individual analysts, or communities of analysts, incorrectly believe that their forecasts are reasonably accurate, then they have little incentive to change their analytic practices. Indeed, they would rightfully resist change. Second, even if there were a successful push for change, it would be impossible to determine whether the change resulted in more or less accurate forecasts, since, as noted above, measurably incorrect lessons learned are often widely accepted.

All of this leads to a simple conclusion: there is a clear need to methodically score the accuracy of analytic forecasts.

In this light, we now reconsider the above-paraphrased objections to measuring accuracy. Some of the objections rested on the notion that the world is so complex, dynamic and contrarian that a simple scoring of accuracy is misleading. We argue the opposite. The more difficult it is to evaluate accuracy, the greater is the opportunity for self-assessments to be unintentionally biased and therefore the greater the need for rigorous measurement of accuracy. The fact that assessing accuracy on a case-by-case basis is difficult and nuanced is precisely why it is necessary to measure accuracy with scientific rigor.

Method

With few exceptions, forecasts in intelligence products are judgment based and probabilistic, but the degree of certainty is rarely expressed quantitatively. Yet metrics for the accuracy of probabilistic forecasts in scientific studies assume such quantitative expression. Our method for measuring the accuracy of products where forecasts are expressed with verbal imprecision is

described in detail in Lehner et al. (2010). However, in essence our approach is founded on two basic ideas: *inferred probabilities* and *blind retrospective assessments* of ground truth.

As an illustration, consider the following forecast statement from the declassified key judgments in the 2007 National Intelligence Estimate (NIE) on *Prospects for Iraq Stability*: "… the involvement of these outside actors is not likely to be a major driver of violence …" The forecast event (as we coded it) is "The involvement of outside actors will not be a major driver of violence in Iraq in the January 2007 to July 2009 time frame." In the first step, five different reviewers read the NIE and on the basis of what was written inferred probabilities of 80%, 85%, 75%, 85%, and 70% for the forecast event. As occurred in this case, if multiple reviewers infer similar probabilities, then the average of those inferred probabilities is a fair representation of how intelligence consumers would consistently interpret the product. On the other hand, if multiple reviewers infer very divergent probabilities then that divergence is measurable evidence that the forecast statement was largely meaningless.

Once inferred probabilities have been assigned to the forecast events a retrospective analysis is used to estimate ground truth. As in any scientific investigation, there is never an error-free determination of ground truth. Rather, estimates or measurements of ground truth are carefully designed to be unbiased relative to the hypotheses being evaluated. It is unbiased error in ground truth measures that makes data amenable to scientific/statistical analysis.

In the second step of our approach we asked SMEs to estimate retrospectively whether each event occurred, and ensured that the SMEs did not see the original forecasts or inferred probabilities. In some cases they were unsure as to whether the event actually occurred ("I *think* it happened") and in some cases they found the event statement so poorly worded that they could not answer it at all ("What the heck does 'major driver' really mean?"). These responses were perfectly acceptable, since our goals were to evaluate overall accuracy, infer trends and draw generalizations; and not to definitively evaluate the accuracy of any one forecast or forecast product. In the case cited here on outside influence we had two SMEs independently assess ground truth. One SME retrospectively rated the forecast event as definitively false and a second SME could not answer.

As described in Lehner et al. (2010), we have applied the inferred probability method to ten products:[2]

Jane's 2006 Forecast for Iran
STRATFOR 2006 Forecast for Iran
STRATFOR 2006 Forecast for South Africa
STRATFOR 2006 Forecast for Sudan
Jane's: US and Iran: Road Map to Conflict (Feb 2007)
STRATFOR 2007 Forecast for Iran
STRATFOR 2007 Forecast for South Africa

[2] The STRATFOR documents can be found at www.stratfor.com. The Jane's documents can be found at www.janes.com. A paid subscription is required to access the documents at both web sites. The two NIEs examined in this study may be found at http://www.dni.gov/nic/special_keyjudg_iraq_2007.html and http://www.dni.gov/press_releases/Declassified_NIE_Key_Judgments.pdf.

STRATFOR 2007 Forecast for Sudan
NIE 2006 Prospects for Iraq Stability (declassified key judgments)
NIE 2006 Trends in Global Terrorism (declassified key judgments)

The results are summarized in Figure 1. The figure shows a *calibration curve*, which is a standard method for depicting forecast accuracy. Across the ten documents there were 144 forecast events for which the SMEs could assess ground truth. The inferred probabilities for these events ranged from 10% to 100% certainty. For example, there were 15 events where the average inferred probability rounded to 60%. If inferred probabilities were perfectly calibrated, then exactly 9 of those 15 events would have occurred (9/15 = 60%). In fact, 7 of those 15 events (47%) actually occurred.[3]

This data shows several clear and interesting trends. First there is a strong correspondence between the strength of the certainty language used in the documents and the frequency with which forecast events occurred, at least as measured by inferred probabilities.[4] Second, the calibration curve is reasonable, by which we mean that the certainty expressions provide useful information for readers. The least calibrated level is 50%, but there were only 8 data points at this level. By contrast, there were 34 forecast events where the inferred probability was 80%, and 82% (28 of 34) of those events occurred. Overall across the different forecast certainty levels there is about a 10% difference between the level of forecast certainty and the relative frequency of occurrence.

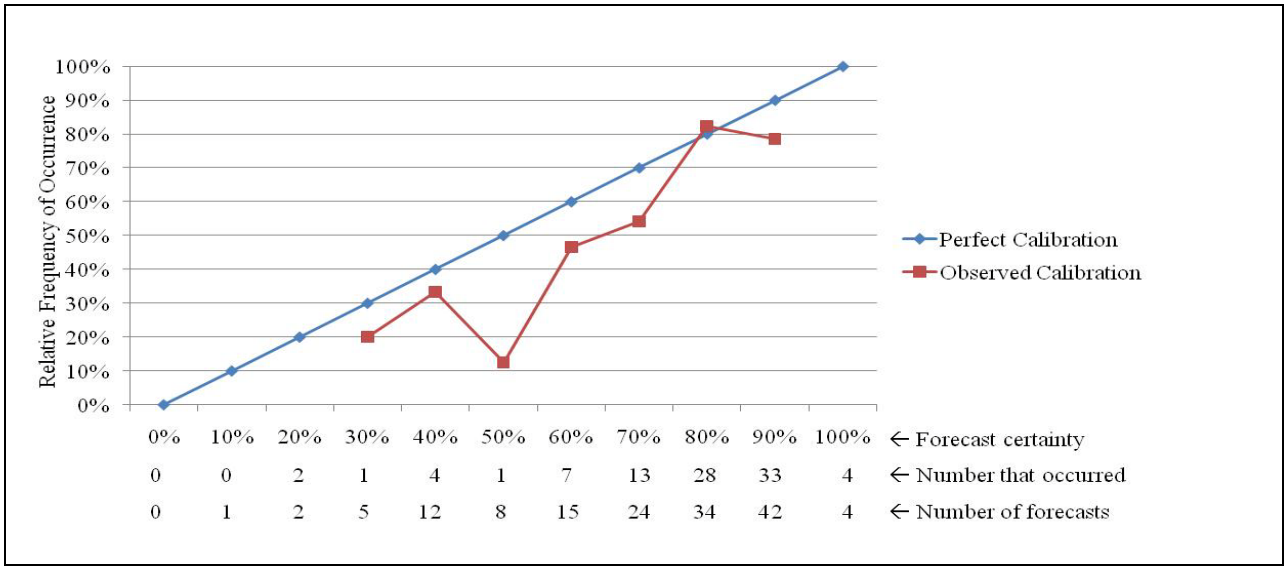


Figure 1: Calibration Curve for the Combined NIE, Jane's and STRATFOR Forecast Events

[3] The line of Observed Calibration only includes probability levels with 5 or more forecasts.
[4] For the seven levels with 5 or more forecast events, the correlation between forecast certainty and relative frequency of occurrence is .89, which is statistically significant (p=.002).

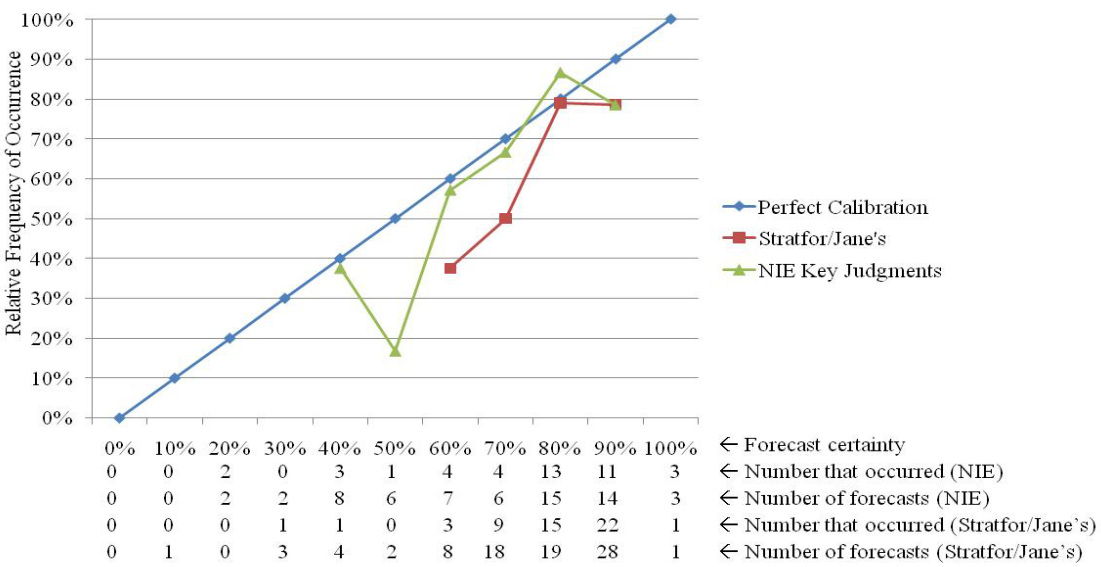
Third, there is some evidence of overconfidence. For six of the seven forecast certainty levels, the observed relative frequency is below the level of perfect calibration.[5] This suggests that in general the forecast certainties expressed in these products were greater than was warranted.

Fourth, forecast certainties in this collection appear somewhat conservative. Only 3% of the forecast events (4 out of 147) had a forecast certainty of 100%, and many were close to 50%. By comparison, in a study by Mandel (2009) where intelligence analysts directly expressed certainty with numerical probabilities, around 32% (159 out of 560) of the forecast events were expressed with 100% certainty.

In summary, the inferred probability method shows the forecasts in these intelligence products to be somewhat conservative, reasonably calibrated, yet still slightly overconfident. All in all the results provide a coherent picture of forecast accuracy.

Comparing the accuracy of different collections of forecasts

As an example of a comparative analysis, Figure 2 shows the difference between the NIEs, which originally were produced as classified products and the open source products. Visually they appear similar, and the description of the forecasts as somewhat conservative, reasonably calibrated yet still a little overconfident applies equally to both.



0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	
0	0	2	0	3	1	4	4	13	11	3	← Number that occurred (NIE)
0	0	2	2	8	6	7	6	15	14	3	← Number of forecasts (NIE)
0	0	0	1	1	0	3	9	15	22	1	← Number that occurred (Stratfor/Jane's)
0	1	0	3	4	2	8	18	19	28	1	← Number of forecasts (Stratfor/Jane's)

Figure 2: Calibration Curves for NIE and STRATFOR/Jane's Inferred Probabilities

We urge the reader not to draw strong substantive conclusions from this comparison. Only ten documents were examined, with only two of them NIEs. Instead, readers should simply

[5] This effect is statistically significant (p=.036 in a 2-tailed paired t-test comparing difference between perfect and observed calibration).

recognize that this type of accuracy measurement and comparative accuracy analysis are straightforward to perform.

Revisiting the Paradox of Warning

Overall we conclude that the inferred probability method represents a significant advance toward determining whether or not the accuracy of analytic forecasts can be measured. Obviously they can. But two of the objections listed in the introduction merit further examination. The first is that forecasts affect policies, which in turn affect the events being forecast, thereby invalidating the accuracy of any accuracy measurements. This is sometimes called the "paradox of warning." In our view, this is a poor argument against data collection, simply because it conclusively interprets data before any data is actually collected. We do, however, consider it an interesting *hypothesis* that can be both tested (Does this effect exist?) and measured (If it exists, how large is the impact?) with good data collection. In the limited empirical data we collected, we saw no hint of the warning paradox. The product forecasts were reasonably calibrated, and to argue that the forecasts substantially affected events is to also suggest that the observed calibration represented a misinterpretation and that actual forecast accuracy was much worse than measured. Second, it would be possible to measure the impact of the "paradox of warning" on forecast accuracy by, for example, separating forecasts into those that clearly had an impact on events, those that may have had an impact on events, and those that clearly had no impact on events, and then comparing forecast accuracy across these groups. *If* results show that the paradox of warning has a serious impact, then that would be an interesting general finding. Furthermore one could statistically adjust accuracy measures to account for this effect by obtaining independent/blind assessments of the forecast impact and treating those assessments as a statistical covariate.

Even if the paradox of warning has no measurable impact on forecast accuracy, we recognize that useful early warning usually addresses improbable events. Waiting until the event is assessed as probable often means waiting too long because policy makers no longer have sufficient time to prepare. This actually poses no problem for measuring accuracy: if early warnings are expressed as improbable events (e.g. "There is a small chance that …") then the forecasts are calibrated if most of those events *do not* occur.

Integrating two approaches to improve analysis

Consider a document that forecast "There is only a very small chance that the peace process will succeed. War will almost certainly break out between the two factions." Now suppose that the peace process does succeed. How should this forecast be evaluated? Should the analytic process be critically reviewed to determine "What went wrong?" Or should it be evaluated with the type of accuracy measures described in this paper?

We think the answer is clear: in this case, the critical review is the appropriate method. Measures such as calibration contribute very little to understanding the quality of any one forecast or forecast document. Instead, these measures are designed to draw inferences in the aggregate about collections of forecasts.

But now consider the policy maker who complains, "You misled me last time, why should I trust your estimates next time?" Little in the critical review of a single case can answer this question. The best that could be offered is the assurance that "We will not make the same mistake again" – if indeed there was a mistake. By contrast, on the basis of aggregate accuracy measures such as calibration it would be fair to respond "… when our products assert 'very small chance' the event only occurs around 10% of the time; and when our products assert 'Almost certainly' the event occurs around 90% of the time. There are no guarantees, but in general our forecast certainties track well with eventual reality."

Reviews of individual analytic products provide information about what may or may not have worked in one particular instance. In our view they create a good foundation for developing new ideas and *hypotheses* about how to improve tradecraft. But lessons learned from individual analyses should not be considered definitive. As documented earlier in this paper, lessons learned in this way often turn out to be measurably incorrect. Rather, they should be treated as tentative conclusions that still require rigorous testing either through experiment (when that is feasible) or through rigorous field measurement of approaches that were put into practice. In this paper we have tried to show that in the area of analytic forecasting such rigorous field measurements are eminently achievable.

Epilogue: Testing our own hypothesis

In the introduction we hypothesized, based on the available science, that analysts naturally develop a robust illusion of forecasting accuracy. However, when we measured the accuracy of a collection of forecasts in ten analytic products the results showed that analytic forecasts were somewhat conservative, reasonably calibrated and only a little overconfident. Overall, this accuracy profile was completely *inconsistent* with our hypothesis. While more data is needed to reach definitive conclusions, the demonstrated feasibility of collecting data that tested our hypotheses about intelligence analysis is itself testimony to the value of measuring accuracy with scientific rigor.

REFERENCES

Alicke, M. and Govorun, O. "The Better-than-Average Effect." In Alicke, M. and Dunning D. (eds.), *The Self in Social Judgment*, Psychology Press, New York, 2005.

Armstrong, J. S. "How to make better forecasts and decisions: Avoid face-to-face meetings," *Foresight - The International Journal of Applied Forecasting*, 5, 3–8, 2008.

Cross, P. "Not can but will college teachers be improved?" *New Directions for Higher Education*, 17, 1-15, 1977

Hazlett, G. "Research on Detection of Deception: What We Know vs. What We Think We Know." In Fein, R., Lehner, P. and Vossekuil, B. (eds.) *Educing Information: Interrogation--Science and Art: Foundations for the Future: Phase I Report.* National Defense Intelligence College Press, December 2006.

Henrion, M. and Fischhoff, B. "Assessing Uncertainty in Physical Constants," *American Journal of Physics*, 54(9), 791–798, 1986.

Kassin, S., Messner, C. and Norwick, R. "I'd Know a False Confession if I Saw One": A Comparative Study of College Students and Police Investigators," *Law and Human Behavior*, 29(2), 2005.

Lehner, P., Adelman, L., Cheikes, B. and Brown, M. "Confirmation Bias in Complex Analyses," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(3), 584–592, 2008.

Lehner, P., Adelman, L., DiStasio, R., Erie, M., Mittel, J. and Olson, S. "Confirmation Bias in the Analysis of Remote Sensing Data," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 39(1), 218–226, 2009.

Lehner, P., Michelson, A., Adelman, L. and Goodman, A. "Using inferred probabilities to measure the accuracy of imprecise forecasts," submitted to *Judgment and Decision Making*, 2010.

Lin, S. and Bier, V. "A Study of Expert Overconfidence," *Reliability Engineering and System Safety*, 93, 711–721, 2008.

Mandel, D.R. (February, 2009). *Canadian perspectives: A calibration study of an intelligence assessment division*. Paper presented at the Global Futures Forum Community of Interest for the Practice and Organization of Intelligence Ottawa Roundtable "What Can the Cognitive and Behavioural Sciences Contribute to Intelligence Analysis? Towards a Collaborative Agenda for the Future." Meech Lake, Quebec.

Nolte, W., "Preserving Central Intelligence: Assessment and Evaluation in Support of the DCI," *Studies in Intelligence*, 48(3), 2004.

Nickerson, R.S. “Confirmation bias: A ubiquitous phenomenon in many guises,” *Review of General Psychology*, 2, 175–220, 1998.

Odean, T. “Volume, volatility, price, and profit when all traders are above average,” *Journal of Finance*, 53(6), 1887–1934, 1998.

Schacter, D. *The Seven Sins of Memory*. Houghton Mifflin, 2001.

Smith, A. "On the Accuracy of National Intelligence Estimates." *Studies in Intelligence*, 13(4), 25-35, 1969.

Svenson, O. "Are we all less risky and more skillful than our fellow drivers?" *Acta Psychologica* **47**(2), 143–148, 1981

Tetlock, P. *Expert Political Judgment.* Princeton University Press, 2005.

Tsai, C., Klayman, J. and Hastie, R. (2008) Effect of amount of information on judgment accuracy and confidence *Organizational Behavior and Human Decision Processes*, 107(2), 97–105, 2008.

Wampold, B. *The Great Psychotherapy Debate: Models, Methods and Findings.* Lawrence Erlbaum Associates: Mahwah, New Jersey, 2001.

Wheaton, K. and Chido, D. “Evaluating Intelligence,” *Competitive Intelligence Magazine*, 10(5), 2007.