*International Studies Quarterly* (2018) **62**, 410–422

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

Jeffrey A. Friedman
*Dartmouth College*

Joshua D. Baker, Barbara A. Mellers, and Philip E. Tetlock
*University of Pennsylvania*

and

Richard Zeckhauser
*Harvard University*

Scholars, practitioners, and pundits often leave their assessments of uncertainty vague when debating foreign policy, arguing that clearer probability estimates would provide arbitrary detail instead of useful insight. We provide the first systematic test of this claim using a data set containing 888,328 geopolitical forecasts. 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 finding does not depend on extreme probability estimates, short time horizons, particular scoring rules, or individual 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. More broadly, our findings advance long-standing debates over the nature and limits of subjective judgment when assessing social phenomena, showing how explicit probability assessments are empirically justifiable even in domains as complex as world politics.

Before President 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 30 percent. But when the Joint Chiefs conveyed this view to the president in writing, they stated only that "[t]his plan has a fair chance of success." The report's author, Brigadier General David Gray, claimed that "[w]e thought other people would think that 'a fair chance' would mean 'not too good.'" Kennedy, by contrast, interpreted the phrase as indicating favorable odds. Gray later concluded 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 kind of aversion to clear probabilistic reasoning is common throughout foreign policy analysis (Lanir and Kahneman 2006; Dhami 2013, 3–5; Marchio 2014; Barnes 2016, 328–39). Figure 1, for example, shows how the US Intelligence Community encourages analysts to communicate probability using qualitative phrases. US 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 US 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. Phrases like "a fair chance of success" would often be more precise than the language that policy advocates use to justify placing lives and resources at risk (Tetlock 2009; Gardner 2011, 118–41).

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,

Jeffrey A. Friedman is an Assistant Professor of Government at Dartmouth College

Joshua D. Baker is a Ph.D Candidate in Psychology & Marketing at the University of Pennsylvania

Barbara A. Millers is the I. George Heyman University Professor at the University of Pennsylvania

Philip E. Tetlock is the Leonore Annenberg University Professor at the University of Pennsylvania

Richard Zeckhauser is the Frank P. Ramsey Professor of Political Economy at Harvard University

*Authors' note:* 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. Support through ANR-Labex IAST is gratefully acknowledged. 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 US government.

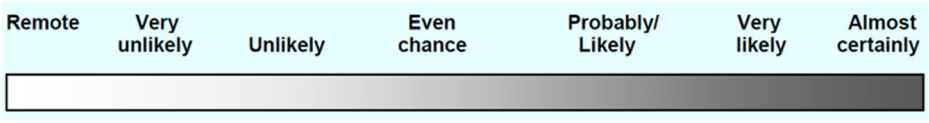
[1] Wyden writes that "in 1977, General Gray was still severely troubled about his failure to have insisted that figures be used. He felt that one of the key misunderstandings in the entire project was the misinterpretation of the word 'fair' as used by the Joint Chiefs."

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

[3] For scholarship on complexity in world politics, see Beyerchen (1992/93), Jervis (1997), and Betts (2000). On the connection between complexity theory and debates about strategic assessment, see Connable (2012, 1–36) and Mattis (2008, 18–19).

Friedman, Jeffrey A. et al. (2018) The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament. *International Studies Quarterly,* doi: 10.1093/isq/sqx078

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




*Director of National Intelligence, 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% |

*Intelligence Community Assessment 2017-01D, "Assessing Russian Activities and Intentions in Recent US Elections" (January 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.

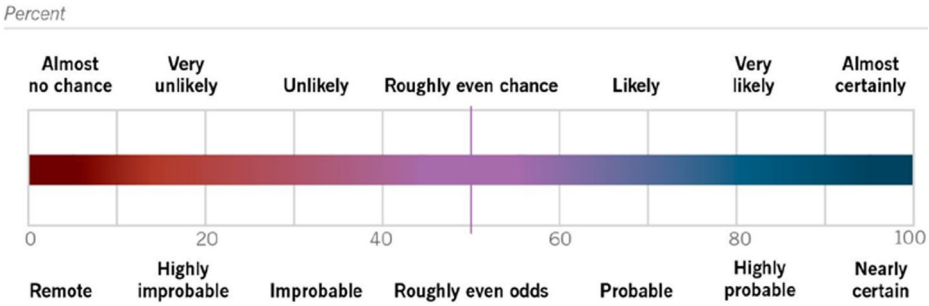



**Figure 1.** Guidelines for communicating probability assessments in the US Intelligence Community

clearer probability estimates convey arbitrary detail instead of useful insight.[4] Some scholars and practitioners even see explicit assessments of uncertainty as counterproductive, imparting illusions of rigor to subjective judgments, enabling analysts' natural tendencies toward 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. But does making such judgments more precise, in fact, also make them more accurate? To our knowledge, no one has tested this claim directly.

This article employs a 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 predict a forecaster's ability to parse probability assessments. 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. 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 this analysis in six parts. The first section frames debates about assessing uncertainty in world politics in relation to broader controversies about subjective judgment in the social sciences. The second section introduces our data set. The third section describes our empirical methodology. The fourth section shows how commonly

[4] Thus, 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 toward overconfidence, see Johnson (2004) and Tetlock (2005, 67–120). The US 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 (2017).

used qualitative expressions systematically sacrifice 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 in this field reflect subjective beliefs rooted in professional opinion, not algorithmic output (Tetlock 2010, 483). These are the cases in which analytic precision often seems hardest to justify. According to Mill (1882, 539), probability assessments "are of no real value" unless analysts derive them from large volumes of reliable data. Keynes (1937, 213–14) wrote that "[a]bout these matters, there is no scientific basis on which to form any calculable probability whatsoever. We simply do not know."

The notion that some concepts are inherently qualitative or otherwise resistant to precision has a long-standing pedigree in the social sciences. Popper (1972, 207) suggested that social phenomena 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 widespread articulation of this view states that foreign policy involves "nonlinear" dynamics, where small changes to a system's inputs can cause huge swings in that system's outputs (Beyerchen 1992/93, Jervis 1997; Betts 2000; Mattis 2008). This framework casts doubt on the notion that foreign policy analysts can draw anything beyond coarse distinctions when assessing uncertainty.

Meanwhile, a large number of empirical studies show that subject-matter experts often struggle to outperform simple algorithms when making probability estimates (Dawes, Faust, and Meehl 1989; Tetlock 2005). One explanation for this finding is that heuristics and biases warp the ways in which individuals perceive uncertainty (Jervis 1976; McDermott 1998; Yarhi-Milo 2014; Hafner-Burton, Haggard, Lake, and Victor 2017). The most consequential of these biases for our purposes is overconfidence (Johnson 2004; Tetlock 2005, 67–120). Coarsening probability estimates could actually *improve* predictive accuracy if this prevents foreign policy analysts from making their judgments too extreme.[6]

Yet, just because foreign policy analysts struggle to assess uncertainty, this does not mean it is desirable to leave their judgments vague. Several bodies of research suggest that foreign policy analysts may in fact possess a reliable ability to parse subjective probability assessments in detail. For instance, prediction markets can often extract meaningful, fine-grained probability estimates from the wisdom of crowds (Meirowitz and Tucker 2004; Arrow, Cropper, Gollier, Groom, Heal, Newell, and Nordhaus 2008). Mandel and Barnes (2014) provide evidence that geopolitical forecasts in Canadian intelligence reports are surprisingly well calibrated. In an even larger study, using predominantly nongovernmental respondents, Mellers, Stone, Murray, Minster, Rohrbaugh, Bishop, and Chen (2015b) identify a group of "superforecasters" who predicted international events with considerable accuracy. 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 judgments vague could still 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 prevents analysts from assessing uncertainty with meaningful detail. At a methodological level lie debates about the extent to which subjective judgment can sustain the kinds of precision that scholars generally reserve for algorithmic analyses. At a practical level, 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 systematically addressed this controversy.

### *Probabilistic Precision and Estimative Language*

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. Note that 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 0 percent and 100 percent). The guidelines for expressing uncertainty shown in Figure 1 each divide the number line into seven bins. Those guidelines reflect an implicit assumption that foreign policy analysts cannot 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, critics observed that intelligence analysts had framed their judgments using "estimative verbs" such as "we assess," "we believe," or "we judge."[7] The guidelines at the bottom of Figure 1 explain how estimative verbs indicate that judgments are uncertain. However, these terms

[6] 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.

[7] For example, the 2002 National Intelligence Estimate on *Iraq's Continuing Program for Weapons of Mass Destruction* (National Intelligence Council 2002) states the following: "[w]e assess that Baghdad has begun renewed production of [the chemical weapons] mustard, sarin, GF (cyclosarin), and VX." Then, "[w]e judge that all key aspects—R&D, production, and weaponization—of Iraq's offensive BW [biological weapons] programs are active." See Jervis (2010, 123–55) and Wheaton (2012) for critiques of this practice.

provide little information about what the relevant uncertainty entails, other than implying that a judgment 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*. Probability and confidence technically reflect different concepts. However, many foreign policy analysts appear to conflate these concepts or to use them interchangeably.[8] For example, the lexicon at the bottom of Figure 1 appears in an Intelligence Community Assessment describing Russian interference in the 2016 US presidential election (National Intelligence Council 2017). These guidelines explain 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 US democratic process. The report then assesses that Putin staged this intervention in order to help Donald Trump to defeat his opponent, Hillary Clinton. The Central Intelligence Agency and the Federal Bureau of Investigation placed "high confidence" in this judgment. The National Security Agency, by contrast, 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.[9]

Of course, analysts can divide the number line into however many bins they prefer when assessing uncertainty. Yet, most existing recommendations for expressing probability in 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).[10] The next two sections describe the data and method we use to evaluate how these systems of expressions influence the predictive accuracy of geopolitical forecasts.

## Data

Our study employs data gathered by the Good Judgment Project (GJP). The GJP 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. Forecasters logged responses to those questions online using numeric probability estimates.[11] Supplementary material provides extensive descriptions of GJP's data and methods.

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 contemporary foreign policy decision-making. IARPA made no attempt to ensure that these questions were suitable for econometric, game-theoretic, or other algorithmic analyses.[12] The main exception to the ecological validity of IARPA's question list was the requirement that each question be written precisely enough so that judges could eventually record its outcome.

This article focuses on the performance of individuals who registered at least twenty-five predictions in a given tournament year.[13] The resulting data set spans 1,832 individuals who registered 888,328 forecasts in response to 380 questions administered between 2011 and 2015. Participants tended to be males (83 percent) and US citizens (74 percent). Average age was forty. Sixty-four percent of respondents held a bachelor's degree. Fifty-seven percent had completed postgraduate training. The article's penultimate section explores the extent to which education and other individual attributes predict analysts' abilities to extract meaningful returns to precision.

The GJP randomly assigned forecasters to work alone or in collaborative teams. Another random subset of forecasters received a one-hour training module covering techniques for effective forecasting. This training module included topics such as base rates, cognitive biases, and extrapolating trends from data (Chang, Chen, Mellers, and Tetlock 2016).[14] These experimental conditions help to ground our analysis. We expect that untrained forecasters who worked alone should make the lowest-quality forecasts and that those forecasts should 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 collaborate closely with their peers and almost all receive some kind of formal training.[15] Forecasters who work in groups and who received training should therefore be more relevant for judging the abilities of professional foreign policy analysts.

At the end of each year, GJP identified the top 2 percent of performers as superforecasters. One of the GJP's principal findings was that superforecasters' predictions

[8] 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 50 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, 834–41) and Friedman and Zeckhauser (forthcoming).

[9] It is possible that the authors of this report intended to convey probability through the estimative verb "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.

[10] For descriptions of additional experiments with quantifying subjective probabilities in foreign policy analysis, see Nye (1994, 88–92), Lanir and Kahneman (2006), Marchio (2014), and Barnes (2016).

[11] The GJP 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.

[12] Indeed, the principal goal of the tournament was to encourage participants to develop whatever techniques they believed would be most effective for addressing a broad range of questions. The GJP 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, Baron, Foster, Mellers, Tetlock, and Ungar (2014).

[13] 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.

[14] We describe the purpose and results of this training in greater detail below when analyzing how returns to precision varied across individual forecasters. We also included a version of the GJP's training manual with this article's supplementary files.

[15] The US Intelligence Community, for example, sends incoming analysts into training programs that can last several months; the US military sends officers to multiple training programs (including masters'-level education for officers who reach the rank of colonel or commander).

generally remained superior to those of other respondents in subsequent tournament years.[16] 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 data are uniquely positioned to evaluate returns to precision in geopolitical forecasting due to the sheer volume of forecasts that the GJP 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 GJP did not randomize the response scale that forecasters employed. Thus, GJP data do not offer a true experimental comparison of numeric percentages versus words of estimative probability, confidence levels, or estimative verbs. Nonetheless, we do not believe that this threatens our inferences. In order to choose appropriate terms from Figure 1, for instance, foreign policy analysts must already determine where their judgments fall along the number line. Moreover, randomizing modes of expressing probability would introduce a fundamental measurement problem. When analysts use terms like "high confidence," there is no reliable way to know whether they mean probabilities closer to 70 percent or to 90 percent (Beyth-Merom 1982; Dhami and Wallsten 2005). Thus, we cannot tell whether a "high confidence" forecast was closer to the truth than a forecast of 80 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. The next section describes our strategy for doing so.

A second caveat for interpreting our results is that the GJP 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 multiyear time horizons. However, the next section demonstrates that our findings are robust across time horizons within GJP data.

Third, GJP only asked respondents to assess the probability of future events. Of course, 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. We expect that analysts should generally 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. Yet, we lack the data necessary to substantiate this claim directly. We therefore emphasize that our empirical analysis measures return to precision in geopolitical *forecasting*, which is a subset of foreign policy analysis writ large.

Finally, because the GJP's respondents volunteered their participation, we cannot say that these individuals comprise a representative sample of foreign policy analysts. Since the GJP gathered extensive information on its participants, however, we can examine how 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 methodology for estimating returns to precision 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. This section explains each stage of that process 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. If an analyst estimates a 70 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 70 percent likely to be true, we can see whether those conclusions were actually true roughly 70 percent of the time (Rieber 2004; Tetlock 2005, 13; Mandel and Barnes 2014, 10985).

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 scoring rule.[17] The Brier score measures the mean squared error of an assessor's judgments. This mean squared error is relative to the judgments that assessors could have made had they known the future with certainty. For example, consider a forecaster who predicts a 60 percent chance that Bashar al-Assad is ousted from Syria's presidency by the end of 2018. If Assad is ousted, then the forecaster's score for this estimate would be 0.16. If Assad remains, then the forecaster's score for this estimate would

[16] Mellers, Ungar, Baron, Ramos, Gurcay, Fincher, and Scott (2014), Mellers (2015a), and Mellers et al. (2015b), and Tetlock and Gardner (2015) describe the superforecasters in more detail.

[17] The Brier score is more appropriate for our purposes because of the severe penalties that the logarithmic scoring rule assigns to misplaced extreme estimates. Logarithmic scoring requires changing estimates of 0.00 and 1.00 (comprising 19 percent of our data points), since an error on these estimates imposes an infinite penalty.

**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%*** |

*Notes:* (1) Table 1 shows rounding errors for different groups of respondents, depending on the degree of imprecision to which we round their forecasts. (2) 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). (3) Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. (4) †Currently recommended by the Office of the Director of National Intelligence (see Figure 1).

be 0.36.[18] Since the Brier score measures judgmental error, lower Brier scores reflect better forecasts. In other words, lower Brier scores indicate that respondents assign higher probabilities to events that occur and lower probabilities to events that do not occur.

### *Step 2: Translating Numeric Forecasts into Corresponding Verbal Expressions*

We translate numerical forecasts into corresponding verbal expressions by rounding probability assessments to the midpoint of the bin that each verbal expression represents. For example, the Director of National Intelligence defines the phrase *even chance* as representing probabilities between 45 and 55 percent. Absent additional information, the expected value of a probability estimate falling within this range is 50 percent. In practice, a decision maker may combine this estimate with other information and prior assumptions to justify a prediction that is more than or less than 50 percent. However, saying that a probability is equally likely to fall anywhere within a range conveys the same expected value as stating that range's midpoint.[19]

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.[20] When forecasts fall on boundaries between bins, as with a forecast of 50 percent when $B = 2$, we randomize the direction of rounding.[21]

### *Step 3: Calculating "Rounding Errors"*

Our data set contains 888,328 forecasts. However, these forecasts are correlated within questions and within individuals who updated forecasts before questions closed for evaluation.[22] 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 888,328 forecasts to 380 forecasting questions. Evaluating individual forecasts returns similar estimates of returns to precision, albeit with inappropriate levels of statistical significance.[23]

We calculate *rounding errors* on forecasting questions by measuring proportional changes in Brier scores. For example, assume that the average Brier score for all untrained individuals on a particular question is 0.160. After rounding off this group's forecasts to the midpoints of three bins, we might find that its average Brier score climbs to 0.200. We would then say that rounding these estimates into three bins induced a rounding error of 25 percent. In the analysis below, we analyze both mean and median rounding errors.[24] Analyzing means and medians together ensures that outliers do not drive our inferences.

## 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 31 percent worse when rounded into two bins. Outliers do not drive this change, as the median rounding error is 22 percent. Even the worst-performing group

[18] 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 $[(0.00-0.60)^2+(1.00-0.40)^2]/2=0.36$.

[19] 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).

[20] We also experiment with rounding probabilities to the empirically weighted mean of each range, so that if responses to a particular question clustered near 100 percent, then we would round numeric estimates to a point that was higher than if those responses clustered near 70 percent. Our findings are robust to this alternative approach.

[21] 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.

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

[23] 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.

[24] 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.

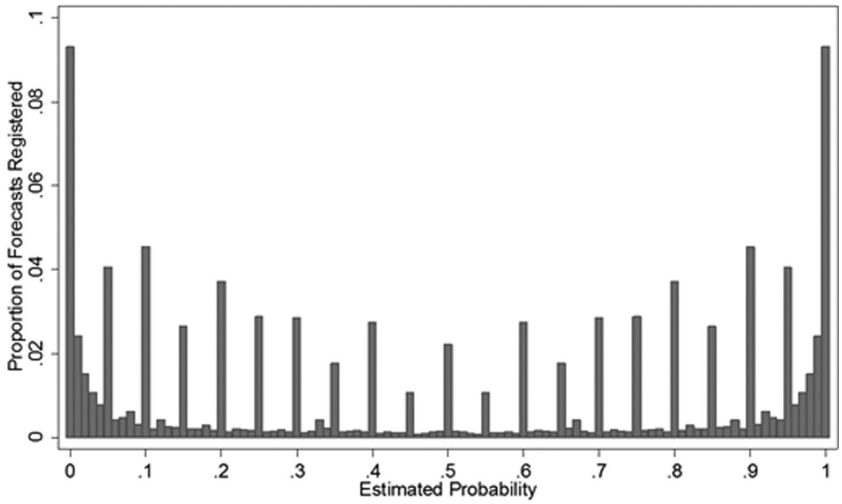



**Figure 2.** Histogram of forecasts in GJP data

of forecasters, untrained individuals, incurred an average rounding error of 15 percent when we rounded their forecasts to estimative verbs. The penalty for superforecasters was far worse, with average rounding errors exceeding 500 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 10 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, every subgroup in our analysis encountered consistent losses of predictive accuracy when we rounded forecasts according to the lexicon currently recommended by the US Director of National Intelligence. The National Intelligence Council's current guidelines for expressing uncertainty induce greater variance; rounding errors here tend to be larger but also less consistent.[25] 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. Lowenthal (2008, 314), a scholar with three decades of experience in the US Intelligence Community, observes that "[n]o one has yet come up with any methodologies, machines, or thought processes that will appreciably raise the Intelligence Community's [performance]."[26] Fingar (2011, 34, 130), formerly the US Intelligence Community's top analyst, writes that "[b]y 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 arguing that foreign policy analysts could consistently improve the accuracy of their judgments by making their probability estimates more precise.

### *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 and time horizon. More extreme probability estimates reflect greater certainty, which may correlate with easier questions. Nearer-term events may also be easier for analysts to predict. We additionally examine the extent to which questions pertaining to particular regions or topics might have generated special returns to precision. We will show that none of these factors drives our main findings.

Figure 2 presents a histogram of GJP forecast values.[27] As a general rule, GJP forecasters assigned estimates at intervals of five percentage points.[28] This pattern alone is important. It indicates that, when GJP forecasters were left without restrictions on how finely to parse their predictions, they naturally 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 divide forecasts into seven bins according to the National Intelligence Council guidelines shown in Figure 1. We separately examine 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 find mixed results from rounding nonsuperforecasters' most extreme estimates—Table 2 shows how coarsening these estimates degrade their accuracy on average but improve 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.

[25] 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.

[26] Tetlock and Mellers (2011, 6–9) and Tetlock (2010, 481–83) further explain how, even when foreign policy analysts possess empirically validated methods, those methods are rarely tested directly against rival approaches.

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

[28] Forty-nine percent of forecasts in the data set are multiples of 0.10, and twenty-five percent of forecasts are additional multiples of 0.05.

**Table 2.** Rounding errors across the probability scale (Brier and logarithmic scoring)

| *Group* | | *Remote (0.00–0.14)* | *Very unlikely (0.15–0.28)* | *Unlikely (0.29–0.42)* | *Even chance (0.43–0.56)* | *Likely (0.57–0.71)* | *Very likely (0.72–0.85)* | *Almost certain (0.86–1.0)* |
|---|---|---|---|---|---|---|---|---|
| *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%*** |

*Notes*: (1) Table 2 examines how rounding forecasts into seven equal bins influences predictive accuracy for forecasts within different segments of the number line. (2) 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). (3) Statistical significance: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 2 also shows that our results do not hinge on the Brier score's particular properties. When we recalculate rounding errors using a logarithmic scoring rule,[29] we again find that superforecasters exhibit reliable returns to precision in every category and that rounding sacrifices predictive accuracy for nonsuperforecasters in every category besides the extremes. [30]

### *Returns to Precision across Time Horizons*

To assess how returns to precision varied across time horizons, we code 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 5 percent probability or no less than 95 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 judgments. We divided all other forecasts into three categories with equal numbers of observations.[31] In the supplementary material, we show how our findings are generally consistent across each time period that we analyze. We thus see no indication that our conclusions rely on easy questions with short time horizons where foreign policy analysts can justify special precision.

### *Returns to Precision across Question Types*

To determine how returns to precision vary across topics, we generate question-specific estimates of returns to precision. We derive these estimates by calculating respondents' Brier scores after rounding each forecast into progressively larger numbers of bins. We defined each question's *threshold of estimative precision* ($B^*$) as the smallest number of bins for which the median rounding error was not statistically greater than 0.[32] This represents another deliberately conservative method for describing returns to precision, given how we set these $B^*$ thresholds at the lowest possible value at which 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.

Among the 375 questions that we analyzed in this way,[33] the mean $B^*$ threshold is 6.1 bins, with a standard deviation of 4.4 bins. These $B^*$ thresholds exceed seven bins for 42 percent of GJP's questions. We thus find that employing official guidelines for expressing uncertainty using qualitative expressions systematically reduces 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 for which forecasters happened to make particularly informative estimates. Supplementary material further shows how classifying questions according to eleven regions and fifteen topics provided little traction for predicting $B^*$ thresholds. In other words, our data do not indicate that returns to precision belong to specific kinds of geopolitical forecasts.

## Examining Variation across Individuals

The previous section shows that forecasters can assess subjective probabilities more precisely than what the conventional wisdom and standard procedures allow. This section examines how individuals vary in their ability to achieve returns to precision. We focus particularly on whether this appears to be an innate capability or whether it correlates with attributes that foreign policy analysts could feasibly cultivate.

International relations scholars often question the numeracy of foreign policy analysts and decision makers. Kent (1964) famously divided intelligence analysts into "mathematicians" and "poets" based on their comfort with quantitative expressions. Nye (1994, 88–92), Johnston (2005, 17–28) and Marchio (2014, 36–39) argue that this cultural divide is a consistent trend in the US foreign policy

[29] 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.

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

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

[32] 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.

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

community.[34] Meanwhile, empirical research on political forecasting by Tetlock (2005) or Tetlock and Gardner (2015) has found 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 individual-level returns to precision, then that would have two important practical implications. First, it would suggest that "poets" and low-numeracy analysts might justifiably opt out of making clear probability assessments. Second, this would suggest that organizations that wish to improve 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 improve their probability assessments through training and feedback.[35] For example, the GJP found that just one hour of randomly assigned training had a substantial, persistent impact on improving forecasters' Brier scores (Chang et al. 2016). Moreover, even if foreign policy analysts initially struggle to translate their subjective judgments into explicit language, this problem might diminish over time as analysts become more comfortable using numeric expressions. If that is true, then it suggests that returns to precision might correlate with variables we call *targets for cultivation*. This 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.

The remainder of this section examines six targets for cultivation and six targets for selection. The first of our targets for cultivation variables is forecasting skill, as measured by each respondent's median Brier score. Higher-quality forecaster should incur greater penalties from having their forecasts rounded. 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 good 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. These guidelines reflect the implicit assumption that foreign policy analysts should be able to discriminate *among* rough categories, but that they should not be able to draw meaningful differences *within* those categories.

We use five additional variables to 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 measure the *granularity* of each respondent's forecasts by calculating 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.[36]

*Probabilistic training* takes a value of 1 if GJP trained the forecaster in probabilistic reasoning. We expect that respondents who received this training 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 from teammates' assessments.

The first two of our targets for selection variables capture respondents' education. *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).[37] 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 (Cokely, Galesic, Schulz, and Ghazal 2012, 45–46). Respondents who are better able to reason numerically might parse probabilities more effectively.[38] In principle, organizations can cultivate both education and numeracy. However, these attributes are substantially more expensive to increase than the effort and training variables described above.

GJP data also include several indices of cognitive style. Higher scores on *Raven's Progressive Matrices* indicate greater abstract reasoning abilities and fluid intelligence (Arthur, Paul, and Sanchez-Ku 1999). Higher scores on the expanded Cognitive Reflection Test (*Expanded CRT*) indicate a greater propensity to suppress misleading intuitions in favor of more accurate, deliberative answers (Baron, Scott, Fincher, and Metz 2015). Higher scores on the *Fox-Hedgehog* scale reflect respondents' self-assessed tendency to rely on ad hoc reasoning versus simplifying frameworks (Mellers, Stone, Atanasov, Rohrbaugh, Metz, Ungar, and Bishop 2015a, 94). Higher scores on *need for cognition* reflect respondents' self-assessed preference for effortful mental activity (Cacioppo and Petty 1982).[39] Throughout our analysis, we also control for age, for gender, and for whether a respondent was designated as a superforecaster in any tournament year. Supplementary files contain full descriptive statistics for each of the variables described in this section.

[34] On numeracy and probability assessment more generally, see Peters, Västfjäll, Slovic, Mertz, Mazzocco, and Dickert (2006) and Dieckmann, Slovic, and Peters (2009).

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

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

[37] If a respondent participated in multiple years of the forecasting competition, we average education values across years.

[38] GJP changed numeracy tests between years two and three of the competition. We standardize numeracy test results so that they represent comparable indices. If a respondent participated in multiple years of the forecasting competition, we average numeracy values across years.

[39] If a respondent participated in multiple competition years, we average values across years. GJP changed CRT tests after year two of the competition, so we standardize 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 (0.15)*** | –1.57 (0.16)*** | –1.80 (0.24)*** | –1.74 (0.26)*** | –1.77 (0.26)*** |
| Number of questions | | 1.15 (0.09)*** | | 1.09 (0.10)*** | 1.09 (0.10)*** |
| Average revisions per question | | 0.34 (0.17)* | | 0.41 (0.26) | 0.41 (0.26) |
| Granularity | | –0.19 (0.10) | | –0.06 (0.14) | –0.05 (0.14) |
| Probabilistic training (dummy) | | 0.66 (0.15)*** | | 0.65 (0.19)*** | 0.63 (0.19)*** |
| Group collaboration (dummy) | | 0.38 (0.16)* | | 0.52 (0.20)* | 0.50 (0.20)* |
| *Targets for selection* | | | | | |
| Numeracy | | | –0.00 (0.10) | –0.04 (0.09) | |
| Education level | | | 0.05 (0.10) | –0.02 (0.10) | |
| Raven's progressive matrices | | | 0.12 (0.11) | 0.04 (0.11) | |
| Cognitive reflection test | | | 0.04 (0.11) | 0.04 (0.11) | |
| Fox-Hedgehog | | | 0.06 (0.09) | 0.02 (0.09) | |
| Need for cognition | | | 0.12 (0.10) | 0.15 (0.09) | |
| *Additional controls* | | | | | |
| Age | 0.17 (0.07) | –0.01 (0.07) | 0.43 (0.10)*** | 0.16 (0.10) | 0.14 (0.09) |
| Female (dummy) | –0.23 (0.19) | 0.01 (0.18) | –0.21 (0.24) | 0.13 (0.23) | 0.12 (0.23) |
| Superforecaster (dummy) | 7.05 (0.64) *** | 5.56 (0.59) *** | 7.71 (0.72)*** | 6.01 (0.71) | 6.11 (0.71)*** |
| Constant | 3.64 (0.09) *** | 3.04 (0.14) *** | 3.85 (0.11)*** | 2.94 (0.18)*** | 2.97 (0.18)*** |
| 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 |

*Notes*: (1) Ordinary least squares regression predicting $B^*$ thresholds for individual respondents. Nonbinary independent variables standardized. Robust standard errors in parentheses. (2) Statistical significance: $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. (3) †Model 5 only retains observations available in Models 3–4.

### *Analyzing Individual-Level Returns to Precision*

Table 3 presents ordinary least squares regressions predicting forecasters' $B^*$ thresholds as a function of individual-level attributes. We standardize nonbinary 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.

Model 1 demonstrates that a simple model featuring forecasting skill and our three controls predicts substantial variation in individual-level returns to precision ($R^2 = 0.32$). Model 2 shows that adding our other targets for cultivation variables substantially improves model fit ($R^2 = 0.41$). In particular, the variables for number of questions, average revisions per question, probabilistic training, and groupwork are statistically significant predictors of individual-level returns to precision.[40] By contrast, Model 3 shows that our measures of education and cognitive style predict little individual-level variation in returns to precision when controlling for respondents' Brier scores. None of the targets for selection variables approach statistical significance in Model 3.

When we examine all predictors together in Model 4, the targets for selection variables remain insignificant. Model 5 replicates our analysis of the targets for cultivation using only observations for which we have data on all variables. Model 5 returns an $R^2$ value just 0.002 below that of Model 4. This comparison indicates how little predictive power the targets for selection add to our analysis of individual-level returns to precision.[41]

Across these models, we find no evidence of a systematic correlation between gender and returns to precision. Model 3 suggests 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 does not persist across models.[42] Finally, and as expected, we find that superforecasters demonstrated unusually 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*

Our analysis in this section sustains two principal conclusions. First, we find that returns to precision correlate 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, demonstrate 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 could replicate and

[40] 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.

[41] 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.

[42] 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.

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 find that respondents' experience making forecasts and their willingness to revise those forecasts consistently predict higher returns to precision (though the latter finding fell short of the $p < 0.05$ threshold 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. Professional foreign policy analysts also have much more opportunity and incentive to refine and revise their forecasts in light of new information than did GJP respondents, who revised their forecasts less than twice per question, on average.

It is unsurprising that number of questions predicts returns to precision among GJP respondents. Forecasters who registered more predictions provide more statistical power for calculating $B^*$ thresholds. This allows smaller rounding errors to register as being statistically significant. Thus, to the extent that our "number of questions" variable correlates with returns to precision in our data set, we cannot say how much this results from sample size versus gains from experience. Yet, it is important to note that 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 produce—along with the vast numbers of interviews and essays that make up the marketplace of ideas in foreign policy discourse more generally—the relationship we observe between questions answered 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 when assessing uncertainty.

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 every major foreign policy debate. As of this writing, for example, the US public is sharply divided in assessing the extent to which restricting immigration from Muslim-majority countries could reduce (or potentially exacerbate) the risk of terrorism. One of the foremost controversies facing the United Nations Security Council concerns the extent to which economic sanctions can reduce the probability that North Korea will continue expanding its nuclear arsenal. Debates over policy responses to climate change revolve around different perceptions of the risks that climate change poses and of the extent to which regulations could feasibly reduce those risks. At the broadest level, it is logically impossible to support a high-stakes decision without believing that its probability of success is large enough 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 assess probability in the most meaningful way possible.

We have seen throughout this article how many scholars and practitioners are deeply skeptical of probability assessment. It is easy to understand why this is the case. Many of the events that have shaped world politics over the past two decades—such as 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 political analysts failed to see coming or cases in which experts confidently stated that the opposite would be true. Our ability to predict world politics is clearly less accurate than we would like it to be.

This article nevertheless shows that it is a mistake to believe that probabilistic reasoning is meaningless in world politics or to think there is no cost to leaving these judgments vague. By examining nearly one million geopolitical forecasts, we find that foreign policy analysts could consistently assess probability with numeric precision. We find that rounding off these forecasts into qualitative expressions—including qualitative expressions currently recommended for use by US intelligence analysts—systematically sacrifices predictive accuracy. We see no evidence that these returns to precision hinged on extreme forecasts, short time horizons, particular scoring rules, or question content. We also see little indication that the ability to parse probabilities belonged primarily to respondents who possess special educational backgrounds or strong quantitative skills.

These findings speak to both academic and practical concerns. Great scholars such as Popper, Keynes, and Mill have all expressed doubts about the value of assessing subjective probability. Aristotle himself argued that justifiable precision declines as questions become more complex. Yet, even if that is true, it does not tell us where the frontier of justifiable precision lies in foreign policy analysis or in any other discipline. That is ultimately an empirical question, and to our knowledge, this article represents the first attempt to address that question directly. The results of our analysis are relevant not only for intelligence analysts and military planners, but also for scholars, pundits, and any other participants in the broader marketplace of ideas. In short, our data indicate that it is possible to improve the quality of foreign policy discourse on a widespread and immediate basis, simply by raising standards of clarity and rigor for assessing uncertainty.

Of course, improving assessments of uncertainty will not always improve the quality of decisions. 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 by a few percentage points might not matter. But when policy makers encounter these decisions many times over, these shifts will inevitably prove critical in some instances. The fact that we cannot always know in advance where these differences will be most important is exactly why analysts should avoid discarding information unnecessarily. Refining standards for assessing uncertainty is also far less costly than the kinds of controversial organizational overhauls that the US government regularly undertakes to improve the quality of foreign

policy analysis (Rovner, Long, and Zegart 2006; Betts 2007, 124–58; Bar-Joseph and McDermott 2008; Pillar 2011, 293–330).

Finally, while our article focuses 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. For 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, often prove reluctant to express probabilistic judgments explicitly (Braddock, Edwards, Hasenberg, Laidley, and Levinson 1999, 2318). Climate scientists similarly debate the value of precision in communicating predictions to the public (Budescu, Broomell, and Por 2009). The application of criminal justice in the United States revolves, in large part, around the ways in which jurors interpret the vague probabilistic standard of guilt beyond a reasonable doubt (Tillers and Gottfried 2006).

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 be extended to estimate the value of precision when assessing other quantifiable aspects of uncertainty, such as how much a policy might cost.[43] And while empirical findings from one domain do not directly translate into others, foreign policy analysis is widely considered an unusually difficult arena for probability assessment. International affairs involve a large number of variables that interact in nonlinear ways within highly specific contexts. Foreign policy analysts generally lack access to broadly accepted theoretical models or to large, well-behaved data sets for grounding their inferences. 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.

## Supplementary Information

Supplementary information is available at http://scholar.dartmouth.edu/friedman and the *International Studies Quarterly* data archive.

## References


Aristotle, . 1985. *Nicomachean Ethics*. Translated by Terence Irwin. Indianapolis, IN: Hackett.

Arrow, Kenneth J., Maureen L. Cropper, Christian Gollier, Ben Groom, Geoffrey M. Heal, Richard G. Newell, and William D. Nordhaus, et al. 2008. "The Promise of Prediction Markets." *Science* 320 (5878): 877–78.

Arthur, Winfred, Don S. Paul, and Maria L. Sanchez-Ku. 1999. "College-Sample Psychometric and Normative Data on a Short Form of the Raven Advanced Progressive Matrices Test." *Journal of Psychoeducational Assessment* 17 (4): 354–61.

Barnes, Alan. 2016. "Making Intelligence Analysis More Intelligent: Using Numeric Probabilities." *Intelligence and National Security* 31 (1): 327–44.

Baron, Jonathan, Sydney Scott, Katrina Fincher, and S. Emlen Metz. 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): 265–84.

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): 127–45.

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.

Beyth-Merom, Ruth. 1982. "How Probable is Probable? A Numerical Translation of Verbal Probability Expressions." *Journal of Forecasting* 1 (3): 257–69.

Braddock, Clarence H., Kelly A. Edwards, Nicole M. Hasenberg, Tracy L. Laidley, and Wendy Levinson. 1999. "Informed Decision Making in Outpatient Practice." *Journal of the American Medical Association* 282 (24): 2313–20.

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, John T., and Richard E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42 (1): 116–31.

Chang, Welton, Eva Chen, Barbara Mellers, and Philip E. Tetlock. 2016. "Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments." *Judgment and Decision Making* 11 (5): 509–26.

Cokely, Edward T., Mirta Galesic, Eric Schulz, and Saima Ghazal. 2012. "Measuring Risk Literacy: The Berlin Numeracy Test." *Judgment and Decision Making* 7 (1): 25–47.

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 (4899): 1668–74.

Dhami, Mandeep K. 2013. *Understanding and Communicating Uncertainty in Intelligence Analysis*. London, UK: H.M. Government.

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. "Improving Intelligence Analysis with Decision Science." *Perspectives in Psychological Science* 10 (6): 753–57.

Dhami, Mandeep K., and Thomas Wallsten. 2005. "Interpersonal Comparison of Subjective Probabilities." *Memory and Cognition* 33 (6): 1057–68.

Dieckmann, Nathan F., Paul Slovic, and Ellen M. Peters. 2009. "The Use of Narrative Evidence and Explicit Likelihood by Decisionmakers Varying in Numeracy." *Risk Analysis* 20 (10): 1473–88.

Ellsberg, Daniel. 1961. "Risk, Ambiguity, and the Savage Axioms." *Quarterly Journal of Economics* 75 (4): 643–69.

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

Friedman, Jeffrey A., Jennifer S. Lerner, and Richard Zeckhauser. 2017. "Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals." *International Organization* 71 (4), doi.org/10.1017/S0020818317000352.

Friedman, Jeffrey A., and Richard Zeckhauser. 2012. "Assessing Uncertainty in Intelligence." *Intelligence and National Security* 27 (6): 824–47.

———. "Analytic Confidence and Political Decision Making: Experimental Evidence from National Security Professionals." *Political Psychology* (forthcoming), doi.org/10.1111/pops.12465.

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

Hafner-Burton, Emilie M., Stephan Haggard, David A. Lake, and David G. Victor. 2017. "The Behavioral Revolution and the Study of International Relations." *International Organization* 71 (S1): S1–S31.

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

———. 1997. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press.

———. 2010. *Why Intelligence Fails*. Ithaca, NY: Cornell University Press.

Johnson, Dominic D. P. 2004. *Overconfidence and War*. Cambridge, MA: Harvard University Press.

Johnston, Rob. 2005. *Analytic Culture in the US Intelligence Community*. Washington, DC: Center for the Study of Intelligence.

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

Keynes, John Maynard. 1937. "The General Theory of Employment." *Quarterly Journal of Economics* 51 (2): 209–23.


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

Lanir, Zvi, and Daniel Kahneman. 2006. "An Experiment in Decision Analysis in Israel in 1975." *Studies in Intelligence* 50 (4), https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol50no4/an-experiment-in-decision-analysis-in-israel-in-1975.html

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

———. 2008. "Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues?" *Intelligence and National Security* 23 (3): 303–15.

Mandel, David R., and Alan Barnes. 2014. "Accuracy of Forecasts in Strategic Intelligence." *Proceedings of the National Academy of Sciences* 111 (30): 10984–89.

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.

McDermott, Rose. 1998. *Risk-Taking in International Politics: Prospect Theory in American Foreign Policy*, Ann Arbor: University of Michigan Press.

McDermott, Rose, and Philip G. Zimbardo. 2007. "The Psychological Consequences of Terrorism Alerts." In *Psychology of Terrorism*, edited by Bruce Bongar, Lisa M. Brown, Larry A. Beutler, James N. Breckinridge and Philip G. Zimbardo. New York: Oxford University Press.

Meirowitz, Adam, and Joshua A. Tucker. 2004. "Learning from Terrorism Markets." *Perspectives on Politics* 2 (2): 331–37.

Mellers, Barbara A., Eric Stone, Pavel Atanasov, Nick Rohrbaugh, Emlen S. Metz, Lyle Ungar, and Michael M. Bishop, et al. 2015a. "The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics." *Journal of Experimental Psychology: Applied* 21 (1): 1–14.

Mellers, Barbara A., Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, and Eva Chen, et al. 2015b. "Improving Probabilistic Predictions by Identifying and Cultivating 'Superforecasters.'" *Perspectives on Psychological Science* 10 (3): 267–81.

Mellers, Barbara A., Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, and Sydney E. Scott, et al. 2014. "Psychological Strategies for Winning a Geopolitical Forecasting Tournament." *Psychological Science* 25 (5): 1106–15.

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, DC: National Intelligence Council.

———. 2002. *Iraq's Continuing Programs for Weapons of Mass Destruction*. Washington, DC: National Intelligence Council.

———. 2017. *Assessing Russian Activities and Intentions in Recent U.S. Elections*. Washington, DC: National Intelligence Council.

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

Peters, Ellen, Daniel Västfjäll, Paul Slovic, C. K. Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. "Numeracy and Decision Making." *Psychological Science* 17 (5): 407–13.

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

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

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.

Rovner, Joshua, Austin Long, and Amy B. Zegart. 2006. "How Intelligent Is Intelligence Reform?" *International Security* 30 (4): 196–208.

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 (2): 344–56.

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): 121–54.

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

———. 2009. "Reading Tarot on K Street." *National Interest* 103: 57–67.

———. 2010. "Second Thoughts about *Expert Political Judgment*." *Critical Review* 22 (4): 467–88.

Tetlock, Philip E., and Daniel Gardner. 2015. *Superforecasting: The Art and Science of Prediction*. New York: Crown.

Tetlock, Philip E., and Barbara A. Mellers. 2011. "Intelligent Management of Intelligence Agencies: Beyond Accountability Ping-Pong." *American Psychologist* 66 (6): 542–54.

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) 135–57.

US Army. 1997. *Field Manual 101–5: Staff Organization and Operations*. Washington, DC: Department of the Army.

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

US 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): 330–49.

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

Yarhi-Milo, Keren. 2014. *Knowing the Adversary: Leaders, Intelligence, and Assessment of Intentions in International Politics*. Princeton, NJ: Princeton University Press.