How Good Are Strategic Intelligence Forecasts?

David R. Mandel

Former CIA veteran, Sherman Kent, who is widely regarded as the father of modern intelligence analysis, noted a half-century ago that most substantive intelligence is comprised of human judgment rather than cold, hard fact. Kent further pointed out that the judgments that matter the most to decision-makers are those that provide an indication of future conditions. Such forecasts can help decision-makers, such as state leaders or military commanders, anticipate future events, make informed decisions, and avoid strategic surprises.

Seven years ago, Alan Barnes and I set out to address an important but previously unanswered question: How accurate are the forecasts made by strategic intelligence analysts? We studied actual forecasts that analysts made in their day-to-day work. Some examples (edited to remove sensitive information) include: **"**It is very unlikely [1/10] that either of these countries will make a strategic decision to launch an offensive war in the coming six months**"** and **"**The intense distrust that exists between Country X and Group Y is almost certain [9/10] to prevent the current relationship of convenience from evolving into a stronger alliance.**"**

The results of that inquiry were recently reported in a paper entitled "Accuracy of Forecasts in Strategic Intelligence" that was published this year ahead of print in the *Proceedings of the National Academy of Sciences*. The question—*How good*?—is important for several reasons. Let me name just two. First, intelligence is a costly enterprise. I don't know the Canadian figures, but it runs in the tens of billions each year in the US. Second, intelligence is a consequential enterprise, affecting not only national security (and hence, to some extent, personal security), but also national opportunity. Capitalizing on opportunities requires leaders and policy makers to have good foresight, and intelligence can support that.

Alan and I came at the topic from different perspectives. I am a behavioral scientist who studies judgment and decision-making under conditions of uncertainty. At the time, Alan directed a division of strategic analysts focused on the Middle East and Africa within the Privy Council Office's Intelligence Assessment Secretariat. Following Kent's lead, he had implemented several standards in his division that were designed to foster analytical rigor. Among them was a set of verbal terms for communicating uncertainty in forecasts, which had numerical equivalents that were recorded. For instance, the term "very unlikely" that appears in the first example I gave earlier was equated with a 10% probability (which is what the 1/10 in brackets referred to), while "almost certain" was equated with a 90% probability (or 9/10). Although the numbers were not included in the final reports that were sent to consumers, they did serve two important functions. First, they reminded analysts what the words they were using were supposed to mean. Second, they provided an invaluable source of information that could be used to answer the accuracy question.

When we met, Alan was already tackling that question by keeping track of how events in the world that were forecasted actually turned out. To know how accurate forecasts are, one must also know what eventually happened. Yet such *ex post facto* tracking is hardly ever done by intelligence organizations. That's not too surprising either, given how much time it takes to keep careful, systematic track of the past.

At a fortuitous meeting where Alan described his early efforts, I pointed out that he had effectively set up the conditions for a quantitative study of forecasting accuracy, and I offered to turn the statistical crank. As the study developed, we refined the methods for doing the research, and what started out as purely a quality control exercise developed into a broader applied research question. It was a wonderful evolution.

Drawing on a comprehensive set of forecasts from Alan's division over a roughly six-year period, we extracted just over 1,500 forecasts whose outcomes could be unambiguously coded as having occurred or not occurred for which we also had the analysts' numerical probability estimates of the events' occurrence. We also kept track of other factors, such as whether the forecaster was a junior or a senior analyst, whether it was an easier or harder call to make, and whether it was deemed to be of higher or lower importance for policy decision-making.

We examined several quantitative measures of forecasting quality, but the main results can be summarized non-technically in two key points. First, we found very good accuracy. Excluding a small number of cases where the analyst assigned a 50-50 probability—the proverbial coin-toss that points to no specific outcome—94% of the forecasts pointed decision-makers in the right direction. That is, when the events actually didn't occur, most forecasts were issued with probabilities lower than 50%. Likewise, when the events actually did occur, most were issued with probabilities higher than 50%.

Secondly, we found that intelligence forecasts were very well calibrated. In other words, if we look at a set of forecasts of a given probability level, say 75%, we find that about 75% of those forecasted events actually occurred. That is, the forecasters' probabilities are proportional to the relative frequencies of event occurrence observed in the real world. The slight imperfections in calibration that we observed primarily reflected underconfidence. This is unsurprising given the high accuracy rate we observed. It is difficult to be overconfident when you are correct most of the time. We also showed how intelligence organizations could take steps to minimize miscalibration by remapping forecasts in such a way that debias them.

Interesting patterns of variability in forecast quality also emerged. Senior analysts were better than junior analysts at discriminating events that occurred from those than didn't, indicating that analytical experience benefits forecasting accuracy. A somewhat surprising finding was that harder forecasts showed even greater underconfidence than easier forecasts. As Baruch Fischhoff noted in the May-June 2014 issue of *Policy Options*, a robust finding from his research was that people show overconfidence for harder judgments and underconfidence for easier ones (the so-called hard-easy effect). We observed the opposite, a finding that we attribute to the accountability pressures on analysts to get it right, especially on the harder and more important calls. In line with that explanation, we found that analysts were more underconfident when forecast importance

to policy makers was especially high. In such cases, analysts seem to exercise greater caution.

Our findings could not have been easily inferred from the literature on forecasting quality, which suggests that even expert forecasters are overconfident, often eking out performance levels only nominally better than a dart-throwing chimp. For instance, Philip Tetlock found that political science experts who made geopolitical forecasts could only account for about a third of the outcome variance explained by strategic analysts in our study. Indeed, the analysts in our study far surpassed the performance of even the best statistical models that Tetlock had tested. The many differences between the studies, however, make quick explanations of differences unreliable. I am skeptical of overconfident assertions on both sides: those that would claim an unambiguous forecasting victory for the intelligence community and those that seek to explain away the positive results.

What I am more certain of is that this study provides a useful policy-option model for intelligence accountability in at least one sphere of its activity—forecasting. Intelligence communities rely mainly on process-accountability practices (such as the U.S. Office of Director of National Intelligence's Intelligence Community Directive 203). This study offers, in contrast, a clear example of how managers could use outcome-accountability practices to keep score of intelligence product quality. How best to set the balance between process and outcome accountability criteria is ultimately a question for policy makers to address, but there too behavioral science could help clarify the costs, benefits, and implementation challenges.