*Intelligence and National Security*, 2014
http://dx.doi.org/10.1080/02684527.2014.980534

ARTICLE

# Why Assessing Estimative Accuracy is Feasible and Desirable

JEFFREY A. FRIEDMAN* AND RICHARD ZECKHAUSER

ABSTRACT The US Intelligence Community (IC) has been heavily criticized for making inaccurate estimates. Many scholars and officials believe that these criticisms reflect inappropriate generalizations from a handful of cases, thus producing undue cynicism about the IC's capabilities. Yet there is currently no way to evaluate this claim, because the IC does not systematically assess the accuracy of its estimates. Many scholars and practitioners justify this state of affairs by claiming that assessing estimative accuracy would be impossible, unwise, or both. This article shows how those arguments are generally unfounded. Assessing estimative accuracy is feasible and desirable. This would not require altering existing tradecraft and it would address several political and institutional problems that the IC faces today.

The US Intelligence Community (IC) has been heavily criticized for making inaccurate estimates, especially following its flawed assessments of Iraq's weapons of mass destruction programs. It is clearly unfair to judge intelligence capabilities based on a handful of perceived failures. Such selective judgment creates undue cynicism and neglects the inherent limitations of grappling with uncertainty in world politics.[1] Yet if we wish to view a more balanced picture of how accurate intelligence estimates are, there is essentially nowhere to turn.

Since its inception, the IC has not systematically evaluated the accuracy of its estimates. Product evaluation teams grade the tradecraft by which intelligence reports are produced, and post-mortem studies seek explanations for why the IC 'got it wrong' in salient cases, but these efforts do not support general inferences about estimative accuracy. Thus, even as the IC maintains that it receives insufficient credit for its performance, it does not measure that performance in a manner that might indicate how much credit is due. 'Just

*Corresponding author. Email: jeffrey.a.friedman@dartmouth.edu

[1] On this bias in debates about intelligence, see Richard K. Betts, *Enemies of Intelligence: Knowledge and Power in American National Security* (NY: Columbia 2007) and Paul R. Pillar, *Intelligence and US Foreign Policy: Iraq, 9/11, and Misguided Reform* (NY: Columbia 2011).

how bad is it out there?' asks one scholar: 'The short answer is, no one knows'.[2]

This state of affairs is problematic, especially because evaluating estimative accuracy would produce several positive benefits. Analysts and managers could better track, improve, and receive credit for their performances. Career incentives could foster the writing of accurate estimates rather than encouraging more problematic strategies for advancement. Decision makers could better appreciate the reliability of the material they receive. Congressional oversight could focus on patterns rather than on outliers, and the broader public might learn that the IC performs better than what headlines and hearings imply.

This topic deserves greater study because, as this article demonstrates, the most prominent objections to assessing estimative accuracy are conceptual, not technical. What does 'accuracy' even mean in this context? How can we evaluate the accuracy of estimates that deal with different subjects? How can we distinguish between poor judgment and bad luck? These questions demand careful consideration, but they are far more tractable than many scholars and practitioners believe.

Section 1 defines the scope of this analysis; it then describes the status quo and its flaws. Section 2 explains why assessing estimative accuracy is feasible, and why prominent arguments to the contrary are mistaken. In contrast to previous work on this subject, we argue that it is possible to evaluate the accuracy of intelligence reports without requiring major changes to existing tradecraft. Section 3 then disputes the widespread notion that assessing estimative accuracy would harm the IC or its personnel. Indeed, the efforts we describe would help to mitigate several long-standing political and institutional problems.

Section 4 concludes by providing more detail on what a theoretically-grounded and practically-oriented system for evaluating estimative accuracy might look like. This system would not generate a 'one-size-fits-all' standard for evaluating intelligence. As we emphasize throughout, accuracy is only one component of an intelligence estimate's overall quality, and all information about performance should be handled with care. But that does not mean that this information is not valuable to collect. In this article, we explain how assessing estimative accuracy is both more feasible and more desirable than conventional wisdom allows.

## Section 1. The Current System and its Flaws

This article proposes assessing the accuracy of intelligence estimates. We use the term 'intelligence estimates' to comprise published assessments of uncertainty. This includes predictions about the future as well as judgments about existing affairs. For an example of the latter, a controversial 2007

[2] Woodrow J. Kuhns, 'Intelligence Failures: Forecasting and the Lessons of Epistemology' in Richard K. Betts and Thomas G. Mahnken, *Paradoxes of Strategic Intelligence: Essays in Honor of Michael I. Handel* (Portland, OR: Frank Cass 2003) p.82.

National Intelligence Estimate (NIE) assessed that Iran had halted its nuclear weapons program.[3] Even though this dealt with a previous event that might have been confirmed with certainty, there was insufficient information to prove this judgment, thus it constituted an intelligence estimate.

NIEs are the most prominent form of intelligence estimates, but we use the term 'estimates' broadly to include assessments of uncertainty in many other forms. For instance, the IC publishes such products as the World Intelligence Review (WIRe) and the President's Daily Brief (PDB). At lower echelons, individual analysts or teams write and circulate a wide range of reports. Briefings and written products can contain many distinct estimates. For example, the 'Key Judgments' section of the 2007 Iran NIE assessed when Iran had halted its nuclear weapons program, whether it had restarted that program, and how much fissile material Iran had imported. Since the term 'intelligence estimate' is often applied to a written report as a whole, it is more precise to define the scope of this article as evaluating the accuracy of 'estimative statements', though we use the terms interchangeably.

Estimates are but a subset of intelligence reporting. The transcript of an intercepted phone conversation or an eye-witness account of a terrorist's whereabouts might be important pieces of information, but if they represent known facts, they lie outside the scope of our analysis. (Any uncertain extrapolations from those facts *would* be estimates.) 'Estimative intelligence' is sometimes contrasted with 'current' or 'tactical' intelligence, but many statements about current events or tactical affairs involve uncertainty and constitute estimates for our purposes. As Sherman Kent described it: 'Estimating is what you do when you do not know … In this broad sense, scarcely an intelligence document of any sort goes out to its consuming public that does not carry some sort of estimate'.[4]

Some intelligence estimates (and almost all high-level estimates) are corporately authored. For example, NIEs reflect input from the IC as a whole, and they are usually coauthored by representatives from all relevant agencies. Other estimates are written by analysts or teams, such as briefings and memoranda responding to specific questions from policymakers, or reports disseminated in the WIRe and the PDB. Authorship is rarely exclusive to any person or group and different forms of intelligence reporting will obviously have different numbers of authors. When assessing estimative performance, one could focus attention on different units of analysis: individuals, teams, divisions, agencies, or even the IC as a whole.

Accuracy is a critical component of evaluating intelligence, but it is hardly the only one.[5] For instance, it is important to know how well intelligence

[3] *Iran: Nuclear Intentions and Capabilities* (November 2007).

[4] Sherman Kent, 'Estimates and Influence' in Donald P. Steury (ed.) *Sherman Kent and the Board of National Estimates: Collected Essays* (Washington, DC: Center for the Study of Intelligence 1994) p.53.

[5] On the multidimensional nature of intelligence 'performance', see Stephen Marrin, 'Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?', *Intelligence and National Security* 27/6 (2012) pp.896–912.

estimates anticipate strategic surprises, as simply identifying plausible catastrophes is the first step toward preventing them. Moreover, even if intelligence reports are accurate and prescient, there is still the question of how much they influence decision makers: insight without influence has limited value. Notwithstanding these other important considerations, this article focuses on accuracy. Unless otherwise noted, all mentions of 'performance' refer to this specific issue. (Section 4, however, explains how a system for evaluating estimative accuracy could easily be extended to examine how well the IC anticipates certain kinds of events and how timely are its assessments, two topics that are tractable and important in their own right.)

The IC has never established a system for evaluating the estimative accuracy of its products writ large.[6] Thomas Fingar, who served as Deputy Director of National Intelligence for Analysis from 2005 to 2008, observes that 'there is no mechanism to evaluate collective performance' and that it is currently impossible to determine 'how well individual analysts and analytic units perform'.[7] The lack of systematic data on estimative accuracy feeds the perception that debates on the subject are politicized and uninformed: 'relentlessly partisan second-guessing' instead of empirically grounded critique.[8]

This is not to say that analysts and teams escape evaluation of any kind. For example, the IC systematically evaluates the process by which many estimates are produced. In 2007, the Office of the Director of National Intelligence (ODNI) created the position of Assistant Deputy Director for Analytical Integrity and Standards, whose office is responsible for reviewing intelligence products and ensuring that they 'meet the highest standards of integrity and rigorous analytic thinking'. Intelligence Community Directive (ICD) 203 defines these standards as objectivity, independence from political considerations, timeliness, use of all available sources of intelligence, and demonstration of 'proper standards of analytic tradecraft'.[9] Each of these attributes addresses high-profile criticisms of how the IC develops and articulates its judgments, and it is reasonable to assume that improvement in these attributes would lead to more accurate estimates. Yet these attributes do

[6] At least, there is none that is known to the public or that has played any meaningful role in public debates.

[7] Thomas Fingar, *Reducing Uncertainty: Intelligence Analysis and National Security* (Stanford, CA: Stanford Security Studies 2011) p.34.

[8] Philip E. Tetlock and Barbara A. Mellers, 'Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong', *American Psychologist* 66/6 (2011) p.544. Cf. Ibid., p.34.

[9] The standards of proper analytic tradecraft, in turn, are defined as: description of the quality and reliability of sources; expression of uncertainty or confidence in analytic judgments; differentiation of facts from assumptions; incorporation of alternative analysis where appropriate; relevance to national security; logical argumentation; and reference to previous analyses. ICD 203 includes accuracy as an eighth tradecraft standard, but this is problematic. One could assume that better tradecraft will correlate with better accuracy, but this presupposes that tradecraft and accuracy are different things (and, indeed, this connection can only be evaluated if tradecraft and accuracy are defined separately).

not directly measure accuracy: it is perfectly plausible that a group of analysts could employ excellent tradecraft and still produce inaccurate estimates.[10]

Post-mortem reviews of estimates that proved to be off the mark represent another prominent form of product evaluation in the IC. By identifying reasons 'why intelligence fails', these reviews offer lessons for improving future analysis, but this is not the same as gauging overall estimative accuracy.[11] In recent years, the ODNI has also sponsored political forecasting competitions. These competitions have solicited hundreds of thousands of predictions about geopolitical events from participants inside and outside the IC.[12] The results of these competitions provide valuable insight into the nature of 'good judgment', and they demonstrate how subjective assessments can be evaluated in structured ways. Yet assessing the accuracy of forecasts submitted to these competitions is different from assessing the accuracy of published intelligence estimates.[13]

The intelligence literature describes other, informal, mechanisms for judging the quality of analysts and teams. One such practice described by some scholars and practitioners is the so-called 'bean count': the volume of publications that an analyst or team has produced, perhaps implicitly weighted for length or perceived significance.[14] Such assessments are obviously problematic, but it is not clear that there are presently any better metrics to work with. For instance, there is the subjective standard of 'I know good work when I see it', but it is hard to know how much to trust people's

[10] See Fingar, *Reducing Uncertainty*, pp.109–11, 129–31 on ICD 203, related reforms, and how they do not directly address estimative accuracy. Tetlock and Mellers summarize the problem in stating that: 'The IC has tacitly placed a massive institutional bet on the validity of its home-grown theory of good judgment: namely, that accuracy should be a positive function of how well analysts conform to the process standards embodied in its performance-management guidelines … It would be flattering to the IC if its official theory were validated. But there is no guarantee that it is right or, if right, that it has been implemented effectively'. Tetlock and Mellers, 'Intelligence Management of Intelligence Agencies', p.8.

[11] On IC post-mortems, see Robert Jervis, *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War* (Ithaca, NY: Cornell University Press 2010).

[12] For results based on the judgment of respondents outside the IC, see Lyle Ungar et al., 'The Good Judgment Project: A Large Scale Test of Different Methods of Combining Expert Predictions', AAAI Technical Report FS-12-06 (2012).

[13] For example, as mentioned earlier, forecasts comprise only a subset of intelligence estimates. In addition, as we discuss later, published intelligence estimates are corporate products which undergo an extensive review process that may affect their accuracy (for better or for worse), relative to judgments solicited from individuals for these forecasting competitions. Neither of these points diminishes the value of these competitions: our point is simply that their results do not directly translate into inferences about estimative accuracy in the IC writ large.

[14] See Douglas J. MacEachin, *The Tradecraft of Analysis: Challenge and Change in the CIA* (Washington, DC: Consortium for the Study of Intelligence 1994) pp.4–6; Welton Chang, 'Getting it Right: Assessing the Intelligence Community's Analytic Performance', *American Intelligence Journal* 30/2 (2012) p.104; and Richard L. Russell, *Sharpening Strategic Intelligence: Why the CIA Gets It Wrong and What Needs to Be Done to Get It Right* (NY: Cambridge 2007) pp.124–5.

intuitions in this regard, and there is always a risk that managers and consumers will simply favor estimates that reinforce their preconceptions.

Perverse incentives – along with job dissatisfaction[15] – can emerge when analysts and teams are not evaluated on the accuracy of their work. The intelligence literature frequently describes how the current system encourages people to focus on quantity and not quality, while subjective performance assessment encourages analysts to write reports that please their superiors, and allows managers to favor personnel for reasons unrelated to performance. Meanwhile, rewarding analysts for contributions to the PDB may encourage personnel to seek assignments based on an issue's visibility, rather than on their knowledge of the subject matter.[16]

Certain analysts fare poorly in this system. Consider the analyst who generally does an excellent job, but has the misfortune to get a 'big one' wrong. This will presumably draw negative attention. Of course, even the best analysts can get unlucky. But since the IC does not track performance systematically, it is hard to know who deserves the benefit of the doubt. Systematically evaluating performance addresses this problem directly. Moreover, Section 4 shows how it is possible to assign and track the priority of each estimate before the fact so as to identify how well certain offices or analysts perform under pressure. This is another judgment that can be based on systematic evidence and not on selective examples.

Analysts who challenge prevailing views may also be disadvantaged by the status quo. Managers operating on tight deadlines may grow weary of analysts who sidetrack discussions by questioning basic assumptions. (The frustration can run in the opposite direction, too, as some managers make a point of requiring analysts to perform extensive alternative assessments.) These challenges may refine some estimates but, since accuracy is not systematically measured, personnel may simply prefer to complete more products and to avoid the organizational frictions that can accompany challenges to existing views.[17]

All of these incentives and disincentives apply to the performance of analysts, teams, and the IC writ large. Just as negative events can overshadow

[15] A 2010 report from the CIA Inspector General found that 51 per cent of employees who had left the Agency or who were considering leaving the Agency cited 'lack of appropriate recognition for my contributions' as a major factor. *Report of a Follow-Up Inspection: Retention at the Agency* (Central Intelligence Agency 2010) p.49.

[16] Perhaps the most scathing description of these and other problems is John A. Gentry, *Lost Promise: How CIA Analysis Misserves the Nation* (Lanham, MD: University Press of America 1993). See also Russell, *Sharpening Strategic Intelligence*, pp.119–48.

[17] One can even imagine giving analysts special credit for staking out positions that challenge prevailing viewpoints that ultimately prove to be correct; this is the most valuable kind of contribution that an analyst can make. Contrarianism and alternative assessment are core principles of intelligence analysis, yet, perversely, this can make it difficult to distinguish between well-considered, unorthodox viewpoints and arguments that challenge existing views simply for contrarianism's sake. The CIA Inspector General (*Report of a Follow-Up Inspection*, pp.17, 49) found that 'lack of management support for prudent risk-taking' was one of the most prevalent reasons for dissatisfaction among analysts.

an unlucky analyst's body of quality work, congressional committees are most likely to convene oversight hearings when intelligence fails. The Arab Spring provides a good example. Few people (including leaders in the region) anticipated that these uprisings would be so widespread or so consequential. The degree of difficulty in predicting this chain of events was extremely high. An inability to foresee this phenomenon may say more about the inherent unpredictability of Middle Eastern politics than about US intelligence capabilities in that region. Yet the Arab Spring predictably launched congressional hearings in 2011, with critical lawmakers framing the situation as yet another intelligence failure.[18] Here, the IC would have been well-served had it been able to demonstrate that it had consistently anticipated other regional developments over the years. Without tracking estimative accuracy systematically, however, the IC leaves itself exposed to being judged by developments which are salient because of their negative consequences, but which are not representative of intelligence capabilities overall. As Fingar observes: 'such criticism has a badly corrosive effect on the confidence in – and the confidence of – the analytic community'.[19]

## Section 2. The Feasibility of Evaluating Estimative Accuracy

The most common justification for why the IC has never attempted to evaluate estimative accuracy is that such evaluation would simply be impossible.[20] Yet this section explains not only that evaluating estimative accuracy is feasible, but also that it would require no significant changes to existing tradecraft.

Our emphasis on practical implementation is important because most existing ideas about evaluating estimative accuracy require that standard procedures be significantly changed. For instance, the IC and the U.S. Defense Department have experimented with using prediction markets to gauge uncertainty. Philip Tetlock and other researchers have demonstrated that it is possible to gauge the accuracy of political forecasts, but they generally require participants to state probabilities numerically, which would require widespread changes to existing practice.[21] To be sure, our arguments in this section and in Section 4 build on many of the same concepts that underpin

[18] Greg Miller, 'Senators Question Intelligence Agencies' Anticipation of Egypt Uprising', *Washington Post*, 3 February 2011, p.A17; Marcus Baram, 'CIA's Mideast Surprise Recalls History of Intelligence Failures', *Huffington Post*, 11 February 2011.

[19] Fingar, *Reducing Uncertainty*, p.35.

[20] To quote Sherman Kent, the chairman of the Board of National Estimates from 1952 to 1967 who is often called the founding father of intelligence estimation in the United States, the question of how accurate intelligence estimates are is 'almost impossible to answer'; or 'it cannot be answered in a way to satisfy an outside questioner', and, despite the flaws with existing analysis of estimative accuracy, 'we can do no better for the outsider'. Kent, 'The Law and Custom of the National Intelligence Estimate', pp.114, 127.

[21] Philip E. Tetlock, *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton, NJ: Princeton University 2005).

these previous efforts, but our explicit goal here is to leverage these concepts for evaluating traditional intelligence estimates.

Three concepts shape this discussion: *calibration*, *discrimination*, and *proper scoring rules*. Calibration captures how well estimated likelihoods compare to actual rates of occurrence. For instance, when estimates state that certain events are unlikely to occur, are those events actually unlikely to occur?[22] Discrimination measures how effectively analysts vary their assessments across cases, distinguishing, for example, high-probability outcomes from low-probability outcomes.[23] Proper scoring rules provide metrics for ranking performance that gives analysts the incentive to report their true beliefs, such that any attempts to 'game' the system would not increase average scores. Many scholars and practitioners argue that it is impossible to evaluate estimative accuracy in ways that elicit truthful reporting, but this and other objections are unfounded.

### *"Since estimates are probabilistic, we can never really say whether they are 'right' or 'wrong'"*

It is possible to draw a sound inference that something is unlikely, and for that outcome nevertheless to happen by chance. For instance, one should think it unlikely that a well-shuffled deck will have an ace on top. Occasionally an ace will appear, but it would not mean that the people who predicted otherwise were 'wrong'. Similarly an intelligence estimate can be right for the wrong reasons, and vice versa.

The solution, however, is not to give up on evaluating accuracy, but rather to examine the accuracy of many estimates together. A poker player who is poor at estimating odds will win on occasion, but over time his poor judgment will become obvious. The same is true with intelligence analysts and their predictions; the more estimates we examine, the better we can distinguish skill from luck.

For example, if we examine a range of estimates, we can observe how well-calibrated those assessments are. When estimates say that events are likely or unlikely to occur, or that certain hypotheses are likely or unlikely to be true, does that actually tend to be the case on balance? This question is fairly easy to answer, and the results could be important. A tendency towards overconfidence would appear in the data when judgments that analysts say are extremely likely (unlikely) prove correct much less (more) often than anticipated. Another common shortcoming is the tendency to misapply

[22] See Steven Rieber, 'Intelligence Analysis and Judgmental Calibration', *International Journal of Intelligence and CounterIntelligence* 17/1 (2004) pp.97–112.

[23] To see the difference between calibration and discrimination, imagine you are given a random list of names and asked to say if they belonged to men or to women. Your response would be well-calibrated if you assigned each one a 50 per cent probability of being female. But this would not represent good assessment, as one can predict the gender of people with names like 'John' or 'Elizabeth' with near certainty. Discrimination measures your ability to 'split' probabilities in this fashion rather than simply to report the base rate in each instance.

assessments of '50/50' or 'even chance' to issues about which analysts are simply unsure, although the odds are actually far from 50 per cent.[24] Such biases would be obvious in large samples, and the IC could then take steps to mitigate the problem.

Studying a large volume of estimates would also indicate how effectively those estimates discriminate between likely and unlikely possibilities. An analyst could be well-calibrated if she simply made the same prediction every time, as would a meteorologist who always predicted a 10 per cent chance of rain in a region where it rains every tenth day. Such judgments convey little information. A good analyst would use information in a way that would allow her to vary assessments (for instance, predicting a high chance of rain on a few days, and a low chance on most days). Ideally, we would like to know both how much assessments vary and how closely they correlate with actual probabilities. Once again, a single estimate may tell us little; but if we examine many estimates at once, the general reliability of judgments becomes evident.

### *"Every intelligence question is unique. Therefore, even if we wanted to evaluate accuracy across estimates, broad patterns are not meaningful"*

Some scholars are skeptical of evaluating large numbers of intelligence estimates together. One scholar writes:

> To have perfectly calibrated intelligence products, analysts would have to be able to say that, if a thing is sixty percent likely to happen, then it does happen sixty percent of the time. But most intelligence questions (beyond the trivial ones) are unique, one of a kind. The exact set of circumstances that led to the question being asked in the first place, and much of the information relevant to its likely outcome, are impossible to replicate, thereby making it difficult to keep score in a meaningful way.[25]

If the purpose of the exercise were to gain an objective sense of how likely a particular phenomenon actually is, then replication would indeed be crucial for deriving statistical estimates. Yet the goal here is to evaluate analysts and teams, not the probabilities of individual events.[26] Are intelligence reports typically overconfident, such that events deemed very likely only happen half

[24] Baruch Fischhoff and Wandi Bruine de Bruin, 'Fifty-Fifty = 50%?', *Journal of Behavioral Decisionmaking* 12/2 (1999) pp.149–63.

[25] Kristan J. Wheaton, 'Evaluating Intelligence: Answering Questions Asked and Not', *International Journal of Intelligence and CounterIntelligence* 22/4 (2009) p.619.

[26] We recognize that the composition of analytic teams often varies, as do the assignments given to individual analysts. These facts provide important context for evaluating the way that performance changes over time. But this is by no means unique to intelligence. Hospitals' medical staffs and sports teams' rosters change frequently, too; but that does not mean one cannot identify certain units, or indeed individuals, that consistently perform better than others.

the time? Do estimates stating that some events are as likely as not to transpire tend to be fair assessments of the odds? Do estimative statements discriminate effectively between high- and low-probability occurrences, or do they instead tend to give similar assessments in most cases?

These are empirical questions that can be answered by tracking estimates systematically. In fact, we must evaluate intelligence *because* analysts and teams usually deal with unique questions. If the IC dealt with the same questions over and over, then estimative skill would be irrelevant because we could estimate the probability of any given phenomenon directly from past results. It is because intelligence deals with unique issues that analysts must rely on judgment. Given this, it is important to know how reliable those judgments tend to be.

### *"Even if we wanted to measure estimative accuracy, there is no rigorous way to keep score"*

Some concepts in the intelligence literature provide poor scoring rules for evaluating estimative accuracy. One of the most common metaphors is the 'batting average', a concept intended to dampen expectations of what the IC should be able to achieve. In baseball, the best players rarely get hits more than 40 per cent of the time; in the Major Leagues, anything over 30 per cent is excellent. By drawing on this analogy, intelligence scholars hope to encourage holding intelligence estimates to a reasonable standard.[27]

The batting-average metaphor, whatever its rhetorical value, is a poor way to evaluate intelligence. It is easy to distinguish hits from outs in baseball, but no such dichotomy exists for intelligence estimation. For instance, if an event occurs, then a prediction that it was almost certain to happen is closer to the truth than a prediction that it was only likely to happen. The batting-average metaphor is ill-equipped to handle such issues, and attempts to stretch the concept can create perverse incentives.[28]

Some intelligence scholars note these problems with the batting-average concept and essentially stop there. But flaws in one metric do not imply that others are invalid. For instance, a Brier Score averages the squared difference between an analyst's predicted probabilities and actual outcomes, which are scored as 1 if the outcome occurs and 0 if it does not.[29] Thus if a report said that some outcome was 70 per cent likely and this judgment proved correct, the score for this estimate would be $(1 - 0.70)^2 = 0.09$. If the judgment proved false, then the score for this estimate would be $(0 - 0.70)^2 = 0.49$. Lower Brier Scores are superior, because they indicate that the analyst is assigning higher probabilities to statements that are true and lower probabilities to statements that are false. The Brier Score is a proper scoring

[27] Marrin, 'Evaluating the Quality of Intelligence Analysis', pp.902–4.

[28] Jeffrey A. Friedman and Richard Zeckhauser, 'Assessing Uncertainty in Intelligence', *Intelligence and National Security* 27/6 (2012) pp.827–8.

[29] In some cases, it is arguable whether or not some event occurred, and the 'true' outcome can thus be represented as a fraction. See Section 4.

rule because analysts wishing to achieve the best rating will adopt the strategy of reporting their genuine beliefs.

Another proper scoring rule rates analysts' assessments by the logarithm of the predicted probability for each true outcome. In this method, analysts provide probabilities for alternative scenarios. We then give each analyst a score that is the logarithm of her estimated probability for the scenario that ultimately occurred. Thus, if she assessed the probability of an observed event as being 40 per cent, her score for that assessment would be $\ln(0.40) = -0.92$. The higher the probability she assigned to that event, the better her score for that assessment would be. This scoring method also elicits honest reporting.

Although these methods induce truthful assessments, they share an important shortcoming: when analysts work on issues that have very different degrees of predictability, it is not appropriate to compare their performances directly using such scores. For instance, if you were asked to predict the chances that each in a series of coin flips would come up heads, then you could not do any better in expectation than saying heads was 50 per cent likely every toss; your average Brier Score would be 0.25. By contrast, if you were asked to predict the odds that the top card on a well-shuffled deck would be an ace, then you could not do any better in expectation than estimating this probability as being one in 13; over time your average Brier Score would converge to 0.077. In each case you are making the best possible judgment, but your Brier Score will be much better when predicting the more extreme odds associated with the top card being an ace. This example emphasizes that while analysts should always strive to improve estimative accuracy, there never will be a uniform standard for what 'acceptable' levels of accuracy entail.

Yet, one can certainly compare Brier Scores for analysts and teams that work on similar issues, and the challenge of comparing performance across groups is by no means unique to intelligence analysis.[30] When evaluating the quality of physicians, no one would compare patient survival rates for heart surgeons to those for orthopedists. Similarly, no one should compare teachers' abilities to prepare students for standardized tests without taking account of classroom size or students' educational backgrounds.

We do not advocate a one-size-fits-all approach for ranking analysts based on single statistics. (Indeed, one potential benefit of evaluating estimative accuracy in the manner we propose is that it may reveal systematic patterns in

[30] The problem could also be addressed in structured ways if the IC wished to do so. One approach worth considering would be to weight Brier Scores for individual estimates based on perceived 'base rates' for relevant outcomes. The idea would be to provide a rough guideline, ex ante, of how predictable certain events tend to be. Since events with probabilities closer to the extremes of 1 and 0 are more predictable, they lead to lower Brier Scores on average. It might make sense to employ independent rating analysts to assess categories of predictability (such as 'balanced', 'strong tendency', or 'extreme tendency'). Such a categorization would be relatively crude, thereby limiting the effort required by the rating analyst. Nevertheless, it could provide a rough standard for judging or weighting Brier Scores.

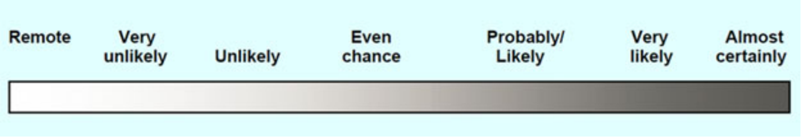

**Figure 1.** Words of Estimative Probability.
*Source*: Graphic displayed in the 2007 National Intelligence Estimate, *Iran: Nuclear Intentions and Capabilities*, as well as in the front matter of other recent intelligence products.

what kinds of topics are easier or harder to estimate than others.) Our goal is simply to show that it is possible to gather information that will allow the IC to understand and improve its performance more rigorously than it does now. Brier Scores and logarithmic payoffs are two valid tools for measuring performance, and both provide incentives for honest reporting.

### *"These methods all require analysts to state probabilities quantitatively"*[31]

Having analysts state explicit probabilities would certainly make it easier to evaluate estimative intelligence. It would also make estimates clearer. But analysts need not quantify probabilities in published work in order for the IC to evaluate probabilistic statements.

In recent years, the IC has defined specific 'Words of Estimative Probability' (WEPs) for intelligence officials to use when conveying assessments of likelihood. Figure 1 shows how these WEPs are presented in the front matter of a recent NIE. It is an open (and potentially very informative) question as to how effectively intelligence estimates employ these terms.

For example, do events described as 'unlikely' actually occur significantly less often than events predicted to have an 'even chance'? Are possibilities judged to be 'very likely' noticeably more frequent than those that are seen to be 'likely'? The IC can begin answering these questions immediately in order to determine which estimates do a better job of discriminating relative likelihoods. Figures 2a and 2b present hypothetical examples of what one might observe by dividing up estimative statements according to which WEPs were employed, and observing how frequently the estimated scenarios actually occurred. It is obvious that the distribution in Figure 2b indicates better estimative performance. The percentages of actual occurrences of

[31] For example, one scholar recently wrote that: 'Words such as "probably", "likely", and "may" are scattered throughout intelligence publications and prevent easy assessment of accuracy. For example, if CIA analysts had said Iraq probably had weapons of mass destruction, was that analysis accurate or inaccurate? There is no way to tell, given the use of the world "probably" which qualified the statement to incorporate the analysts' uncertainty'. Marrin, 'Evaluating the Quality of Intelligence Analysis', p.898.

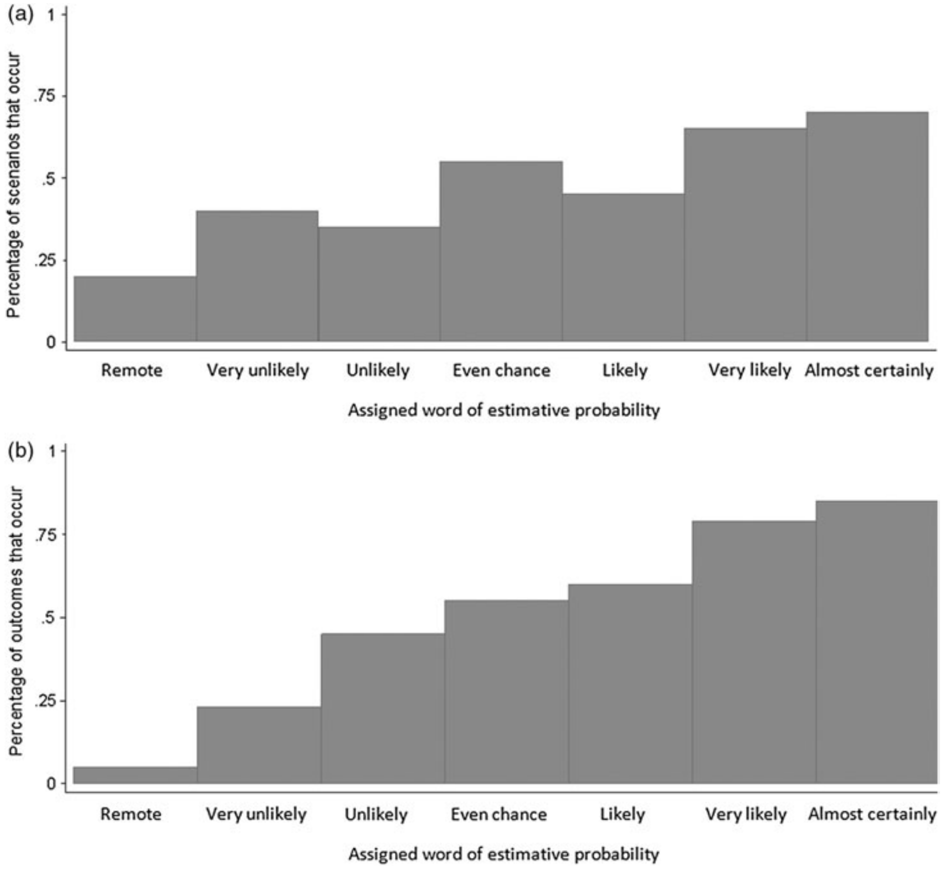

**Figure 2.** (a) An Example of Mediocre Estimation and (b) An Example of More Informative Estimation.

estimated scenarios corresponding to each WEP ascend in the appropriate order, in reasonably even fashion, with the percentage of occurrences of events that had been judged 'remote' being very small and the percentage of occurrences of events that had been judged 'almost certain' being very large. By contrast, the distribution in Figure 2a is noticeably less calibrated and less discriminating. Simply by graphing this information, an analyst or team would see several areas for improvement, and decision makers would have a better idea of what the estimates mean.

It is possible to gather more specific information. For example, the seven WEPs displayed in Figure 1 are evenly spaced on a spectrum from 0 per cent to 100 per cent. One could, therefore, interpret each WEP as spanning roughly 14 percentage points. In order to rate the accuracy of these estimates, it might be reasonable to quantify each WEP as implying the middle of its respective range. The term 'Very Unlikely' thus implies a probability of about

21 per cent, the term 'Probably/Likely' implies a probability of about 65 per cent, and so on.[32] This is obviously far from the ideal way to capture probabilities; more precise inputs would lead to more precise outputs. Our point is simply that if intelligence practitioners wished to continue with existing tradecraft, one could still evaluate the accuracy of their products.

One recent study used a clever approach to generate more precise estimates of probability.[33] Subjects in this study were given estimates, and each time they encountered a qualitative statement of likelihood, they were asked to decide what they believed the statement meant numerically. Several raters were assigned to each estimate; their inferred probabilities were averaged and then compared to observed outcomes. This approach is too labor-intensive to apply broadly, but it reinforces the point that existing tradecraft does not preclude rigorous evaluation of estimative accuracy.[34]

This issue would vanish, of course, if analysts expressed probabilities numerically. Numerical estimates have traditionally been opposed on the grounds that decision makers would misinterpret such estimates as being overly scientific.[35] Yet, even if we take this objection at face value, analysts could keep their reports unchanged and quantify estimative language simply for internal use (as with the IC's forecasting competitions mentioned above). Section 4 explains what this approach might look like in practice, and why it would not burden the estimative process; analysts and teams could always simply report the midpoint of the range corresponding to the relevant WEP if they saw no reason to shift their estimates in either direction.

### *"Evaluating estimative accuracy would be too expensive"*

Given the magnitude of the stakes in intelligence estimation, even measures that marginally improve US intelligence capabilities and the public debate surrounding them would merit a substantial outlay. Moreover, the cost critique is usually made by people who have attempted to evaluate estimative intelligence as individual scholars, and not as part of an institutional effort.

[32] In principle, one might interpret each WEP as conveying the expected probability of an event conditional on the event's falling within each bin. If probabilities were distributed normally around 50 per cent, for instance, this would mean that those expected probabilities would fall towards the high end of each bin on the left side of the spectrum and towards the low end of each bin on the right side of the spectrum.

[33] Paul Lehner et al., 'Using Inferred Probabilities to Measure the Accuracy of Imprecise Forecasts' (MITRE 2012). For related approaches, see Frederick Mosteller and Cleo Youtz, 'Quantifying Probabilistic Expressions', *Statistical Science* 5/1 (1990) pp.2–34.

[34] See also David R. Mandel and Alan Barnes, 'Accuracy of Forecasts in Strategic Intelligence', *Proceedings of the National Academy of Sciences* 111/30 (2014) pp.10984–9. Mandel and Barnes encoded verbal statements of likelihood into one of nine possible numbers, calibrated the results, and drew lessons for improving the accuracy of estimative intelligence.

[35] For a critical discussion of arguments against using numeric probabilities in intelligence estimation, see Friedman and Zeckhauser, 'Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden', *Intelligence and National Security*, forthcoming.

Intelligence official Abbot Smith, for instance, wrote: 'I am sure that if one were to try and work out an accuracy score covering the product of nearly twenty years he would have to scan not less than 25,000 judgments, and probably far more'.[36] Smith may have estimated the scale of the task correctly, but why should it be a one-person job? If just four people were assigned to gauge estimative accuracy, and they each evaluated just one judgment per day, they would stay ahead of the workflow Smith described. In Section 4, we suggest a system of delegating most reporting requirements of an evaluation system, which would further ensure that such a system could keep pace while requiring minimal bureaucratic overhead.

## Section 3. The Benefits of Evaluating Estimative Accuracy

Even if it were generally accepted that evaluating estimative accuracy is feasible, such assessments would still face considerable skepticism. As sociologist Rob Johnston writes, there has been 'long-standing bureaucratic resistance to putting in place a systematic program for improving analytical performance'. The underlying reason for this pattern, Johnston argues, is that, 'simply put, a program explicitly designed to improve human performance implies that human performance needs improving, an allegation that risks considerable political and institutional resistance'.[37] It is important to note that discomfort with evaluating professional judgment is widespread and by no means idiosyncratic to the IC.[38]

Yet while this cultural resistance must be acknowledged, it should not be confused with the widespread notion that evaluating estimative accuracy would politically or professionally harm intelligence personnel. Our goal in this section is to disentangle these arguments. As we detail below, many analysts and managers would actually benefit from having estimative accuracy evaluated. Establishing such a system would help to address several contemporary political and institutional problems that the IC faces.

One constituency that would directly benefit from evaluating estimative accuracy includes managers, trainers, analytic methodologists, and anyone else whose goal is to improve the IC's performance. Most suggestions for improving analytic performance offer marginal improvements which are important to capture but difficult to recognize. For example, the intelligence literature offers a range of 'structured analytic techniques' for improving

[36] Abbot E. Smith, 'On the Accuracy of National Intelligence Estimates', *Studies in Intelligence* 13/3 (1969) p.25.

[37] Rob Johnston, *Analytic Culture in the US Intelligence Community* (Washington, DC: Center for the Study of Intelligence 2005) p.xvi.

[38] On this issue in general, see Paul E. Meehl, 'Causes and Effects of My Disturbing Little Book', *Journal of Personality Assessment* 50/3 (1986) pp.370–5; and Robyn M. Dawes, 'The Robust Beauty of Improper Linear Models in Decision Making', *American Psychologist* 34/7 (1979) pp.571–82.

performance.[39] All of them have strengths and weaknesses (including varying levels of complexity and requirements of time), and it is often unclear how these costs and benefits would balance out. Mark Lowenthal observes that 'No one has yet come up with any methodologies, machines or thought processes that will appreciably raise the Intelligence Community's [performance]'.[40] Yet this says more about the IC's ability to recognize gains than it does about the ability of structured analytic techniques to produce them. If accuracy were measured systematically, it would indeed be feasible to ascertain which measures produce more accurate estimates than others.

Similar questions emerge when evaluating structural elements of the intelligence process. For example, some intelligence estimates undergo an extensive review process. That process can be lengthy, and it is often criticized as being a forum for bureaucratic politics, while potentially exposing estimates to the dangers of 'groupthink'. Many analysts see review as inefficient, even damaging.[41] But is that true? If estimative accuracy were systematically measured, it would then be straightforward to determine how the review process correlates with estimative accuracy. And if some forms of review systematically improve performance, they would then be easier to justify to skeptics.

Absent systematic evaluation, however, it is virtually impossible to make confident diagnoses and prescriptions relating to review or to any other part of the intelligence process. Of course, the benefits of fine-tuning structure and tradecraft may prove modest, which is precisely why it is so difficult to identify such benefits through casual trial and error. But measures that modestly improve decision makers' abilities to protect the nation are worth substantial investment.

Moreover, certain benefits of evaluating estimative accuracy are bound to emerge organically. For several decades, decision theorists have trained people in assessing uncertainty. One of the most consistent findings in this field is that few people estimate probabilities effectively without such training.[42] Untrained estimates tend to be swayed by factors such

[39] See, for instance, Richards J. Heuer, Jr., and Randolph H. Pherson, *Structured Analytic Techniques for Intelligence Analysis* (Washington, DC: CQ 2011).

[40] Mark M. Lowenthal, 'Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues?', *Intelligence and National Security* 23/3 (2008) p.314. Cf. Fingar, *Reducing Uncertainty*, pp.34, 130: 'By and large, analysts do not have an empirical basis for using or eschewing particular methods'.

[41] As CIA official Martin Petersen describes, critics of the review process 'present a bill of particulars that alleges the process does little to improve the product, reduces judgments to the lowest common denominator, stifles creativity, and takes analysis out of the hands of experts. Those who defend the review process counter that it sharpens focus, guarantees that the piece addresses policymaker concerns, taps all relevant expertise, and ensures a corporate product'. Martin Petersen, 'Making the Analytic Review Process Work', *Studies in Intelligence* 49/1 (2005). Available at https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol49no1, accessed November 11, 2014.

[42] In international politics specifically, see Tetlock, *Expert Political Judgment*.

as how easily scenarios come to mind or how dreadful their consequences might be.[43] People tend to be overconfident in their assessments, assigning far too little weight to scenarios they deem unlikely, quickly overriding large amounts of objective information based on misleading 'gut instincts', and failing to consider more than a subset of relevant possibilities.[44]

Another common finding is that structured feedback often promotes substantial improvement. Most people simply have a poor intuitive sense of their strengths and weaknesses when assessing uncertainty. Uncertainty is abstract, and a person's skill at dealing with the problem can only be judged reliably when looking at a volume of data. Most people process unstructured feedback selectively in a self-deluding manner, one that supports an illusion of skill. Giving people structured feedback (even simple feedback along the lines of Figure 2) dispels such illusions, usually for the better.[45]

The impact of performance evaluation would fall unevenly on some groups. For instance, evaluating the accuracy of intelligence estimates will make clear how analysts and teams rank on this dimension. This should help high performers. To the extent that all analysts and teams can benefit from incorporating systematic feedback, this should also raise the performance of analysts overall.

Of course, some analysts and teams will perform below the mean. While it seems unassailable that personnel should be rewarded based on performance, this can obviously be overdone, and any system of performance incentives should be carefully designed. Yet the challenge of designing fair performance incentives is common to all large organizations. Hospitals and health plans, for instance, face tough questions when deciding how to handle surgeons with high error rates or general practitioners with poor track records in diagnosis, but no one would suggest ignoring these issues. Moreover, as mentioned in Section 1, the IC already deals with tricky questions relating to performance reviews; existing methods for dealing with these questions are already problematic. Evaluating estimative accuracy does not create the challenge of incentivizing good performance. It simply allows the IC to confront the issue with more relevant information.

Almost all decision makers would benefit from having a clearer sense of the IC's capabilities. Decision makers weighing the value of a given estimate

[43] Daniel Kahneman, Paul Slovic, and Amos Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases* (NY: Cambridge 1982); Heuer, *Psychology of Intelligence Analysis* (Washington, DC: Center for Intelligence Analysis 1999); and Paul Slovic (ed.), *The Perception of Risk* (Sterling, VA: Earthscan 2000).

[44] Paul E. Meehl, *Clinical Versus Statistical Prediction* (Minneapolis, MN: Minnesota 1954); Marc Alpert and Howard Raiffa, 'An Interim Report on the Training of Probability Assessors', in Kahneman, Slovic, and Tversky, *Judgment Under Uncertainty*; and Robyn M. Dawes, *Rational Choice in an Uncertain World* (NY: Harcourt Brace 1988).

[45] See the references in the previous note, as well as; Sarah Lichtenstein and Baruch Fischhoff, 'Training for Calibration', *Organizational Behavior and Human Performance* 26/2 (1980) pp.149–71; Daniel Kahneman, *Thinking, Fast and Slow* (NY: FSG 2011) pp.209–21.

would benefit from understanding the IC's general estimative accuracy, and how its performance might vary across topics or regions. Anecdotal evidence suggests that decision makers often hold strong views on these subjects on the basis of limited evidence.[46] Meanwhile, certain briefers and officials can develop reputations for being overly cynical, cautious, or optimistic in their advice. This effectively requires decision makers to correct for perceived biases one way or another, but without access to systematic information.[47] Decision makers could better utilize the information at their disposal by replacing such heuristics with well-founded assessments of analytic performance.

Finally, evaluating estimative accuracy could very well elevate the public's assessment of the IC as a whole. While we cannot say how well the IC will fare in systematic evaluation before the fact, intelligence scholars and practitioners have long lamented that perceptions of the IC are tarnished by selective attention to major mistakes. As Kristan Wheaton describes it: 'Imagine a football coach who watched the game film only when the team lost and ignored lessons from when the team won. Doing so is clearly stupid, yet it is very close to what happens to the IC every five to ten years. From the Hoover Commission to today, "intelligence failures" are investigated while intelligence successes are largely overlooked or ignored'.[48]

Examining the full spectrum of intelligence products (and not just outliers at one end of that spectrum) may well dampen such biased perceptions of the IC. At the very least, judging intelligence performance with systematic assessments would diminish the superficiality of debates on this subject.[49] Moreover, once something can be measured, it becomes possible to demonstrate and reinforce improvement. As analysts, offices, and agencies better identify their strengths and weaknesses and adapt accordingly, it should be possible to show that these adaptations make a difference. In the meantime, however, it is hard to accept the idea that the IC's performance is underappreciated *and* to say that its performance does not merit systematic assessment.

[46] In one of the earliest articles on estimative accuracy, Abbot Smith wrote that: 'The hard fact of life is that the high-level consumer of [National Intelligence Estimates or NIEs] … is apt to judge the whole output on the basis of two or three estimates which strike home to him. If they prove correct, NIEs are good; if incorrect, they are bad'. Smith, 'On the Accuracy of National Intelligence Estimates', p.26.

[47] For example, analysts who deal with strategic warning have incentives to exaggerate the probability of certain threats in order to gain policymakers' attention. Interpreting a warning estimate may thus implicitly involve downgrading threat assessments. But how much correction is appropriate? Systematically measuring the accuracy of these assessments would provide a structured baseline for addressing this issue, while mitigating analysts' incentives to exaggerate in the first place.

[48] Wheaton, 'Evaluating Intelligence', p.629.

[49] As Tetlock and Mellers describe, such 'superficial accountability demands' have predictably negative effects on organizations, producing 'shell-shocked, blame-averse organizational culture that tries to shield itself from a capricious environment' rather than seeking genuine improvement. Tetlock and Mellers, 'Intelligent Management of Intelligence Agencies', p. 543.

## Section 4. Components of Evaluating Estimative Accuracy

As we have described in previous sections, estimative statements predict the future (for example, the prospective result of Syria's civil war), or make uncertain judgments about existing affairs (for example, the current status of Iran's nuclear program). Such statements define possible states of the world and indicate their likelihoods. Evaluating these statements requires tracking the alternative scenarios being estimated, the estimated likelihood for each scenario, and whether those scenarios actually transpire. This section describes what a system for recording this information might entail.

### *Tracking Alternative Scenarios*

The first step in evaluating estimative accuracy is to track the alternative scenarios that analysts and teams are estimating. The easiest way to do this would be for analysts to log key estimative statements into a central database any time they submit a report. Decentralizing the process would help to ensure that evaluation kept pace with workflow while minimizing bureaucracy. This task could also be performed at higher levels; teams and managers could systematically track their estimates in the same way we suggest for individual analysts. Additional methods for cataloguing estimative statements would be for managers and agencies to collate reports they disseminate in a more centralized manner, or for the ODNI to select a subset of estimative products for ongoing evaluation.

Cataloguing estimative statements should not be onerous. Making clear to consumers exactly what is being estimated is already a core principle of analytic tradecraft.[50] Many reports already open with a section summarizing 'Key Judgments'. Tracking these statements for purposes of evaluation would essentially involve taking what is already highlighted and recording it for future reference. This would only require a more than modest effort when estimative statements are not already clearly laid out in a report. In those cases, requiring authors to enumerate their conclusions would be a useful way to promote clarity.

Recording estimative statements would produce two additional benefits. First, the system we propose would make it possible to assess how well the IC anticipates strategic surprises. Decision theorists distinguish between *uncertainty*, where people do not know what probabilities to attach to certain states of the world, and *ignorance*, where people cannot even define what some of the relevant states of the world might be. Similarly in intelligence, analysts inevitably anticipate some possibilities but not others; it is important to know how extensive these gaps might be. As mentioned in Section 1, anticipating strategic surprise is one of the IC's principal functions. If estimative statements were catalogued as we recommend, it would be possible to see whether the IC had a

[50] See ICD 208, 'Write for Maximum Utility' (December 2008).

particular event 'on its radar' before it occurred, even if that event had been judged unlikely.

Second, the system we propose would make it possible to evaluate the timeliness of intelligence reporting. Systematically recording estimative statements would make it clear not just *whether* intelligence officials consider certain scenarios, but also *when* they do so, and how far in advance they are able to anticipate significant developments. Measuring this attribute properly requires theoretical attention in its own right. For our purposes here, the point is simply that by gathering the information necessary to evaluate estimative accuracy, it would also become feasible to address two other important components of intelligence performance.

### *Tracking Estimative Probabilities*

To evaluate estimative accuracy, analysts must record the probabilities they assign to each scenario. As mentioned in Section 2, the process is simplest when probabilities are expressed numerically, though this does not mean that published estimates must use quantitative language themselves. There are three main options when an estimative statement invokes one of the so-called WEPs discussed above.

First, the IC could establish rules of thumb for interpreting what WEPs mean quantitatively. For example, consider a statement such as 'Bashar al-Assad will likely be deposed from power within two years'. According to Figure 1, the term 'likely' corresponds to an estimate of perhaps 65 per cent.[51] A meta-analysis averaging the results of 20 studies found that people typically interpret the word 'likely' to mean about 69 per cent.[52] For the purpose of evaluating estimates, these or any other figures could be designated the baseline interpretation of the word 'likely'.

Second, analysts could record their own assessments of how the qualitative language they use maps onto a numerical spectrum. If the IC adopted our suggestion of defining a baseline interpretation for each WEP, that could be the default choice. Analysts would retain the freedom to change that interpretation to anything they felt more appropriate for the purposes of evaluation.

A third approach would simply keep estimative probabilities qualitative throughout. This would sacrifice precision, but it would not preclude assessing calibration and discrimination. For instance, it would be very useful to know how often estimated outcomes occur which were initially considered 'remote', or 'likely', or 'almost certain', and so on. Existing doctrine already orders these terms on the probability spectrum. How well do they actually correlate with the probabilities of observed outcomes? Do some analysts or

[51] If the seven WEPs are spaced evenly along the probability spectrum such that each is roughly 14 percentage points wide, then the term 'Probably/Likely' essentially covers anything between 58 and 72 per cent, and the midpoint of this range is 65 per cent (though see Note 32 on why one might prefer a different interpretation).

[52] Mosteller and Youtz, 'Quantifying Probabilistic Expressions', p. 4.

teams consistently excel at using such terms in a fashion that discriminates among likely and unlikely events? Even if the IC resists assigning numerical probabilities for the purposes of evaluation, it is still desirable and feasible to answer these questions. (See Figure 2 examples.)

### *Tracking Outcomes*

The third step in evaluating estimative accuracy is to identify at the specified time whether each enumerated state of the world has occurred. We readily acknowledge that some estimative statements can never be evaluated, even after the fact. For instance, some intelligence estimates deal with leaders' preferences and mindsets, and such statements can rarely be verified.[53] But that problem does not affect the utility of examining estimative accuracy in those places where possible, and it is by no means unique to intelligence analysis. Baseball teams, for instance, generally struggle to evaluate players' defensive skills. This hardly reduces the importance of measuring their value on offense, which is easier to measure. It is hard for hospitals to track many subjective aspects of patient care, yet they can still record survival rates, error rates, cost-per-procedure, or tests ordered per patient in a particular category. In almost any profession, it is important to measure performance whenever one can, with the caution that such information should not be treated as being more definitive than it really is.

Recording outcomes could again be delegated to individual analysts, or to other levels of the IC.[54] This process is simplest when outcomes are recorded in a binary fashion: '1' if the state of the world came to pass or '0' if it did not. It is only possible to render this kind of judgment if the estimative statements being graded are worded clearly. For instance, if an analyst writes that 'Bashar al-Assad will likely be deposed from power soon', reasonable people could disagree about whether this prediction proved true if he were exiled from Syria two years after the statement was made. If the analyst instead wrote that 'Bashar al-Assad will likely be deposed from power by 31 December 2015', then one could evaluate the statement unambiguously.

Lacking access to a representative sample of current intelligence estimates, we cannot say what fraction of estimative statements describe states of the world in sufficiently precise terms to be evaluated in this way.[55] However, it is

[53] The IC could also consider scoring such outcomes probabilistically. For instance, if evaluators later believed that it was just as likely as not that an estimated outcome had occurred, then the outcome could be entered as '0.50' for the purposes of calculating Brier Scores or other metrics.

[54] One way to keep the workflow manageable would be to date-stamp estimative statements when they go into a database such that, at the appointed time, analysts (or managers, evaluators, etc.) would be prompted to indicate whether or not relevant outcomes had been observed.

[55] However, this problem need not preclude systematic assessments. For example, Mandel and Barnes, 'Accuracy of Forecasts in Strategic Intelligence', were able to calibrate 1943 of the 2897 forecasts in their sample (67 per cent).

important to keep in mind that the issue of specifying a date certain is only necessary for predictions; estimative judgments of current or past states of the world do not require this information. And to the extent that some estimative statements are currently too vague to be judged after the fact, it is also hard to imagine that they could provide ideal information to decision makers *before* the fact. In both instances, greater precision is better.

### *Additional Information and Conclusion*

We have specified all of the information that is required to establish straightforward evaluations of estimative accuracy, including proper scoring rules and basic measures of calibration and discrimination of judgments. Gathering additional indicators could then allow evaluations to go further, in order to examine which subsets of estimates tend to be more accurate than others. That information facilitates interpreting performance in the proper context, and would foster improvement in that performance where possible.

For example, the IC might wish to rate certain estimative statements by level of priority or extent of analytic review.[56] Designating such measures even in a rough fashion could assist in determining whether certain analysts, teams, or organizations perform better or worse on the most important issues, and whether peer review actually improves estimative accuracy. (One could also use these designations to weight certain estimative statements more heavily when calculating scoring rules, in much the same way that home runs count for more than singles when computing a baseball player's slugging percentage.) Additional measures that might be worth gathering include information about the regions and topics being studied (to see whether the IC performs better on some issues than on others), and the time frames of the estimates (to see whether and how swiftly analytic accuracy diminishes the further out analysts are asked to predict the future). These additional measures could be coded into the system by analysts, or added by administrators, depending on what variables the IC found most important to gather and how it preferred to spread the effort.

Many additional possibilities merit examination, depending on which aspects of analytic performance the IC and its consumers would like to see addressed. Generally speaking, there are many interesting and important questions that the IC could address rigorously once some semblance of the system we describe has been put in place. As we have shown, gathering this information would require little change to existing tradecraft, and the process

[56] The IC already has guidelines for establishing the priority of different intelligence products in accordance with a matrix updated by the DNI. See Intelligence Community Directive 204 (September 2007), which describes the National Intelligence Priorities Framework (NIPF). In addition to the information provided by the NIPF, managers could grade the priority of specific estimative assignments on an ad hoc basis. Since clearly conveying priorities is a common principle of good management, this is another place where gathering information for evaluating estimative accuracy would help to reinforce the estimative process itself.

could be decentralized so that a simple evaluation system could be established quickly and with modest bureaucratic overhead.

In closing, there is an important role that scholars can play in getting this process off the ground. Those of us outside the IC lack access to the 'raw data' required to evaluate estimative accuracy directly. But many scholars are familiar with the concepts and ideas underpinning such an effort. As Section 2 showed, the main obstacles to evaluating estimative accuracy in the IC today are issues neither of logistics nor of technical expertise. They are predominantly problems of theory: the building of a conceptual foundation for a system of evaluating intelligence and the dispelling of common misconceptions that this cannot be done. Scholars have analytic tools that, if well-articulated and properly applied, could make a real difference in evaluating, managing, improving, and understanding intelligence analysis. At the moment, however, the United States, its leaders, and the IC have no systematic way to assess estimative accuracy, even though it would be feasible and desirable to do so.

## Acknowledgements

For helpful comments on this manuscript, the authors thank three anonymous reviewers, Welton Chang, Seth Goldstein, Peter Mattis, Joseph Nye, Stefan Trautmann, Gregory Treverton, and several Intelligence Community officials who offered critical insight

## Notes on Contributors

Jeffrey A. Friedman is an Assistant Professor of Government at Dartmouth College, USA. He conducted research for this article while he was a postdoctoral fellow at the Dickey Center for International Understanding.

Richard Zeckhauser is the Frank P. Ramsey Professor of Political Economy at the Harvard Kennedy School, USA. He is the author of more than 280 professional articles and a dozen books, focusing mainly on problems across policy fields where risk, uncertainty, and ignorance play a role.