The Value of Precision in Probability Assessment:
Evidence from a Large-Scale Geopolitical Forecasting Tournament

Jeffrey A. Friedman, *Assistant Professor of Government, Dartmouth College*
Joshua D. Baker, *Ph.D. Candidate in Psychology & Marketing, University of Pennsylvania*
Barbara A. Mellers, *I. George Heyman University Professor, University of Pennsylvania*
Philip E. Tetlock, *Leonore Annenberg University Professor, University of Pennsylvania*
Richard Zeckhauser, *Frank P. Ramsey Professor of Political Economy, Harvard University*

Forthcoming in *International Studies Quarterly*

Abstract. This article employs a unique data set containing 888,328 geopolitical forecasts to examine the extent to which analytic precision improves the predictive value of foreign policy analysis. Scholars, practitioners, and pundits often prefer to leave their assessments of uncertainty vague when debating foreign policy, on the grounds that clearer probability estimates would provide arbitrary detail instead of useful insight. However, we find that coarsening numeric probability assessments in a manner consistent with common qualitative expressions – including expressions currently recommended for use by intelligence analysts – consistently sacrifices predictive accuracy. This result does not depend on extreme probability estimates, short time horizons, particular scoring rules, or individual-level attributes that are difficult to cultivate. At a practical level, our analysis indicates that it would be possible to make foreign policy discourse more informative by supplementing natural language-based descriptions of uncertainty with quantitative probability estimates. Most broadly, our findings advance long-standing debates over the limits of subjective judgment when assessing social phenomena, showing how explicit probability assessments are empirically justifiable even in domains featuring as much complexity as world politics.

Acknowledgments. Thanks to Pavel Atanasov, Michael Beckley, William Boettcher, David Budescu, Michael Cobb, Shrinidhi Kowshika Lakshmikanth, David Mandel, Angela Minster, Brendan Nyhan, Michael Poznansky, Jonah Schulhofer-Wohl, Sarah Stroup, Lyle Ungar, Thomas Wallsten and Justin Wolfers for valuable input on previous drafts. This work benefited from presentations at Dartmouth College, Middlebury College, the University of Pennsylvania, the University of Virginia, the 2015 meeting of the American Political Science Association, the 2015 ISSS-ISAC joint annual conference, and the 2015 National Bureau of Economic Research Summer Institute. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) Contract No. D11PC20061. The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. government.

The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament

Before John F. Kennedy authorized the Bay of Pigs invasion in 1961, he asked the Joint Chiefs of Staff to evaluate the plan. The Joint Chiefs found it unlikely that a group of Cuban exiles could topple Fidel Castro's government. Internally, they agreed that this probability was about thirty percent. But when the Joint Chiefs conveyed this view to the president in writing, they stated only that "This plan has a fair chance of success." Though the report's author, Brigadier General David Gray, claimed that "We thought other people would think that 'a fair chance' would mean 'not too good,'" President Kennedy interpreted this phrase as indicating favorable odds. Afterwards, Gray believed that his vague language had enabled a strategic blunder, while Kennedy resented the fact that his military advisers did not offer a clearer expression of doubt (Wyden 1979, 88-90).[1]

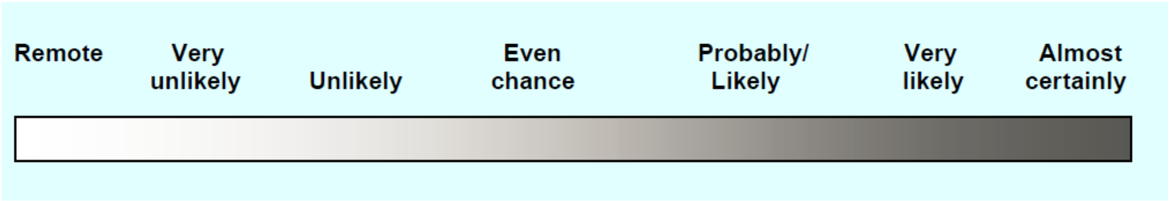
This aversion to clear probabilistic reasoning is common and deliberate in foreign policy analysis (Lanir and Kahneman 2006; Dhami 2013; Marchio 2014; Barnes 2016). Figure 1, for example, displays recent U.S. Intelligence Community guidelines for expressing uncertainty, which encourage analysts to communicate probability using qualitative phrases. U.S. military doctrine instructs planners to identify courses of action that minimize risk and that offer the highest chances of success, but not necessarily to identify what those risks and chances are.[2] From 2003 to 2011, the U.S. Department of Homeland Security communicated the probability of terrorism to the public using a vague, color-coded scale (Shapiro and Cohen 2007; McDermott and Zimbardo 2007). Many scholars and pundits are just as reluctant to describe the uncertainty surrounding their judgments when debating foreign policy in the public sphere. Quite often, phrases like "a fair

[1] Wyden writes that "in 1977, General Gray was still severely troubled about his failure to have insisted that figures be used. He felt that one of the key misunderstandings in the entire project was the misinterpretation of the word 'fair' as used by the Joint Chiefs."

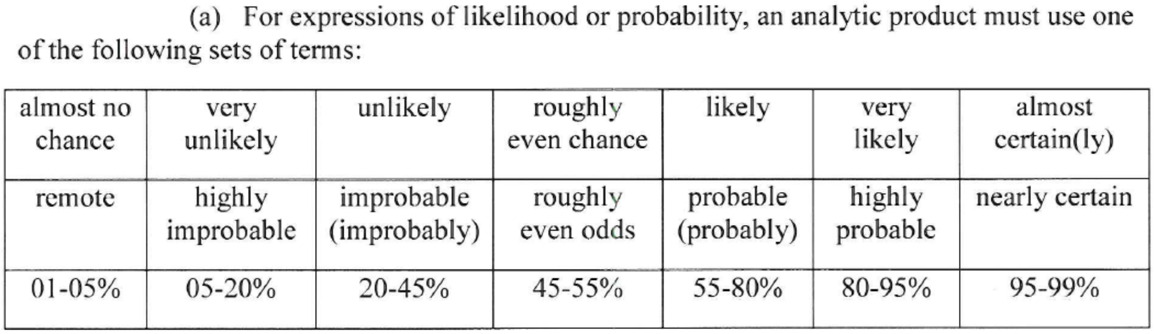
[2] See, for example, U.S. Army (2009, 2-19, B-173); U.S. Army (1997, 5-24); U.S. Joint Forces Command (2006, 3-14).

Figure 1. Guidelines for communicating probability assessments in the U.S. Intelligence Community

In the National Intelligence Estimate, "Iran: Nuclear Intentions and Capabilities" (November 2007)



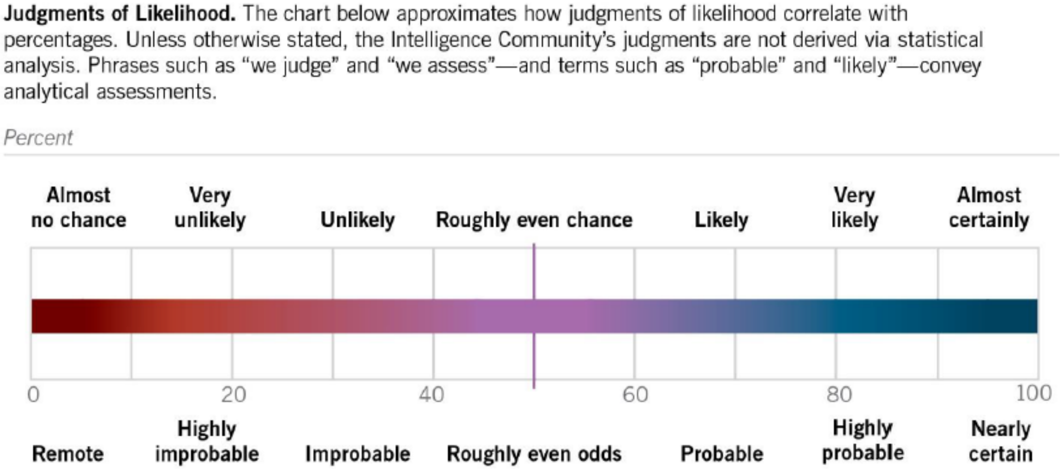
Director of National Intelligence Guidelines: Intelligence Community Directive 203, "Analytic Standards" (January 2015)



(a) For expressions of likelihood or probability, an analytic product must use one of the following sets of terms:

almost no chance	very unlikely	unlikely	roughly even chance	likely	very likely	almost certain(ly)
remote	highly improbable	improbable (improbably)	roughly even odds	probable (probably)	highly probable	nearly certain
01-05%	05-20%	20-45%	45-55%	55-80%	80-95%	95-99%

In the Intelligence Community Assessment, "Assessing Russian Activities and Intentions in Recent US Elections" (2017)



Judgments of Likelihood. The chart below approximates how judgments of likelihood correlate with percentages. Unless otherwise stated, the Intelligence Community's judgments are not derived via statistical analysis. Phrases such as "we judge" and "we assess"—and terms such as "probable" and "likely"—convey analytical assessments.

chance of success" would be more precise than the arguments that policy advocates use to justify placing lives and resources at risk (Tetlock 2009; Gardner 2011).

Foreign policy analysts typically defend these practices by arguing that world politics is too complex to permit assessing uncertainty with meaningful precision.[3] In this view, clearer probability estimates convey arbitrary detail instead of useful insight.[4] Some scholars and practitioners even argue that explicit assessments of uncertainty are counterproductive, imparting illusions of rigor to subjective judgments, enabling analysts' natural tendencies towards overconfidence, or otherwise degrading the quality of foreign policy analysis.[5] The notion that foreign policy analysts should avoid assessing subjective probabilities holds implications for writing any intelligence report, presenting any military plan, or debating any major foreign policy issue. Yet it is ultimately an empirical question as to whether making these judgments more precise would also make them more accurate, and that hypothesis has never been tested directly.

[3] For scholarship on complexity in world politics, see Beyerchen (1992/93), Jervis (1997), Pape (1997/98), Betts (2000). On the connection between complexity theory and debates about strategic assessment, see Connable (2012). Current Secretary of Defense James Mattis (2008, 18-19) famously drew on complexity theory in suggesting that "it is not scientifically possible to predict the outcome of an action. To suggest otherwise runs contrary to historical experience and the nature of war."

[4] Thus Mark Lowenthal (2006, 129) writes in arguably the most important textbook for intelligence studies that numeric probabilities "run the risk of conveying to the policy client a degree of precision that does not exist. What is the difference between a 6-in-10 chance and a 7-in-10 chance, beyond greater conviction? In reality, the analyst is back to relying on gut feeling."

[5] On foreign policy analysts' natural tendencies towards overconfidence, see Johnson (2004) and Tetlock (2005). The U.S. National Intelligence Council (2007, iv) thus explained its use of qualitative probability phrasings by writing that "assigning precise numerical ratings to such judgments would imply more rigor than we intend." For a recent theoretical and empirical examination of this "illusions of rigor" thesis, see Friedman, Lerner, and Zeckhauser (forthcoming).

This article employs a unique data set containing 888,328 geopolitical forecasts to examine the extent to which analytic precision improves the predictive value of foreign policy analysis. We find that coarsening numeric probability assessments in a manner consistent with common qualitative expressions consistently sacrifices predictive accuracy. This result does not depend on extreme probability estimates, short time horizons, particular scoring rules, or question content. We also examine how individual-level factors predicted a forecaster's ability to parse probability assessments with meaningful detail. Contrary to popular notions that this ability hinges on attributes like education, numeracy, and cognitive style, we find that a broad range of forecasters can reliably parse their forecasts with numeric precision. At a practical level, our analysis indicates that it would be possible to make foreign policy discourse more informative by supplementing natural language-based descriptions of uncertainty with quantitative probability estimates.

We present our analysis in six parts. The first section frames debates about assessing uncertainty in world politics in relation to broader controversies about subjective probability in the social sciences. The second section introduces our data set of geopolitical forecasts. The third section describes our empirical methodology. The fourth section shows how commonly-used qualitative expressions systematically sacrificed predictive accuracy across the forecasts we examined, and demonstrates the robustness of this finding with respect to different scoring rules, time horizons, and question types. The fifth section analyzes how returns to precision varied across participants in our study. We conclude by discussing implications for international relations scholarship, as well as for broader efforts to improve discussions of uncertainty in foreign policy discourse.

How Much Precision Does Foreign Policy Analysis Allow?

Aristotle (1985, 1049b) argued that "the educated person seeks exactness in each area to the extent that the nature of the subject allows." In some areas of world politics, scholars have demonstrated that statistical analyses, game-theoretic models, and other algorithmic techniques can generate rigorous, numeric predictions (Ward 2016; Schneider, Gleditsch, and Carey 2011; Bueno de Mesquita 2009). Yet foreign policy analysts regularly confront questions that do not suit these methodologies. The vast majority of probabilistic judgments offered by intelligence analysts, military planners, scholars, pundits, and other participants in foreign policy debates reflect subjective beliefs that are grounded in professional judgments, not algorithmic output (Tetlock 2010).[6] These are the cases where analytic precision often seems hardest to justify. According to John Stuart Mill (1882, 539), probability assessments "are of no real value" unless analysts derive them from large volumes of reliable data. John Maynard Keynes (1937, 213-14) wrote that "About these matters, there is no scientific basis on which to form any calculable probability whatsoever. We simply do not know" (cf. Keynes 1921).

The notion that some concepts are inherently qualitative or otherwise resistant to precision has a long-standing pedigree in the social sciences. Karl Popper (1972, 207) suggested that social science problems fall on a continuum where one extreme resembles "clocks," which are "regular, orderly, and highly predictable," and the other extreme resembles "clouds," which are "highly irregular, disorderly, and more or less unpredictable." Many international relations scholars believe that world politics lies at the far, disorderly end of that spectrum. One prominent way to articulate this view is the idea that foreign affairs and armed conflicts involve "nonlinear" dynamics, where small changes to a system's inputs can cause huge swings in that system's outputs (Beyerchen 1992/93, Jervis 1997; Pape 1997/98; Betts 2000; Mattis 2008; Connable 2012). This framework casts doubt on the notion that foreign policy analysts can draw anything beyond coarse distinctions when assessing the uncertainty that surrounds consequential issues.

Meanwhile, a large number of empirical studies show that subject matter experts often struggle to outperform simple algorithms when assessing uncertainty (Dawes, Faust, and Meehl 1987, Tetlock 2005). The "behavioral revolution" in the social sciences has demonstrated that a broad range of heuristics and biases can warp assessments of uncertainty in world politics (Hafner-Burton et al. 2017). The most consequential of these biases for our purposes is overconfidence, which is the tendency of probability assessors to attach too much certainty to their estimates (Jervis 1976; Johnson 2004; Tetlock 2005). This bias suggests that coarsening

[6] The role of intuitive judgment in assessing uncertainty is one of the foundational topics in strategic studies, seen most clearly in debates between Carl von Clausewitz (1832) and Antoine-Henri de Jomini (1838) over the value of mathematical logic in military planning. See Beyerchen (1997/98) and Gat (1989) for commentary.

probability estimates might actually make them more *accurate*, by counterbalancing foreign policy analysts' tendencies to make extreme judgments.[7]

Yet just because foreign policy analysts struggle to assess uncertainty, this does not mean it is desirable to leave these judgments vague. Deliberately coarsening these judgments could still reduce their value, and several bodies of research suggest that foreign policy analysts may in fact possess a reliable ability to parse subjective probability assessments in detail. Prediction markets, for example, can often extract meaningful, fine-grained probability estimates from the wisdom of crowds (Meirowitz and Tucker 2004; Arrow et al. 2008). A recent, large-scale study of Canadian intelligence reports found geopolitical forecasts to be surprisingly well-calibrated (Mandel and Barnes 2014). In an even larger study using predominantly non-governmental respondents, researchers identified a group of "superforecasters" who predicted international events with considerable accuracy (Tetlock and Gardner 2015). These research programs suggest that even if subjective probability assessments are not scientific in the sense that Keynes and Mill used that term, leaving these assessments vague could still systematically sacrifice meaningful information.

Debates about the value of precision in probability assessment thus have three main implications for foreign policy analysis. At a theoretical level are long-standing questions about the extent to which the complexity of world politics impedes analysts' capacity to assess uncertainty with meaningful detail. At a methodological level lie debates about the extent to which subjective judgmental processes can sustain the kinds of analytic precision that are generally thought to be reserved for algorithmic analyses. And at a practical level, the standard practice of leaving probability assessments vague could needlessly impair the quality of foreign policy discourse. Given how uncertainty surrounds nearly every intelligence report, military plan, or foreign policy debate, even small improvements in this area could bring major aggregate benefits. Yet to date, no empirical study has examined this subject systematically.

[7] In other words, if analysts who use the qualitative expressions shown in Figure 1 were to make those judgments more precise, they might resolve this ambiguity in a manner that imparts excessive certitude to their estimates.

Probabilistic precision and estimative language

In this article, we use the phrase *returns to precision* to describe the degree to which quantitative probability estimates convey greater predictive accuracy than qualitative terms or quantitative estimates that are systematically less precise than the original forecasts. Returns to precision need not be positive: as described above, analytic precision could enable counterproductive tendencies among overconfident assessors. We define probabilistic precision by segmenting the number line into "bins." When analysts express uncertainty using all integer percentages, this divides the probability continuum into 101 equally-sized bins, including zero and one. The guidelines for expressing uncertainty shown in Figure 1 divide the number line into seven bins with varying definitions. These guidelines reflect the assumption that foreign policy analysts lack the ability to consistently parse their probability assessments into more than seven categories.

Many official reports describe uncertainty more coarsely than this. Following controversy over assessments of Saddam Hussein's presumed weapons of mass destruction programs, for instance, intelligence analysts received criticism for framing their judgments using "estimative verbs" such "we assess," "we believe," or "we judge."[8] As the guidelines at the bottom of Figure 1 explain, estimative verbs indicate that judgments are uncertain. However, these terms provide little information about what the relevant level of uncertainty entails, beyond implying that a statement is likely to be true. In this sense, expressing uncertainty through estimative verbs divides the number line into two bins.

Confidence levels divide the number line into three bins, corresponding to the terms "low confidence," "moderate confidence," and "high confidence." While probability and confidence technically reflect different concepts, many foreign policy analysts appear to conflate these terms

[8] For example, the 2002 National Intelligence Estimate (NIE) on *Iraq's Continuing Program for Weapons of Mass Destruction* states: "We assess that Baghdad has begun renewed production of [the chemical weapons] mustard, sarin, GF (cyclosarin), and VX." Then: "We judge that all key aspects – R&D, production, and weaponization – of Iraq's offensive BW [biological weapons] programs are active." See Jervis (2010) and Wheaton (2012) for critiques of this practice.

or to use them interchangeably.[9] For example, the lexicon at the bottom of Figure 1 appears in a 2017 Intelligence Community Assessment describing Russian interference in the previous year's U.S. presidential election.[10] This lexicon explains that intelligence analysts should communicate probability using fourteen terms grouped into seven equally-spaced segments along the number line. But the report's key judgments do not use any of those terms. The report thus assesses with "high confidence" that Russian President Vladimir Putin interfered with the election in order to undermine faith in the U.S. democratic process.[11] The report then assesses that President Putin staged this intervention with the objective of helping then-Republican candidate Donald Trump to defeat then-Democratic candidate Hillary Clinton. Here, the Central Intelligence Agency and the Federal Bureau of Investigation placed "high confidence" in their judgment, whereas the National Security Agency only made this assessment with "moderate confidence." A statement made with "high confidence" presumably reflects a higher perceived likelihood than a statement made with "moderate confidence," particularly when analysts do not assess probability and confidence independently. In this sense, "confidence levels" effectively divide assessments of uncertainty into three bins.[12]

Of course, analysts of foreign policy or any other field could divide the number line into however many bins they prefer. Yet most existing recommendations for expressing probability in

[9] In principle, "probability" describes the chances that a statement is true and "confidence" describes the extent to which an analyst believes that she has a sound basis for assessing uncertainty. Thus most people would say that a fair coin has a fifty percent probability of turning up heads, and they would make this estimate with high confidence. On the seemingly-interchangeable use of probability and confidence in national security decision making, see Friedman and Zeckhauser (2012, 834-841).

[10] ICA 2017-01D, *Assessing Russian Activities and Intentions in Recent U.S. Elections* (January 2017).

[11] Emphasis added.

[12] It is possible that the authors of this report intended to convey probability through the estimative verbs "we assess." In this case, the judgments would not have conflated probability and confidence, but they would have been even more vague in conveying the chances that these statements were true.

foreign policy analysis employ one of four alternatives: estimative verbs (two bins), confidence levels (three bins), words of estimative probability (seven bins), or integer percentages (101 bins).[13] The next section describes the data and method we use to evaluate how these or any other systems of expression influence the predictive accuracy of geopolitical forecasts.

Data

Our study employs data gathered by the Good Judgment Project (GJP). The Good Judgment Project began in 2011 as part of a series of large-scale geopolitical forecasting tournaments sponsored by the Intelligence Advanced Research Projects Activity (IARPA). IARPA distributed forecasting questions to participants, whom GJP recruited. Forecasters logged responses to those questions on a website using numeric probability estimates.[14] In supplementary material, we provide extensive descriptions of the GJP's data and methods, which have also been documented in several recent studies (e.g., Mellers et al. 2014; 2015a; and 2015b).

IARPA's question list covered topics such as the likelihood of candidates winning Russia's 2012 presidential election, the probability that China's economy would exceed a certain growth rate in a given quarter, and the chances that North Korea would detonate a nuclear bomb before a particular date. IARPA chose these questions to reflect a broad range of issues that impact foreign policy decision making. The organization made no attempt to ensure that these questions were suitable for the use of statistical analysis, game-theoretic models, or other algorithmic techniques. (Indeed, the principal goal of the tournament was to encourage participants to

[13] For descriptions of additional experiments with quantifying subjective probabilities in foreign policy analysis, see Nye (1994); Lanir and Kahneman (2006); Marchio (2014); and Barnes (2016).

[14] The Good Judgment Project also administered a prediction market, but we do not analyze those data in this article. Though prediction markets have been shown to generate reliable forecasts in many settings, they mainly allow individuals to register their opinion that an event's true probability lies above or below a given market price. By comparison, the numeric probability assessments that we analyze in this article provide more direct insight into how each individual estimated the chances that particular events would occur.

develop whatever techniques they believed would be most effective for addressing a broad range of questions.[15]) The main exception to the ecological validity of IARPA's question list was the requirement that each question be written precisely enough so that outcomes could be judged clearly after the fact.

This article focuses on the performance of individuals who registered at least twenty-five predictions in a given tournament year.[16] The resulting data set spans 1,832 unique individuals who registered 888,328 forecasts in response to 380 questions administered between 2011 and 2015. Participants tended to be males (eighty-three percent) and U.S. citizens (seventy-four percent). Their average age was forty. Sixty-four percent of respondents had a bachelor's degree, and fifty-seven percent had completed post-graduate training. In the penultimate section of this article, we explore the extent to which individual attributes predict analysts' abilities to extract meaningful returns to precision.

The Good Judgment Project randomly assigned forecasters to work alone or in collaborative teams. Another random subset of forecasters received a one-hour online training module covering various techniques for effective forecasting. This training module covered topics such as defining base rates, avoiding cognitive biases, and extrapolating trends from data.[17] (We describe the purpose and results of this training in greater detail below when analyzing how returns to precision varied across individual forecasters.) These basic divisions are helpful for grounding our analysis. We expected that untrained forecasters who worked alone would make the lowest-quality forecasts, and that those forecasts would also demonstrate the lowest returns to precision in the data set. Yet most foreign policy analysts – especially those who work for governments or universities – are not untrained individuals. Most foreign policy practitioners

[15] The Good Judgment Project won this competition by employing a technique that pooled opinions from a broad range of forecasters, weighting those opinions based on forecasters' prior performance, and then extremizing aggregate results. See Satopää et al. (2014).

[16] We also limit our focus to questions involving binary, yes-or-no outcomes, in order to avoid potential confounding in our calculation of rounding errors.

[17] We included a version of the GJP's training manual with this article's supplementary files. For discussion, see Mellers et al. (2014).

collaborate closely with their peers, and almost all of them receive some kind of formal training.[18] We therefore expect that forecasters who work in groups and who received training will not only demonstrate higher returns to precision in their forecasts, but also that their performance will be more relevant for judging the abilities of professional foreign policy analysts.

At the end of each year, GJP identified the top two percent of performers as "superforecasters." One of the Good Judgment Project's principal findings was that superforecasters' predictions generally remained superior to those of other respondents in subsequent tournament years, contrary to expectations that high-performers would regress to the mean. Elsewhere, Mellers et al. (2014, 2015a, 2015b) and Tetlock and Gardner (2016) describe the superforecasters in more detail. Generally speaking, superforecasters were relatively numerate and relatively knowledgeable about foreign policy, but they were not necessarily experts in particular subjects or methodologies. Instead, the superforecasters typically shared a willingness to address each forecasting problem in a flexible, ad hoc manner, and to draw on an eclectic range of inputs rather than any particular theoretical or methodological framework. This method of analysis proved surprisingly effective – but, as mentioned above, many scholars and practitioners believe that this style of reasoning is also ill-suited for analytic precision, particularly when analyzing complex phenomena like world politics.

GJP's data are uniquely positioned to evaluate returns to precision in geopolitical forecasting due to the sheer volume of forecasts that the Good Judgment Project collected, the range of individuals that the project involved, and IARPA's efforts to ensure that forecasters addressed questions relevant to practical concerns. Nevertheless, we note four principal caveats for interpreting our results.

First, the Good Judgment Project did not randomize the response scale that forecasters employed. Thus GJP's data do not offer a true experimental comparison of numeric percentages versus "words of estimative probability," confidence levels, or estimative verbs. Nonetheless, we

[18] The U.S. Intelligence Community, for example, send incoming analysts into training programs that can last several months; the U.S. military sends officers to multiple training programs (including masters'-lever education for officers who reach the rank of Colonel or Commander).

do not believe that this threatens our inferences. In order to choose appropriate terms from Figure 1, for instance, analysts must first determine where their judgments fall on the number line. Each of the guidelines shown in Figure 1 thus already requires employing approximate numerical reasoning. Moreover, randomizing modes of expressing probability would introduce a fundamental measurement problem. When analysts use words like "high confidence," there is no reliable way to know whether they mean probabilities closer to seventy percent or to ninety percent. Thus we cannot tell whether a "high confidence" forecast was closer to the truth than a forecast of eighty percent when predicting an outcome that occurred. For these reasons, "rounding off" numerical forecasts in a manner that is consistent with different modes of qualitative expression is the most straightforward way to estimate returns to precision, and we describe our methodology for doing this below.

A second caveat for interpreting our results is that the Good Judgment Project only asked respondents to make predictions with time horizons that could be resolved during the course of the study. The average prediction was made seventy-six days (standard deviation, eighty days) before questions closed for evaluation. Thus GJP data cannot describe the relationship between estimative precision and predictive accuracy on forecasts with multi-year time horizons. However, we will demonstrate our findings' robustness across time horizons within GJP data.

Third, GJP only asked respondents to assess the probability of future events, but foreign policy analysis also requires making probabilistic statements about current or past states of the world, such as whether a state is currently pursuing nuclear weapons or whether a terrorist is hiding in a suspected location. Generally speaking, we expect that analysts should find it more difficult to parse probabilities when making forecasts, as forecasting requires assessing imperfect information while accounting for additional uncertainty about how states of the world may change in the future. If predicting the future is harder than assessing uncertainty about the present and past, then our findings should be conservative in identifying returns to precision when estimating probabilities in international affairs. Without the data necessary to substantiate this claim directly, however, we emphasize that our empirical analysis measures returns to precision in political *forecasting*, which is a subset of political analysis writ large.

Finally, because the Good Judgment Project's respondents volunteered their participation, we cannot say that these individuals comprise a representative sample of foreign policy analysts.

Since GJP gathered extensive information on its participants, however, we can examine whether returns to precision correlate with factors such as education, numeracy, cognitive style, or other individual attributes. In the second-to-last section of this article, we show that none of these attributes predicts substantial variation in returns to precision, especially relative to factors such as skill, effort, and training in probabilistic reasoning.

Methodology

Our basic methodology involves three steps. First, we measure the predictive accuracy of respondents' original, numeric probability assessments. Next, we "round off" those estimates in a manner that makes them less precise. Then we calculate the extent to which coarsening estimates changed their predictive accuracy. In this section, we explain each of these steps in more detail. In particular, we highlight how we adopted deliberately conservative statistical assumptions that presumably understate returns to precision across our data. Supplementary files contain a formal description of our technique.

Step 1: measuring predictive accuracy

It is difficult to evaluate the quality of a single probability assessment.[19] If an analyst estimates a seventy percent chance that some candidate will win an election but then the candidate loses, it is hard to say how much we should attribute this surprise to poor judgment versus bad luck. But when examining a large volume of probability assessments together, we can discern broad trends in their accuracy. Thus if we take a large body of cases where analysts estimated that their conclusions were seventy percent likely to be true, we can see whether those conclusions were actually true roughly seventy percent of the time (Rieber 2004; Tetlock 2005; Mandel and Barnes 2014).

Our main metric for measuring predictive accuracy in this article is the commonly-used Brier Score, though we will also show that our results are robust to using an alternative, logarithmic

[19] Betts (2000) explains how this poses major analytic problems for retrospective evaluations of military strategy.

scoring rule.[20] The Brier Score measures the mean squared error of an assessor's judgments. For example, consider the question, "Will Bashar al-Assad be ousted from Syria's presidency by the end of 2017?" There are two possible outcomes here: either Assad is ousted (in which the "true probability" is eventually realized to be one hundred percent), or he remains (in which case the "true probability" is eventually realized to be zero percent). Assume that our forecaster predicts a sixty percent chance that Assad is ousted and a forty percent chance that he remains. If Assad is ousted, then the forecaster's score for these assessments would be 0.16; if Assad remains, then the forecaster's score for these assessments would be 0.36.[21] Since the Brier Score measures judgmental error, lower Brier Scores reflect better forecasts, indicating that respondents assign higher probabilities to events that occur, and lower probabilities to events that do not occur.[22]

Step 2: translating numeric forecasts into corresponding verbal expressions

To translate numerical forecasts into corresponding verbal expressions, we round probability assessments to the midpoint of the bin that each verbal expression represents. For example, according to the Director of National Intelligence's current analytic standards, the phrase "even chance" implies a predicted probability between forty-five and fifty-five percent. Absent additional information, the expected value of a probability estimate falling within this range is fifty percent. In practice, a decision maker may combine this estimate with other information and prior assumptions to justify a prediction that is higher or lower than fifty percent. However,

[20] The Brier Score is more appropriate for our purposes because of the severe penalties which the logarithmic scoring rule assigns to misplaced extreme estimates. Logarithmic scoring requires changing estimates of 0.00 and 1.00 (comprising nineteen percent of our data points), since an error on these estimates imposes an infinite penalty.

[21] If Assad is ousted, the forecaster's Brier Score is calculated as $[(1.00 - 0.60)^2 + (0.00 - 0.40)^2]/2 = 0.16$. If Assad remains, the forecaster's score is $[(0.00 - 0.60)^2 + (1.00 - 0.40)^2]/2 = 0.36$.

[22] Again, see Tetlock (2005) for more discussion of the Brier Score.

saying that a probability is equally likely to fall anywhere within a range conveys the same expected value as stating that range's midpoint.[23]

We generalize this approach by dividing the number line into B bins, then rounding each forecast to the midpoint of its associated bin. Thus if we divide the number line into three equally-sized bins (which would be consistent with assigning "high," "medium," and "low" confidence levels), then we would round all forecasts within the highest bin (67-100 percent) to the range's midpoint of 83.3 percent.[24] When forecasts fall on boundaries between bins, as with a forecast of fifty percent when $B = 2$, we randomize the direction of rounding.[25]

Step 3: calculating "rounding errors"

Our data comprise 888,328 forecasts. However, these forecasts are correlated within questions and within individuals who updated forecasts before questions closed for evaluation.[26] It would therefore be inappropriate to treat all forecasts in our data set as independent observations. We thus take the forecasting question as our unit of analysis. We do this by identifying a subset of forecasters to evaluate (all forecasters, superforecasters, etc.), and then calculating an *aggregate Brier Score* for that group on each forecasting question. This method represents a deliberately conservative approach to statistical analysis, because it reduces our maximum sample size from

[23] While it is inappropriate to translate uncertainty about quantities into single point estimates, it is the proper way to treat uncertainty about probabilities. For discussion of this point, see Ellsberg (1961).

[24] We also experimented with rounding probabilities to the empirically-weighted mean of each range, so that if responses to a particular question clustered near one hundred percent, then we would round numeric estimates to a point that was higher than if those responses clustered near seventy percent. Our findings are robust to this alternative approach.

[25] Though the Director of National Intelligence's guidelines define "remote" and "almost certain" as comprising assessments of 0.01-0.05 and 0.95-0.99, respectively, we included GJP forecasts of 0.0 and 1.0 in these categories.

[26] Respondents updated their forecasts an average of 1.49 times per question.

888,328 forecasts to 380 forecasting questions. Evaluating individual forecasts returns similar estimates of returns to precision, albeit with inappropriate levels of statistical significance.[27]

We calculate *rounding errors* on forecasting questions by measuring proportional changes in Brier Scores when we round individual forecasts into bins of different widths. For example, we might have found that the average Brier Score for all untrained individuals on a particular question was 0.160. After rounding off this group's forecasts to the midpoints of three bins we might find that their average Brier Score climbs to 0.200. In that case, we would say that rounding the untrained individuals' estimates into three bins induced a rounding error of twenty-five percent. In the analysis below, we will analyze these rounding errors in two ways: by how much coarsening estimates influences predictive accuracy on *average*, and also by examining these changes at the *median*.[28] Analyzing means and medians together is important for ensuring that outliers do not drive our inferences, and so that we can describe how coarsening probability estimates influences the accuracy of the typical judgment.

[27] Our aggregation method has the additional advantage that averaging across days during which a question remained open reduces the influence of forecasts made just before a closing date. Later in the article, we further demonstrate that short-term forecasts do not drive our results.

[28] We report statistical significance in two ways, as well, using standard two-way t-tests when comparing means, and using Wilcoxon signed-rank tests when comparing medians.

Table 1. Estimative Precision and Predictive Accuracy – Aggregated Results

			Rounding Errors			
Reference class		*Brier Scores for Numerical Forecasts*	Words of estimative probability[†] (2015 version)	Words of estimative probability (7 equal bins)	Confidence levels (3 bins)	Estimative verbs (2 bins)
All forecasters	*Mean:*	0.153	0.7%***	1.9%	11.8%***	31.4%***
	Median:	0.121	0.9%***	1.2%***	7.3%***	22.1%***
Untrained individuals	*Mean:*	0.189	0.5%***	0.5%***	5.9%***	15.0%***
	Median:	0.162	0.6%***	0.2%	3.6%***	9.9%***
Trained groups	*Mean:*	0.136	0.8%***	3.3%*	17.8%***	48.6%***
	Median:	0.100	0.9%***	2.4%***	11.0%***	30.1%***
Super-forecasters	*Mean:*	0.093	6.1%***	40.4%***	236.1%***	562.0%***
	Median:	0.032	1.7%***	10.2%***	54.7%***	141.7%***

Table 1 shows rounding errors for different groups of respondents, depending on the degree of imprecision to which we round their forecasts. We estimate whether these rounding errors are statistically distinct from zero using paired-sample t-tests (for differences in means) and Wilcoxon signed-rank tests (for differences in medians). * p<0.05, ** p<0.01, *** p<0.001.

[†] Currently recommended by the Office of the Director of National Intelligence (see Figure 1).

How Vague Probability Assessments Sacrifice Information

Table 1 shows how rounding GJP forecasts to different degrees of (im)precision consistently reduced their predictive accuracy. On average, Brier Scores associated with GJP forecasts become thirty-one percent worse when rounded into two bins. Outliers do not drive this change, as the median rounding error is twenty-two percent. Even the worst-performing group of forecasters, untrained individuals, incurred an average rounding error of fifteen percent when we rounded their forecasts to "estimative verbs." The penalty for superforecasters was far worse, with average rounding errors over five hundred percent. We also see large rounding penalties from shifting GJP forecasts to "confidence levels": on average, this level of imprecision degraded forecast accuracy by more than ten percent, and substantially more for high-performing forecasters.

Rounding forecasts into seven-step "words of estimative probability" (WEPs) recovered some, but not all, of these losses. Despite our conservative approach to estimating statistical significance, and despite analyzing median changes to minimize the influence of outliers, every subgroup in our analysis encountered consistent losses of predictive accuracy when we rounded forecasts according to the lexicon currently recommended by the U.S. Director of National Intelligence. The National Intelligence Council's current guidelines for expressing uncertainty, which divide probability assessments into seven equal bins, induce greater variance: rounding errors here tended to be larger but also less consistent.[29] Superforecasters continued to suffer the largest losses under both systems of expression. Coarsening probability assessments thus prevented the best forecasters from reaching their full potential, sacrificing information disproportionately from the sources that produced the most reliable assessments.

These comparisons are especially meaningful in relation to the challenges that scholars generally face when evaluating methods of intelligence estimation. Mark Lowenthal (2008, 314), a scholar with three decades' experience in the U.S. Intelligence Community, observes that "No

[29] The Director of National Intelligence's spectrum compensates for tightening the "remote" and "almost certain" bins by widening the "likely" and "unlikely" bins. This makes a majority of forecasts worse (and the difference in means more statistically significant) even as average rounding errors decline.

one has yet come up with any methodologies, machines or thought processes that will appreciably raise the Intelligence Community's [performance]."[30] Thomas Fingar (2011, 34, 130), formerly the U.S. Intelligence Community's top analyst, writes that "By and large, analysts do not have an empirical basis for using or eschewing particular methods." By contrast, our results *do* provide an empirical basis for expressing probabilities more precisely than what standard practice allows. And given how questions about expressing probability surround virtually every intelligence report, military plan, and foreign policy debate, even small improvements in the quality of these judgments are worth pursuing.

Rounding errors across the number line

We now examine whether there are specific kinds of forecasts where respondents consistently extracted larger (or smaller) returns to precision. It is important to determine whether returns to precision appear primarily when making "easy" forecasts. Two main indicators of forecasting ease are the forecast's size (as more extreme probabilities reflect greater certainty, which may correlate with easier questions) and its time horizon (as nearer-term events may be easier to predict). We also examine the extent to which questions pertaining to particular regions or topics may have been easier for analysts to address with precision. Our results show that GJP respondents extracted returns to precision across a broad range of forecasts.

Figure 2 presents a histogram of GJP forecast values.[31] As a general rule, GJP forecasters assigned estimates at intervals of five percentage points.[32] This pattern alone is important, indicating that when left without restrictions on how fine-grained their forecasts should be, GJP

[30] Cf. Tetlock and Mellers (2011). As Tetlock (2010) explains, even when foreign policy analysts possess empirically-validated methods, those methods are rarely tested directly against rival approaches.

[31] The histogram is symmetric because predicting that an outcome will occur with probability p implies that the outcome will *not* occur with probability $1 - p$.

[32] Forty-nine percent of forecasts in the data set are multiples of 0.10 and 25 percent of forecasts are additional multiples of 0.05.

respondents preferred to express probabilities with greater detail than what common qualitative expressions allow.

To see how returns to precision varied across the probability spectrum, we divided forecasts into seven bins according to the National Intelligence Council guidelines shown in Figure 1. We separately examined forecasts falling within each of these bins. Table 2 shows that GJP analysts, on the whole, demonstrated returns to precision across the number line. We found that rounding superforecasters' estimates according to National Intelligence Council guidelines consistently sacrificed information within all seven categories. And though we found mixed results from rounding non-superforecasters' most extreme estimates – Table 2 shows how coarsening these estimates degraded their accuracy on average but improved them at the median – this finding only reinforces how our overall estimates of returns to precision are not driven by the most extreme forecasts in our data set.

Figure 2. Histogram of forecasts in GJP data

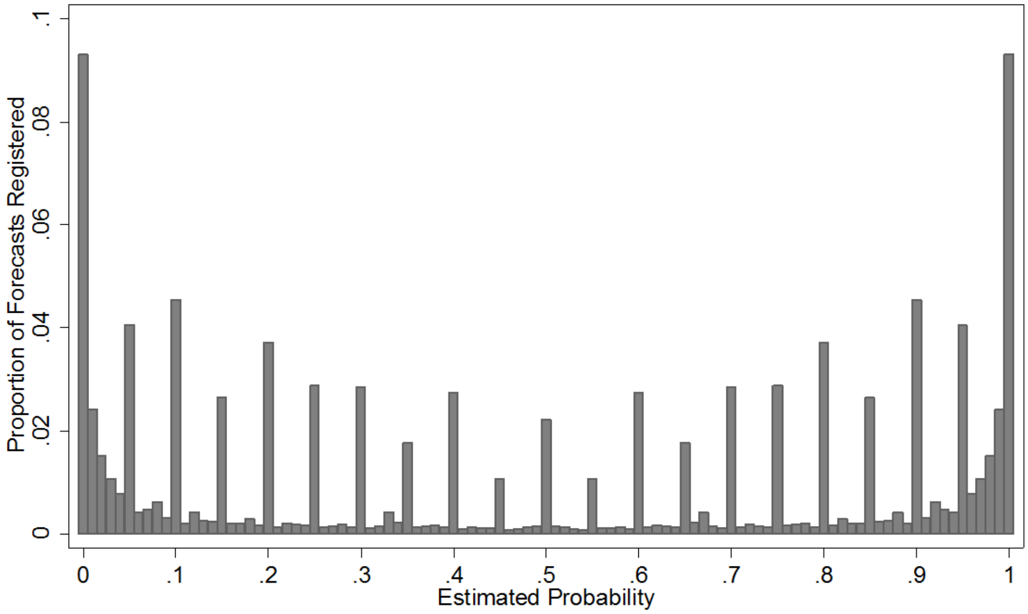


Table 2. Rounding Errors across the Probability Scale (Brier and Logarithmic Scoring)

			Very		Even		Very	Almost
		Remote	Unlikely	Unlikely	chance	Likely	Likely	Certain
		(.00-.14)	(.15-.28)	(.29-.42)	(.43-.56)	(.57-.71)	(.72-.85)	(.86-1.0)
Rounding Errors via Brier Scoring								
All	*Mean:*	3.4%***	4.3%***	2.3%***	1.3%***	2.3%***	4.3%***	3.4%***
Forecasters	*Median*	-0.5%***	3.7%***	2.2%***	1.1%***	2.2%	3.7%***	-0.5%***
Super-	*Mean:*	85.8%***	16.2%***	7.0%***	1.8%***	7.0%***	16.2%***	85.8%***
Forecasters	*Median*	32.2%***	12.1%***	4.1%***	1.0%***	4.1%***	12.1%***	32.2%***
Rounding Errors via Logarithmic Scoring								
All	*Mean:*	-1.1%***	3.5%***	1.6%***	0.9%***	1.6%***	3.5%***	-1.1%***
Forecasters	*Median*	-7.5%***	3.7%***	1.7%***	0.8%***	1.7%***	3.7%***	-7.5%***
Super-	*Mean:*	70.4%	9.9%***	4.4%***	1.2%***	4.4%***	9.9%***	70.4%
Forecasters	*Median*	55.0%***	9.4%***	3.1%***	0.7%***	3.1%***	9.4%***	55.0%***

Table 2 examines how rounding forecasts into seven equal bins influences predictive accuracy for forecasts within different segments of the number line. We estimate whether these rounding errors are statistically distinct from zero using paired-sample t-tests (for differences in means) and Wilcoxon signed-rank tests (for differences in medians). * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

Table 2 also shows that our results do not hinge on the Brier Score's particular properties. When we recalculated rounding errors using a logarithmic scoring rule,[33] we again found that superforecasters exhibited reliable returns to precision in every category, and that rounding sacrificed predictive accuracy for non-superforecasters in every category besides the extremes. [34]

Returns to precision across time horizons

To assess how returns to precision varied across time horizons, we coded the time horizon for each forecast as the number of days between the date when the forecast was registered and the date when the forecasting question was resolved. In our data set, the mean time horizon was seventy-six days (standard deviation, eighty days). We identified forecasts as "lay-ups" if they entailed no more than five percent probability or no less than ninety-five percent probability, and if respondents registered those forecasts within two weeks of a question's closing time. We expected to see special returns to precision on these highly-certain, near-term estimates. We divided all other forecasts into three categories with equal numbers of observations.[35] In supplementary material, we show how our findings are generally consistent across each time period that we analyzed. We thus see no indication that our conclusions rely on easy questions with short time horizons where foreign policy analysts can justify special precision.

[33] This rule scores analysts' predictions according to the natural logarithm of the probability they assigned to the observed outcome. Higher logarithmic scores are better. In order to prevent scores of negative infinity (which is the natural logarithm of zero), we convert estimates of 1.00 and 0.00 to 0.99 and 0.01, respectively.

[34] The benefits to rounding non-superforecasters extreme estimates increase under logarithmic scoring because of the way that this function imposes severe penalties on erroneous estimates made with near-certainty.

[35] There were 109,240 "lay-ups" in our data, leaving 259,696 forecasts in each of the three periods we defined.

Returns to precision across question types

To determine how returns to precision vary across questions, we generated question-specific estimates of returns to precision. We derived these estimates by calculating respondents' Brier Scores after rounding each forecast into progressively larger numbers of bins. For each level of precision, we evaluated whether coarsened forecasts had worse Brier Scores than respondents' original, numeric predictions. We defined each question's *threshold of estimative precision* (B^*) as the smallest number of bins where the median rounding error was not statistically greater than zero.[36] Because we set these B^* thresholds at the lowest possible value where we cannot reject the hypothesis that coarsening did not make forecasts less accurate, and because we tested this hypothesis by comparing median rounding errors instead of mean rounding errors, this represents another deliberately conservative method for describing returns to precision. Among the 375 questions that we analyzed in this way,[37] the mean B^* threshold was 6.1 bins, with a standard deviation of 4.4 bins. These B^* thresholds exceeded seven bins for forty-two percent of GJP's questions.

We thus found that employing official guidelines for expressing uncertainty using qualitative expressions systematically reduced the accuracy of GJP forecasts for nearly half of the questions in our data set. Our results clearly do not hinge on a few questions where forecasters happened to make particularly informative estimates. In supplementary material we further show how classifying questions according to eleven regions and fifteen topics provided little traction for predicting B^* thresholds. In other words, we see little indication that forecasters' ability to extract meaningful returns to precision is confined to particular topics.

Examining Variation across Individuals

Having shown that forecasters can assess subjective probabilities more precisely than what the conventional wisdom and standard procedures allow, we turn to the question of which foreign

[36] We made this determination using a comparison of medians, based on a one-sided, paired-sample Wilcoxon signed rank test with a 0.05 threshold for statistical significance

[37] We dropped five questions from this analysis due to missing data.

policy analysts tend to achieve higher returns to precision than others. We focus our analysis on understanding whether the capacity to achieve these returns to precision appears to be innate, or whether this is an ability that analysts can feasibly cultivate.

The notion that some people are incapable of numeric reasoning is widespread in the literature on foreign policy analysis. In his original essay criticizing intelligence analysts for making vague probability estimates, Sherman Kent (1964) famously divided his colleagues into "mathematicians" and "poets" based on their comfort with quantitative expressions.[38] Meanwhile, empirical research on political forecasting such as Tetlock (2005) or Tetlock and Gardner (2016) finds that individuals' overall forecasting skill tends to be correlated with attributes like education, numeracy, and cognitive style. If these attributes also play a prominent role in predicting analysts' capacities to achieve returns to precision, then this would not only help scholars to understand the cognitive processes underlying subjective judgments, but it would also have two important practical implications. First, it would suggest that "poets" or other quantitatively-averse analysts might justifiably opt out of making clear probability assessments. And second, this would suggest that organizations that wish to improve the returns to precision among their analysts would need to do this primarily through processes for selecting personnel. For the remainder of this section, we therefore refer to numeracy, education, and cognitive style as "targets for selection."

At the same time, decision scientists have produced a large volume of evidence indicating that even low-quality probability assessors can substantially improve their performance through training and feedback.[39] For example, the Good Judgment Project found that just one hour of randomly-assigned training had a substantial, persistent impact on improving forecasters' Brier Scores (Mellers et al. 2014). Moreover, even if foreign policy analysts are initially

[38] Johnston (2004), Nye (1998), and Marchio (2014) provide similar descriptions of this cultural divide among foreign policy analysts. On numeracy and probability assessment more generally, see Peters et al. (2006).

[39] The prospect for improving forecasting skill, even with relatively limited training, is well-established in the decision science literature. See Dhami et al. (2015) for a review of relevant literature with applications to foreign policy analysis specifically.

uncomfortable translating their subjective judgments into explicit language, this problem might diminish over time as analysts became more fluent using numeric expressions. If that is true, then it suggests that returns to precision might correlate with variables we call "targets for cultivation." If the targets for cultivation are the primary predictors of returns to precision, then that would be an optimistic finding. It would suggest that a broad range of foreign policy analysts could begin making their assessments of uncertainty more informative *right now* if they believed it was important to do so.

In the remainder of this section we describe six "targets for cultivation" and six "targets for selection." We then analyze the extent to which these variables correlate with individual returns to precision. The first of our "targets for cultivation" variables is forecasting skill, as measured by each respondent's median Brier Score. Higher-quality forecasters should incur greater penalties from having their forecasts rounded. However, this relationship is not tautological. It is possible for a forecaster to demonstrate excellent discrimination (separating events that are likely from those that are unlikely) even if she is not especially well-calibrated (that is, she cannot make fine-grained distinctions among events that are either likely or unlikely to occur). This forecaster might obtain a relatively low Brier Score without suffering significant penalties from coarsening her estimates. This is, in fact, the hypothesis implied by recommendations that foreign policy analysts express their judgments using coarse language like "estimative verbs" or "confidence levels." The notion is that foreign policy analysts can reliably discriminate *among* rough categories, but that they cannot draw meaningful differences *within* those categories.

Five additional variables capture effort, training, experience, and collaboration – these are all factors over which foreign policy analysts possess substantial control. *Number of Questions* counts the number of distinct questions to which an individual responded throughout all years of the competition. *Average Revisions per Question* captures how many times respondents tended to update their beliefs before each question closed for evaluation. These variables proxy for the effort that respondents expended in engaging with the competition and for their experience responding to forecasting questions. We expect that respondents who score higher on these measures will demonstrate additional returns to precision. We also captured the *Granularity* of each respondents' forecasts by measuring the proportion of those forecasts that were not recorded in multiples of ten percentage points. We expect that respondents who are comfortable expressing their views more precisely, or who took the additional effort to do so, would incur larger rounding penalties than forecasters who provided coarser judgments.[40]

Probabilistic Training takes a value of 1 if the forecaster received training in probabilistic reasoning from GJP. We expect that respondents who received training in probabilistic reasoning would be more effective at parsing their probability assessments. As mentioned above, these training sessions lasted about one hour, and covered basic concepts such as base rates, reference classes, and ways to mitigate cognitive biases. *Group Collaboration* takes a value of 1 if a forecaster was assigned to collaborate with a team in GJP's competition. We expect that respondents working in groups would be exposed to more information that would help in parsing their estimates effectively, including the ability to anchor-and-adjust off teammates' assessments.

In examining "targets for selection," we considered six variables. The first two of these variables capture respondents' education prior to participating in the Good Judgment Project. *Education Level* is a four-category variable capturing a respondent's highest academic degree (no bachelor's degree, bachelor's degree, master's degree, doctorate).[41] Advanced education could enhance a respondent's ability to analyze complex questions and to parse probabilities reliably. *Numeracy* represents respondents' scores on a series of word problems designed to capture mathematical fluency (Peters et al. 2006). Respondents who are better able to reason numerically might parse probabilities more effectively.[42] In principle, organizations can cultivate both of these attributes. However, numeracy and education levels are substantially more expensive to increase than the effort and training variables described above.

GJP data also include several indices of "cognitive style," including *Raven's Progressive Matrices*, where higher scores indicate better reasoning ability (Arthur et al. 1999); an expanded

[40] An index of granularity representing the proportion of forecasts that were not multiples of 0.05 yields similar results.

[41] If a respondent participated in multiple years of the forecasting competition, we averaged Education values across years.

[42] GJP changed numeracy tests between years 2 and 3 of the competition. We standardized numeracy test results so that they represent comparable indices. If a respondent participated in multiple years of the forecasting competition, we averaged Numeracy values across years.

Cognitive Reflection Test (*Expanded CRT*), where higher scores indicate an increased propensity to suppress misleading intuitive reactions in favor of more accurate, deliberative answers (Baron et al. 2015); *Fox-Hedgehog*, a variable on which higher scores capture respondents' self-assessed tendency to rely on ad hoc reasoning versus simplifying frameworks (Mellers et al. 2014); and *Need for Cognition*, an index of respondents' self-assessed preference for addressing complex problems (Cacioppo and Petty 1982).[43] In addition to the primary variables of interest described above, we also control for age, for gender, and for whether a respondent was designated as a superforecaster in any tournament year. We include those variables in all models. Supplementary files contain full descriptive statistics for each of the variables described here.

Analyzing individual-level returns to precision

Table 3 presents ordinary least squares regression analyses predicting forecasters' B^* thresholds as a function of different individual attributes. We standardized non-binary independent variables. Each coefficient in Table 3 thus reflects the extent to which B^* thresholds improve, on average, when each predictor increases by one standard deviation, or when binary variables change from 0 to 1.

[43] If a respondent participated in multiple competition years, we averaged values across years. GJP changed CRT tests after Year 2 of the competition, so we standardized each test's results in order to provide comparable measures.

Table 3. Predicting Individual-Level Returns to Precision

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*†
Targets for cultivation					
Brier Score	-1.62 (.15)***	-1.57 (.16) ***	-1.80 (.24)***	-1.74 (.26)***	-1.77 (.26)***
Number of Questions		1.15 (.09) ***		1.09 (.10)***	1.09 (.10)***
Average Revisions per Question		0.34 (.17) *		0.41 (.26)	0.41 (.26)
Granularity		-0.19 (.10)		-0.06 (.14)	-0.05 (.14)
Probabilistic Training (dummy)		0.66 (.15) ***		0.65 (.19)***	0.63 (.19)***
Group Collaboration (dummy)		0.38 (.16) *		0.52 (.20)*	0.50 (.20)*
Targets for selection					
Numeracy			-0.00 (.10)	-0.04 (.09)	
Education Level			0.05 (.10)	-0.02 (.10)	
Raven's Progressive Matrices			0.12 (.11)	0.04 (.11)	
Cognitive Reflection Test			0.04 (.11)	0.04 (.11)	
Fox-Hedgehog			0.06 (.09)	0.02 (.09)	
Need for Cognition			0.12 (.10)	0.15 (.09)	
Additional controls					
Age	0.17 (.07)	-0.01 (.07)	0.43 (.10)***	0.16 (.10)	0.01 (.01)
Female (dummy)	-0.23 (.19)	0.01 (.18)	-0.21 (.24)	0.13 (.23)	0.12 (.23)
Superforecaster (dummy)	7.05 (.64) ***	5.56 (.59) ***	7.71 (.72)***	6.01 (.71)	6.11 (.71)***
Constant	3.64 (.09) ***	3.04 (.14) ***	3.85 (.11)***	2.94 (.18)***	2.57 (.31)***
N	1,821	1,821	1,307	1,307	1,307
R^2	0.32	0.41	0.37	0.45	0.45
AIC	9,547	9,299	6,905	6,733	6,725

Ordinary least squares regression predicting B^* thresholds for individual respondents. Non-binary independent variables standardized. Robust standard errors. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

†Model 5 only retains observations available in Models 3-4.

Model 1 demonstrates that a simple model featuring forecasting skill and our three controls predicted substantial variation in individual-level returns to precision (R^2=0.32). Model 2 shows that adding our other "Targets for Cultivation" variables substantially improved model fit (R^2=0.41). In particular, the variables for Number of Questions, Average Revisions per Question, Probabilistic Training and Groupwork were statistically significant predictors of individual-level returns to precision.[44] By contrast, Model 3 shows that our education and cognitive style variables predicted little individual-level variation in returns to precision when controlling for respondents' Brier Scores. None of the "Targets for Selection" variables approached statistical significance in Model 3.

When we examined all predictors together in Model 4, the Targets for Selection variables remained insignificant, and the Average Revisions per Question (*p*=0.12) variable lost statistical significance as well. Model 5 then replicates our analysis of the Targets for Cultivation using only observations for which we have data on all variables. Model 5 returned an R^2 value just 0.002 below that of Model 4, indicating how little predictive power the Targets for Selection added to our analysis of individual-level returns to precision.[45]

Throughout these five models, we found no evidence of a systematic correlation between gender and returns to precision. Model 3 suggested that older respondents might have been able to parse their probability estimates more reliably. This finding could be consistent with the idea that older respondents have more knowledge to apply to their predictions (or more time to devote to the tournament, especially for retired respondents), but the pattern did not persist across models.[46] Finally, and as expected, we found that "superforecasters" demonstrated unusually

[44] Adding a squared term for Number of Questions is statistically significant (*p*<0.01), but improves R^2 by less than 0.01. A model containing all targets for cultivation less Brier Score has a model fit of R^2=0.17 for the full sample and for the 1,307 observations for which we have full data.

[45] Estimating Model 1 in a sample with those same 1,307 observations only returns R^2 and AIC scores of 0.37 and 6,898, respectively.

[46] We also found that a dummy variable for "retired" respondents (as proxied by age cutoffs from 60-65) was not a statistically significant predictor of returns to precision.

high returns to precision. The average superforecaster could reliably parse her judgments into 11.3 bins (standard deviation 5.6), once again demonstrating how systems of qualitative expression are particularly constraining for high-quality analysts.

Discussion

Two principal conclusions emerge from this analysis. First, returns to precision correlated with factors that foreign policy analysts and organizations can feasibly cultivate. For example, GJP forecasters who received brief training sessions in probabilistic reasoning, or who collaborated in teams, demonstrated substantially higher returns to precision than their peers, even when controlling for respondents' Brier Scores. Given random assignment to training and to groups, our findings suggest that professional foreign policy analysts can replicate and presumably exceed this benefit. For example, analytic teamwork is much denser among national security professionals than it was among GJP groups who collaborated online. Similarly, foreign policy organizations have opportunities to train their analysts much more extensively than the simple, one-hour training modules that GJP respondents received.

We also found that respondents' experience making forecasts and their willingness to revise those forecasts consistently predicted higher returns to precision (though the latter finding fell outside the $p<.05$ threshold for statistical significance in some models). These findings provide additional grounds for optimism that professional forecasters could replicate and potentially exceed the returns to precision shown in GJP's data. Many national security professionals assess uncertainty on a daily basis over many years, and they have much more opportunity and incentive to refine and revise their forecasts in light of new information than did the GJP respondents (who revised their forecasts, on average, less than twice per question).

It is not surprising that Number of Questions predicted returns to precision among GJP respondents. Forecasters who registered more predictions were not only more experienced and more engaged in the competition, but they also provided more statistical power for calculating B^* thresholds, such that smaller rounding errors would register as being statistically significant. Our analyses cannot isolate how much of this correlation results from sample size versus gains from experience. Yet either interpretation has the same practical implication: the more forecasts analysts make, the more likely it becomes that coarsening those estimates will systematically sacrifice information. Given the vast quantity of judgments that national security officials make, along with the vast numbers of interviews and essays that make up the marketplace of ideas in foreign policy discourse, the relationship we observe between Number of Questions and returns to precision further emphasizes how GJP's data may understate the degree to which scholars, practitioners, and other foreign policy analysts could achieve meaningful returns to precision by quantifying their probability assessments.

The second principal takeaway from this analysis is that we see little evidence that returns to precision belong primarily to forecasters who are especially skilled in quantitative reasoning, who have special educational backgrounds, or who possess particular cognitive styles. Thus while many scholars divide foreign policy analysts into "mathematicians" and "poets," and though we see little reason to doubt the notion that foreign policy analysts range widely in terms of their reasoning styles and methodological preferences, our data suggest that when a broad range of forecasters take the time and effort to make precise forecasts, this consistently adds information to foreign policy analysis.

Conclusion

Uncertainty surrounds virtually every major foreign policy debate. As of this writing, for example, the United States public is engaged in sharp disagreements about the extent to which restricting immigration from specific Muslim-majority countries could reduce (or potentially exacerbate) the risk of terrorism. The permanent members of the United Nations Security Council are currently debating the chances that negotiated settlements might prevent Iran from gaining nuclear weapons, and whether that probability would be higher or lower under a more coercive approach. Debates over policy measures to combat climate change depend on diverging views about the risks that climate change poses, and to what extent regulations could feasibly mitigate the chances that those outcomes will occur. At the most general level, it is logically impossible to support a foreign policy decision without believing that its probability of success is sufficiently large to make expected benefits outweigh expected costs. For that reason, it makes little sense to ask *whether* foreign policy analysts should assess probability. The question is rather *how* they can do this in the most meaningful way possible.

We have seen throughout this article how many scholars and practitioners approach this issue with deep skepticism. Given the complexity of world politics and the inherent subjectivity of assessing uncertainty in this domain, many observers assume that even coarse probability estimates provide arbitrary detail instead of meaningful insight. This is a major reason why many scholars, practitioners, and pundits leave their assessments of uncertainty deliberately vague. The basis for this skepticism is clear. Many of the events that have shaped international affairs over the past two decades – the September 11, 2001 terrorist attacks, mistaken judgments of Iraq's presumed weapons of mass destruction programs, the 2008 financial crisis, the Arab Spring, the rise of ISIS, Brexit, and the election of Donald Trump – were outcomes that most foreign policy analysts either failed to see coming, or where experts confidently stated that the opposite would be true. Needless to say, our ability to predict world politics is much less accurate than we would like it to be.

Yet one could argue that this only makes it *more* important to debate key assumptions clearly and transparently – and, ultimately, the question of how much precision foreign policy analysts can justify is an empirical matter that has never before been tested directly. In this article, we provided the first systematic examination of this issue by drawing on a unique data set containing nearly one million forecasts, based on questions chosen by the U.S. Intelligence Community to be as relevant as possible to practical concerns. Our results demonstrate that foreign policy analysts can consistently assess probability with greater precision than what conventional wisdom supposes and what standard practices allow. We found no evidence that these returns to precision hinged on extreme forecasts, short time horizons, particular scoring rules, or question content. We also saw little indication that the ability to parse probabilities belonged primarily to respondents who possess special educational backgrounds or strong quantitative skills.

These results refute widespread skepticism regarding the value of precision in probability assessment. And while that skepticism has a longstanding academic pedigree, involving great scholars such as Popper, Keynes, and Mill, our research also speaks to eminently practical concerns. Our results suggest that the widespread practice of leaving probability assessments vague systematically sacrifices meaningful information from public discourse. This is not just a matter for intelligence analysts and military planners, but also for scholars, pundits, or any other observers who participate in the broader marketplace of ideas. Our data indicate that it would be possible to begin reaping these gains on an effectively immediate basis, simply by holding foreign policy analysts to higher standards of clarity and rigor.

Our data further suggest that skills in parsing probability assessments can be cultivated. Forecasters who were randomly assigned to receive just one hour of training in probabilistic reasoning achieved significantly higher returns to precision than did their peers. If intelligence agencies or other organizations prioritized such training efforts among their personnel, we expect that they could achieve even better performance than what we observed in our data. Showing that forecasters who worked in teams achieved better returns to precision is also grounds for optimism, in the sense that professional forecasters can achieve much denser (and perhaps more beneficial) collaborations than did GJP respondents who primarily communicated online and in their spare time.

Of course, enhancing predictive accuracy will not always improve decision quality. Yet the difficulty of anticipating where changes in informational quality are most likely to impact decision making is exactly why we believe that it is important to seek broad improvements in foreign policy analysis. When considering drone strikes or Special Forces missions, for example, decision makers continually wrestle with whether the available evidence is sufficiently certain to justify moving forward. In many cases, shifting a probability estimate from, say, seventy percent to eighty percent might not matter. But when policymakers encounter such decisions many times over, there are bound to be instances where such shifts in probability are critical. The fact that we cannot always know in advance where these differences will be most important is a strong justification for ensuring that analysts avoid discarding information unnecessarily.

Moreover, refining analytic standards for expressing probability is a far more cost-effective method for improving foreign policy analysis than other attempted reforms. In previous decades, for example, the U.S. government has repeatedly conducted large-scale organizational overhauls of its Intelligence Community despite ambiguous theoretical and empirical justifications for doing so (Betts 2007; Bar-Joseph and McDermott 2008; Pillar 2011). If such costly measures are justified on such a contested basis, it should also be desirable to implement guidelines for expressing estimative probabilities more precisely if this improves predictive accuracy.

Finally, while our article has focused on the domain of international relations, similar controversies about the value of precision surround assessments of uncertainty in almost any other area of high-stakes decision making. Medicine is a prime example. One of a physician's most important responsibilities is to communicate with patients about uncertain diagnoses and treatment outcomes. Yet medical professionals, like foreign policy analysts, are often reluctant to express probabilistic judgments explicitly (Gigerenzer 2002; Braddock et al. 1999). Similarly, the application of criminal justice in the United States revolves around assessing the vague probabilistic standard of guilt "beyond a reasonable doubt" (McAuliff 1982; Tillers and Gottfried 2006). And questions about the value of precision in communicating probability are also currently a prominent topic of debate among international climate scientists (Budescu, Broomell, and Por 2009).

This article offers a generalizable methodology showing how these disciplines can revisit their own basic skepticism about the value of probabilistic precision. Our methodology can also potentially be extended to estimate the value of precision when assessing other quantifiable aspects of uncertainty, such as how much a policy might cost.[47] And while empirical findings from one domain do not directly translate into others, foreign policy analysis is widely considered to be a field in which probability assessment is unusually difficult. International affairs involve a large number of variables that interact in nonlinear ways within contexts that are frequently unique. Foreign policy analysts generally lack access to broadly-accepted theoretical models or to large, well-behaved data sets for grounding their inferences: that is, in fact, exactly why many people doubt the value of precision when assessing probabilities in this domain. By comparison, analysts in professions such as medicine, law, and climate science often have much stronger bases for defining reference classes, estimating base rates, or employing analytic tools to assist with assessing uncertainty. If foreign policy analysts can reliably parse subjective probability estimates with numeric precision, this suggests that other disciplines may also benefit from scrutinizing their own conventional wisdom about the value of precision when assessing uncertainty.

[47] For a richer description of what "good judgment" might entail in foreign policy and other fields, see Renshon and Larson (2003).

References

Aristotle, *Nicomachean Ethics* tr. Terence Irwin (Indianapolis, Ind.: Hackett, 1985).

Arrow, Kenneth J. et al. 2008. “The Promise of Prediction Markets.” *Science*, 320: 877-878.

Arthur, W. Jr. et al. 1999. “College-Sample Psychometric and Normative Data on a Short Form of the Raven Advanced Progressive Matrices Test.” *Journal of Psychoeducational Assessment* 17 (4): 354-361.

Barnes, Alan. 2016. “Making Intelligence Analysis More Intelligent: Using Numeric Probabilities.” *Intelligence and National Security* 31 (1): 327-344.

Baron, Jonathan et al. 2015. “Why Does the Cognitive Reflection Test (Sometimes) Predict Utilitarian Moral Judgment (and Other Things)?” *Journal of Applied Research in Memory and Cognition* 4 (3): 265-284.

Bar-Joseph, Uri and Rose McDermott. 2008. “Change the Analyst and Not the System: A Different Approach to Intelligence Reform.” *Foreign Policy Analysis* 4 (2): 127-145.

Betts, Richard K. 2000. “Is Strategy an Illusion?” *International Security* 25 (2): 5-50.

-------. 2007. *Enemies of Intelligence: Knowledge and Power in American National Security*. New York: Columbia University Press.

Beyerchen, Alan. 1992/93. “Clausewitz, Nonlinearity, and the Unpredictability of War.” *International Security* 17 (3): 59-90.

Braddock, Clarence H. et al. 1999. “Informed Decision Making in Outpatient Practice,” *Journal of the American Medical Association* 282 (24): 2313-2320.

Brooks, Stephen. 1997. “Dueling Realisms.” *International Organization* 51 (3): 445-477.

Budescu, David V., Stephen Broomell, and Han-Hui Por. 2009. “Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change.” *Psychological Science* 20 (3): 299-308.

Bueno de Mesquita, Bruce. 2009. *The Predictioneer’s Game*. New York: Random House.

Cacioppo, J. T. and R. E. Petty, 1982. “The Need for Cognition.” *Journal of Personality and Social Psychology* 42 (1): 116-131.

von Clausewitz, Carl. 1832/1984. *On War* tr. Michael Howard and Peter Paret. Princeton, N.J.: Princeton University Press.

Connable, Ben. 2012. *Embracing the Fog of War*. Santa Monica, CA: Rand.

Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. “Clinical versus Actuarial Judgment.” *Science* 243: 1668-1674.

Dhami, Mandeep K. 2013. *Understanding and Communicating Uncertainty in Intelligence Analysis*. Report Prepared for H.M. Government, U.K.

-------, David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. “Improving Intelligence Analysis with Decision Science.” *Perspectives in Psychological Science* 10: 753-757.

Ellsberg, Daniel. 1961. “Risk, Ambiguity, and the Savage Axioms.” *Quarterly Journal of Economics* 75 (4): 643-669.

Fingar, Thomas. 2011. *Reducing Uncertainty: Intelligence Analysis and National Security*. Stanford, CA: Stanford Security Studies.

Friedman, Jeffrey A., Jennifer S. Lerner, and Richard Zeckhauser. Forthcoming. “Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals.” *International Organization*.

Friedman, Jeffrey A. and Richard Zeckhauser, 2012. “Assessing Uncertainty in Intelligence.” *Intelligence and National Security* 27 (6): 824-847.

Gardner, Daniel. 2011. *Future Babble: Why Pundits are Hedgehogs and Foxes Know Best*. New York: Plume.

Gat, Azar. 1989. *The Origins of Military Thought*. New York: Oxford University Press.

Gigerenzer, Gerd. 2002. *Calculated Risks*. New York: Simon and Schuster.

Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. Forthcoming. "The Behavioral Revolution and the Study of International Relations." *International Organization*.

Jervis, Robert. 1976. *Perception and Misperception in International Politics*. Princeton, N.J.: Princeton University Press.

-------. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, N.J.: Princeton University Press.

-------. 2010. *Why Intelligence Fails*. Ithaca, N.Y.: Cornell University Press.

Johnston, Rob. 2005. *Analytic Culture in the U.S. Intelligence Community*. Washington, D.C.: Center for the Study of Intelligence.

de Jomini, Antoine-Henri. 1838/2007. *The Art of War* tr. W. P. Craighill and G. H. Mendell. New York: Dover.

Kent, Sherman. 1964. "Words of Estimative Probability." *Studies in Intelligence* 8 (4): 49-65.

Keynes, John Maynard. 1921. *A Treatise on Probability*. London: Macmillan.

-------. 1937. "The General Theory of Employment." *Quarterly Journal of Economics* 51 (2): 209-223.

Lanir, Zvi and Daniel Kahneman. 2006. "An Experiment in Decision Analysis in Israel in 1975." *Studies in Intelligence* 50 (4): np.

Lowenthal, Mark M. 2006. *Intelligence: From Secrets to Policy,* 3rd ed. Washington, D.C.: CQ Press.

-------. 2008. "Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues?" *Intelligence and National Security* 23 (3): 303-315.

Mandel, David R. and Alan Barnes. 2014. "Accuracy of Forecasts in Strategic Intelligence." *Proceedings of the National Academy of Sciences* 111 (30): 10984-10989.

Marchio, James. 2014. “The Intelligence Community’s Struggle to Express Analytic Uncertainty in the 1970s.” *Studies in Intelligence* 58 (4): 31-42.

Mattis, James N. 2008. “USFJCOM Commander’s Guidance for Effects-based Operations.” *Parameters* 38 (3): 18-25.

McAuliff, C. M. A. 1982. “Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional Guarantees?” *Vanderbilt Law Review* 35: 1293-1335.

McDermott, Rose and Philip G. Zimbardo. 2007. “The Psychological Consequences of Terrorism Alerts” in Bruce Bongar et al., eds., *Psychology of Terrorism*. New York: Oxford University Press.

Meirowitz Adam and Joshua A. Tucker. 2004. “Learning from Terrorism Markets.” *Perspectives on Politics* 2 (2): 331-337.

Mellers, Barbara A. et al. 2014. “Psychological Strategies for Winning a Geopolitical Forecasting Tournament.” *Psychological Science* 25 (5): 1106-15.

-------. 2015a. “The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics.” *Journal of Experimental Psychology: Applied* 21 (1): 1-14.

-------. 2015b. “Improving Probabilistic Predictions by Identifying and Cultivating ‘Superforecasters.’” *Perspectives on Psychological Science* 10 (3): 267-281.

Mill, John Stuart. 1882. *A System of Logic, Ratiocinative and Inductive,* 8th ed. New York: Harper and Brothers.

National Intelligence Council. 2007. *Prospects for Iraq’s Stability*. Washington, D.C.: National Intelligence Council.

Nye, Joseph S., Jr. 1994. “Peering into the Future.” *Foreign Affairs* 73 (4): 82-93.

Pape, Robert A. 1997/98. “The Air Force Strikes Back.” *Security Studies* 7 (2): 191-214.

Peters, Ellen et al. 2006. “Numeracy and Decision Making.” *Psychological Science* 17 (5): 407-413.

Pillar, Paul. 2011. *Intelligence and U.S. Foreign Policy: Iraq, 9/11 and Misguided Reform.* New York: Columbia University Press.

Popper, Karl. 1972. *Objective Knowledge: An Evolutionary Approach*. London: Clarendon.

Rathbun, Brian C. 2007. "Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory." *International Studies Quarterly* 51 (3): 533-557.

Renshon, Stanley A. and Deborah Welch Larson eds. 2003. *Good Judgment in Foreign Policy: Theory and Application*. Lanham, Md.: Rowman and Littlefield.

Rieber, Steven. 2004. "Intelligence Analysis and Judgmental Calibration." *International Journal of Intelligence and CounterIntelligence* 17 (1): 97-112.

Satopää, Ville A., Jonathan Baron, Dean P. Foster, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. 2014. "Combining Multiple Probability Predictions Using a Simple Logit Model." *International Journal of Forecasting* 30: 344-356.

Schneider, Gerald, Nils Petter Gleditsch, and Sabine Carey. 2011. "Forecasting in International Relations." *Conflict Management and Peace Science* 20 (1): 5-14.

Shapiro, Jacob N. and Dara Kay Cohen. 2007. "Color Blind: Lessons from the Failed Homeland Security Advisory System." *International Security* 32 (2): 121-154.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, N.J.: Princeton University Press.

-------. 2009. "Reading Tarot on K Street," *The National Interest* 103: 57-67.

-------. 2010. "Second Thoughts about *Expert Political Judgment*." *Critical Review* 22 (4): 467-488.

------- and Daniel Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York: Crown.

------- and Barbara A. Mellers. 2011. “Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong.” *American Psychologist* 66 (6): 542-554.

Tillers, Peter and Jonathan Gottfried. 2006. “Case Comment – *United States* v. *Copeland*, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim That Proof Beyond a Reasonable Doubt Is Unquantifiable?” *Law, Probability, and Risk* 5 (2) 135-157.

U.S. Army. 2997. *Field Manual 101-5: Staff Organization and Operations*. Washington, D.C.: Department of the Army.

-------. 2009. *Field Manual 5-0: The Operations Process*. Washington, D.C.: Department of the Army.

U.S. Joint Forces Command. 2006. *Commander’s Handbook for an Effects-Based Approach to Joint Operations*. Norfolk, VA: Headquarters Joint Forces Command.

Ward, Michael D. 2016. “Can We Predict Politics? Toward What End?” *Journal of Global Security Studies* 1 (1): 80-91.

Wheaton, Kristan J. 2012. “The Revolution Begins on Page Five: The Changing Nature of NIEs.” *International Journal of Intelligence and CounterIntelligence* 25 (2): 330-349.

Wyden, Peter H. 1979. *Bay of Pigs: The Untold Story*. New York: Simon and Schuster.