The Value of Precision in Geopolitical Forecasting: Empirical Foundations for Intelligence Analysis and Foreign Policy Decision Making

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*

Paper prepared for NBER Summer Institute, Program on the Economics of National Security in Cambridge, Mass. (July 20, 2015)

This draft: July 5, 2015 (11,928 words)
All comments welcome

Abstract

Uncertainty surrounds almost all important foreign policy decisions. Scholars and practitioners disagree about how precisely foreign policy analysts should assess this uncertainty, however, particularly when it comes to communicating subjective probability estimates. Common proposals include "estimative verbs," "confidence levels," "words of estimative probability," and quantitative expressions. Evaluating these proposals requires understanding how reliably analysts can parse probabilities when predicting geopolitical events. We provide the first systematic analysis of this subject by analyzing a data set of 764,448 forecasts. We round numerical probability estimates to different degrees of (im)precision and examine how this affects predictive accuracy. Our data indicate that qualitative expressions of probability systematically sacrifice meaningful information in foreign policy analysis. These findings do not depend on extreme probability estimates, short time horizons, or particular question types. At the individual level, returns to precision correlate primarily with forecasting skill, effort, experience, and training as opposed to numeracy, education, or cognitive style. These results suggest that returns to precision can be cultivated, and that it is possible to improve the informational value of intelligence analysis and other forms of geopolitical forecasting simply by using clearer language. Our methodological approach generalizes to other fields, such as medicine and finance, where subjective probability assessments play a crucial role in making consequential decisions.

Acknowledgments

Pavel Atanasov, David Budescu, Shrinidhi Kowshika Lakshmikanth, Angela Minster, Brendan Nyhan, Michael Poznansky, Jonah Schulhofer-Wohl, Sarah Stroup, Lyle Ungar, and Thomas Wallsten provided valuable input. This work has also greatly benefited from feedback provided in seminar presentations at Dartmouth College, Middlebury College, the University of Pennsylvania, and the University of Virginia.

The Value of Precision in Geopolitical Forecasting

Before President John F. Kennedy authorized the Bay of Pigs invasion in 1961, he asked the Joint Chiefs of Staff to assess the plan's feasibility. When the Chiefs reported that "This plan has a fair chance of success," Kennedy took this to be an optimistic assessment. Yet the report's author, Brigadier General David Gray, claims that "We thought other people would think that 'a fair chance' would mean 'not too good.'" Gray believed that his imprecise language enabled a strategic blunder.[1]

At the same time, many foreign policy scholars and practitioners believe that explicit probability assessments represent unjustifiable detail. Foreign policy analysis usually involves subjective judgment. As with analysts in many other disciplines, foreign policy officials often find it uncomfortable to discuss their subjective beliefs precisely, especially if this involves quantification.[2] In 2011, for instance, President Barack Obama's advisers assigned numerical percentages to the chances that Osama bin Laden was living in Abbottabad, Pakistan. Estimates reportedly ranged from 35 percent to 95 percent. Obama himself questioned whether these judgments "disguised uncertainty as opposed to actually providing you with more useful information."[3]

Yet this paper demonstrates that numeric probability assessments actually *do* provide more useful information than estimative verbs, confidence levels, words of estimative probability and other commonly-used qualitative terminology. Using a data set of 764,448 geopolitical forecasts collected by the Good Judgment Project in collaboration with the U.S. Intelligence Community, we show how rounding off numerical estimates of subjective probability to different degrees of (im)precision systematically sacrifices predictive accuracy. These findings do not depend on extreme probability estimates, short time horizons, or particular question types. At the individual level, we find that returns to precision correlate primarily with forecasting skill, effort, experience, and training, as opposed to numeracy, education, or cognitive style. These results suggest that returns to precision in geopolitical forecasting can be cultivated, and that it is possible to improve a wide range of intelligence estimates and foreign policy debates simply by expressing probability using clearer language.

We present this analysis in seven parts. Section 1 describes debates about expressing uncertainty in foreign policy analysis, couched in broader controversies about subjective probability assessment writ large. Sections 2 and 3 describe our data and our basic empirical approach. Section 4 evaluates common modes of expressing probability in geopolitical forecasting. Section 5 examines how returns to precision vary across individuals, and Section 6 examines variation across question types. Section 7 concludes by discussing implications for foreign policy analysis as well as for other fields, such as medicine, where scholars and practitioners also debate proper means for expressing uncertainty.

---

[1] Wyden 1979, 88-90.

[2] John Maynard Keynes (1937, 214) was a particular skeptic of subjective probability, arguing that "About these matters, there is no scientific basis on which to form any calculable probability whatsoever."

[3] Bowden 2012, 160.

Section 1. Expressing Probability in Foreign Policy Analysis

Some scholars and practitioners are so pessimistic about assessing probability in foreign policy analysis that they recommend avoiding such assessments altogether. The U.S. Intelligence Community (IC) deliberately eschews long-term predictions in its *Global Trends* reports on the grounds that these predictions would be effectively meaningless. Similarly, Thomas Fingar (formerly Chair of the National Intelligence Council) writes that "prediction is not – and should not be – the goal of strategic analysis…. The goal is to identify the most important streams of developments, how they interact, where they seem to be headed, what drives the process, and what signs might indicate a change of trajectory."[4]

Such views reflect assumptions that that world politics is so complex that attempts to predict it are effectively meaningless.[5] Yet predictability is a matter of degree. Karl Popper argued that analytic 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."[6] International affairs may be more "cloudlike" than many other disciplines. Yet it is ultimately an empirical question as to how finely foreign policy analysts can parse probabilistic assessments. Scholars have yet to address this empirical question directly.

Questions about the proper level of precision for expressing probability have become especially important within the U.S. Intelligence Community over the past decade. A prominent critique of the 2002 National Intelligence Estimate (NIE) on *Iraq's Continuing Programs for Weapons of Mass Destruction* was that its authors failed to express the uncertainty surrounding their judgments in appropriate detail. The Silberman-Robb Commission thus concluded that analysts must find better ways "to explain to policymakers degrees of certainty in their work" and "strongly urge[d] that such assessments of certainty be used routinely and consistently throughout the [Intelligence] Community."[7] The 2004 Intelligence Reform and Terrorism Prevention Act attempted to mandate such reform by requiring analysts to "properly caveat and express uncertainties or confidence in analytic judgments."[8] Yet there is no consensus on what "properly expressing" uncertainty entails, and there are several common proposals and practices to consider:

Estimative verbs. Phrases such as "we judge," "we estimate," or "we assess" are perhaps the most common way to express probability in intelligence. For example, the 2002 *Iraq* NIE states: "We assess that Baghdad has begun renewed production of [the chemical weapons] mustard, sarin, GF (syclosarin), and VX." Then: "We judge that all key aspects – R&D, production, and weaponization – of Iraq's offensive BW [biological weapons] programs are active." Though estimative verbs indicate that these judgments are uncertain, such phrasings do not parse uncertainty any further than implying that these conclusions are likely to be true.

[4] Fingar 2011, 53, 74. For related views, see MacEachin 1995 and Davis 1997.
[5] For example, Beyerchen 1992/93.
[6] Popper 1972, 207. An analogous concept in intelligence studies is the division of problems into "puzzles," "mysteries," and "complexities."
[7] Silberman-Robb Commission (2005, 419, 409).
[8] IRTPA Section 1019(b)(2)(A).

Confidence levels. Intelligence analysts frequently express judgments with "low," "moderate," or "high" confidence. Though likelihood and confidence are different concepts, intelligence analysts often use both terms to communicate probability.[9] For example, the 2007 NIE on *Iran: Nuclear Capabilities and Intentions* concludes: "We judge with high confidence that in fall 2003, Tehran halted its nuclear weapons program…. We assess with moderate confidence Tehran had not restarted its nuclear weapons program as of mid-2007…. We continue to assess with low confidence that Iran probably has imported at least some weapons-usable fissile material."

Words of estimative probability. Recent NIEs include front matter defining "words of estimative probability" (WEPs), which allow analysts to express probability qualitatively, yet more finely than what confidence levels allow.[10] Figure 1 presents three such spectrums. Over time, this guidance has become increasingly specific. In 2007, the WEP spectrum was expanded from five to seven terms. In 2015, the Director of National Intelligence (DNI) ordered that these terms represent unequal ranges of the probability spectrum, allowing analysts to identify extreme probabilities more precisely.[11]

[Figure 1]

Numerical expressions. Though rare in published intelligence analysis,[12] the debate over bin Laden's location shows how quantitative expressions of probability appear in important settings. Many observers advocate the broader use of quantitative probability, including numerical percentages, bettor's odds (e.g., 5-to-1), and frequency representations (e.g., 1-in-10).[13]

In principle, "Words of Estimative Probability" spectrums have been recommended doctrine for the U.S. Intelligence Community since 2007. However, this guidance has not been followed consistently. For example, even though the 2007 *Iran* NIE contained the seven-step spectrum of WEPs shown in Figure 1, its Key Judgments expressed probability in several ways, including estimative verbs ("Tehran's decision to halt its nuclear weapons program suggests it is less determined to develop nuclear weapons than we have been judging since 2005"), confidence levels ("We assess with high confidence that until fall 2003, Iranian military entities were working under government direction to develop nuclear weapons"), words of estimative probability ("We assess centrifuge enrichment is how Iran probably could first produce enough fissile material for a weapon"), confidence levels *and* words of estimative probability ("We judge with moderate confidence Iran probably would be technically capable of producing enough HEU [highly enriched uranium] for a weapon sometime during the 2010-2015 time frame"), and probabilistic language with no clear definition ("We cannot rule out that Iran has acquired from abroad – or will acquire in the future – a nuclear weapon or enough fissile material for a weapon").

Moreover, the DNI's guidance for assessing probability does not apply to analysis outside the IC. For instance, U.S. military doctrine currently guides planners to assess risks using "five

[9] For more on the conceptual distinction between likelihood and confidence, and on how these terms are often conflated in U.S. intelligence analysis, see Friedman and Zeckhauser 2012 and Friedman and Zeckhauser 2015.
[10] See Wheaton 2012 on the origins of WEP guidelines.
[11] Intelligence Community Directive 203 (ICD-203), *Analytic Standards* (January 2015).
[12] Friedman and Zeckhauser 2012 reviewed 379 declassified intelligence estimates published between 1964-94, finding that four percent of key judgments used quantitative expressions of probability.
[13] Nye 1994, Schrage 2005, Marchio 2014, Barnes 2015.

levels of probability – frequent, likely, occasional, seldom, and unlikely."[14] Meanwhile, foreign policy decision makers presumably range widely in terms of how they use and interpret qualitative terms such as "likely" or "a good chance."[15]

There are several reasons to favor greater consistency in expressing estimative probability. Consistency facilitates analysts' abilities to communicate and compare their views.[16] Clear guidelines also promote accountability.[17] Critics allege that foreign policy analysts use "waffle words" to avoid making predictions that appear clearly mistaken after the fact. Without clear standards, it is difficult to know when vague language represents genuine analytic misgivings as opposed to other incentives.[18] And most importantly for the purposes of this paper, if vague expressions systematically sacrifice information in foreign policy analysis, then consistent guidance can help to ensure that this information is not lost. Given how uncertainty surrounds nearly all important intelligence estimates and foreign policy debates, even marginal gains in this area could bring major aggregate benefits.

Returns to precision

We define *returns to precision* as the degree to which estimating probabilities more precisely increases predictive accuracy. All else being equal, we expect this relationship to be positive and concave. Yet we see no theoretical basis for predicting where returns to precision become negligible. Without empirical analysis, it is impossible to know the extent to which numerical assessments of probability may be more informative than "words of estimative probability," "confidence levels," or other common proposals.

Of course, there are other questions to consider in debates about expressing probability in foreign policy analysis. Making judgments more precise would make it more difficult for analysts to agree on contentious issues. In the IC's time-constrained environment, this is a nontrivial concern. Yet airing disagreements can also be productive in revealing discrepancies among analysts' views and in encouraging careful reasoning.

For example, the CIA initially provided President Obama with an estimate stating that there was a "strong possibility" that bin Laden was living in Abbottabad in spring 2011. (This is another example of how guidelines for expressing probability using WEPs are not consistently followed in important cases.) As the President pushed analysts to define their judgments more explicitly, CIA Deputy Director Michael Morell provided an estimate of 60 percent, while another CIA official provided an estimate of 95 percent. President Obama asked his advisers to explain this disparity. Morell argued that while many counterterrorism officials had understandably grown confident in their targeting abilities, his own priors were shaped by the

[14] Field Manual 5-19, *Composite Risk Management*, paragraph 1-23.
[15] On variations in interpreting qualitative expressions of probability, see Beyth-Merom 1982, Mosteller and Youtz 1990, and Wallsten and Budescu 1995. In national security analysis specifically, see Kent 1964, Wark 1964, Johnson 1973, and Wallsten, Shlomi, and Ting 2008.
[16] Thus the Silberman-Robb Commission report (2005, 409) argued that "A structured Community program must be developed to teach rigorous tradecraft and to inculcate common standards for analysis so that, for instance, it means the same thing when two agencies say they assess something 'with a high degree of certainty.'"
[17] Tetlock and Mellers 2011.
[18] Jervis 2006, 15 suggests that the lack of standards for expressing probability may help to explain flaws in the 2002 Iraq weapons of mass destruction NIE.

IC's mistaken estimates on Iraq's presumed weapons of mass destruction programs, and he was inclined to be more skeptical of judgments based on circumstantial evidence. In this way, unpacking a vague estimate ("a strong possibility") revealed important disagreements while raising fundamental issues about analytic and cognitive constraints in forming judgments under uncertainty.[19]

Though we do no analyze these issues directly, connecting them to practical concerns requires making assumptions about returns to precision. The argument that resolving disagreements about estimative probability is "too difficult" can only be made in relation to the benefits that additional precision would bring. If vague expressions of probability consistently sacrifice meaningful information, then it would be difficult to justify excluding that information on the grounds that analysts are prone to arguing about estimative language. Similarly, the argument that analysts should be encouraged to justify why their probability assessments differ by, say, 10 percentage points implicitly assumes that such differences are not just random noise.

Other relevant issues in broader debates about communicating probability concern how decision makers interpret foreign policy analysis. The distinction between words and numbers is especially significant in this literature. Numerical expressions are unambiguous, whereas even if intelligence estimates include front matter defining words of estimative probability, there is no guarantee that this is how decision makers will intuitively process qualitative language.[20] Advocates of qualitative expression counter that numbers convey inappropriate impressions of scientific rigor. Asking analysts to express both likelihood and confidence could alleviate this problem. But more importantly for our purposes here, one cannot say when analysts are expressing probability more precisely than their capabilities allow without first knowing how precisely analysts can parse probabilities to begin with. Once again, advancing debates about expressing probability requires evaluating assumptions about returns to precision.

Section 2. Forecasting Data from the Good Judgment Project

To our knowledge, this paper provides the first empirical analysis of returns to precision in geopolitical forecasting. Our analysis relies on data gathered by the Good Judgment Project (GJP). GJP began in 2011 as part of several large-scale geopolitical forecasting tournaments sponsored by the Intelligence Advanced Research Projects Activity (IARPA). A total of 1,718 unique individuals[21] registered 764,448 forecasts[22] in response to 288 questions administered between 2011-13.

[19] Morell 2015, 156-161. See Jervis 2006 on the dangers of neglecting the role of such prior assumptions, and see Kent 1964 for further analysis of how vague estimates can conceal meaningful disagreement among analysts.
[20] For striking empirical results to this effect, see Budescu, Por, and Broomell 2012; Budescu, Por, Broomell, and Smithson 2014.
[21] Participation required a bachelor's degree or higher and completion of a battery of psychological and political knowledge tests that took about two hours. Participants tended to be males (85 percent) and U.S. citizens (77 percent). Average age was 39. Sixty-four percent of respondents had post-graduate training. For an overview of GJP, see Mellers et al. 2014 and Mellers et al. 2015b. For policy implications, see Tetlock et al. 2014. For individual difference analyses, see Mellers et al. 2015a. For discussions of statistical method, see Satopää, Baron, et al. 2014, and Satopää, Jensen, et al. 2014.

IARPA and GJP collaborated in writing forecasting questions to ensure their relevance to intelligence analysis.[23] Questions covered issues such as the likelihood of different 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 by a particular date. Respondents recorded estimates on GJP's website using numeric probabilities. They could update forecasts as often as they wished before questions closed for assessment.

To discover how elicitation and training methods influenced forecasting quality, GJP randomly assigned forecasters to work alone or in collaborative teams. Random subsets of forecasters received a one-hour online training module covering various techniques for effective forecasting, such as defining base rates, avoiding cognitive biases, and extrapolating trends from data. This produced four categories of respondents: trained individuals, untrained individuals, trained individuals working in groups, and untrained individuals working in groups. After the first year's tournament, GJP identified the top 2 percent of performers as "superforecasters." Superforecasters remained consistently superior to other GJP respondents in subsequent tournament years.[24]

GJP's data are uniquely well-suited to evaluating empirical claims about the relationship between estimative precision and predictive accuracy in geopolitical forecasting due to the sheer volume of forecasts collected and because of IARPA's efforts to ensure the tournament's relevance to intelligence analysis. Nevertheless, we note three principal caveats for interpreting our results.

First, GJP did not randomize the response scale which forecasters employed. Thus GJP does not provide a true experimental comparison of numerical percentages versus WEPs, confidence levels, or estimative verbs. Yet we do not believe that this is a threat to our inferences. In order to choose appropriate WEPs from Figure 1, for instance, analysts must first determine where their judgments fall on the number line. Though several scholars have explored the ways in which analysts intuitively employ verbal expressions of probability,[25] all of the proposals discussed in Section 1 require approximate numerical reasoning in order to be employed consistently.

Moreover, randomizing modes of expressing probability would generate a fundamental measurement problem, in that when analysts use words like "high confidence," there is no reliable way to know whether they meant probabilities more like 70 percent or 90 percent. Thus we cannot tell whether a "high confidence" forecast was closer to the truth than a forecast of 80

[22] We only use forecasts which GJP respondents recorded numerically. GJP also administered a prediction market, but we do not use that data here because it only allows respondents to indicate whether they believe the probability of an event is higher or lower than the market's existing estimate.
[23] The only intentional exception to ecological validity was the requirement that each question be written sufficiently precisely so that outcomes could be judged clearly after the fact. See Marrin 2012 and Mandel and Barnes 2014 on the degree to which intelligence assessments pass this "clairvoyance test." Exploring a sample of 2,897 Canadian intelligence forecasts, for example, Mandel and Barnes found that 33 percent of predicted outcomes were too vague to score.
[24] Mellers et al. 2014. Data on superforecasters' performance only comprise forecasts from Years 2 and 3 of the competition, so that the definition of high performers is based solely on prior experience.
[25] See the sources in note 15.

percent when predicting an outcome that occurred.[26] For these reasons, "rounding off" numerical forecasts in a manner that is consistent with different modes of expression is the most straightforward way to estimate returns to precision for our purposes. We present our method for doing so in more detail below.

A second caveat for interpreting our results is that 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 72 days (standard deviation 83 days) before it was evaluated. By contrast, some intelligence reports, such as the *Global Trends* series mentioned above, consider scenarios in the more distant future. GJP data cannot directly evaluate the relationship between estimative precision and predictive accuracy on such long-term forecasts. However, we show in Section 4 that our substantive findings are generally robust across time horizons within GJP data. At the very least, we demonstrate that our general findings on returns to precision are not driven by short-term forecasts that critics might say are easier to address than the questions facing intelligence analysts writ large.

Third, GJP only asked respondents to make forecasts, but intelligence analysts and foreign policy officials also make probabilistic statements about current or past states of the world, as in debates about Osama bin Laden's location or the status of Iran's nuclear program. Generally speaking, we expect that analysts find it more difficult to parse probabilities when making forecasts, as forecasting requires assessing imperfect information while also anticipating contingencies that have not yet developed. This assumption implies that our findings should be conservative in identifying returns to precision when estimating probabilities in international affairs. Since this assumption is conjecture, however, we emphasize that our empirical results pertain directly to *predictive* accuracy and to geopolitical *forecasting*, which is a subset of intelligence and foreign policy analysis writ large.

Section 3. Measuring Estimative Precision and Predictive Accuracy

We measure predictive accuracy using Brier Scores, and then show in supplementary material that our results are robust to logarithmic and spherical scoring rules.[27] Brier Scores are a function of predicted probabilities (p_n) and observed outcomes (Y_n), where Y_n takes the value of 1 when outcome n occurs and 0 when it does not. Brier Scores measure mean squared errors across assessments within a forecasting problem: $(1/N) \cdot \sum_1^N (Y_n - p_n)^2$, where N is the number of possible outcomes.

[26] Mandel and Barnes 2015 show that it is possible to evaluate the predictive accuracy of qualitative forecasts, but the analysis here depends on clear counterfactuals as to what forecasters would have chosen to report in numerical formats.

[27] Proper scoring rules give analysts their highest expected payoffs when they report their true beliefs. Brier scoring, logarithmic scoring, and spherical scoring are the three most common proper scoring rules. We use the Brier Score as our main measure because logarithmic scoring assigns severe penalties to extreme probability forecasts (including $-\infty$ for mistaken forecasts of zero percent, which appear several times in GJP data), and because spherical scoring is much less well-understood than Brier scoring. Steyvers et al. (2014) propose a Bayesian signal detection method for evaluating forecasts. This method offers many advantages over traditional scoring rules, but it does not assess forecasters' calibration, which is crucial for evaluating returns to precision in our context.

For example, consider a response to the question, "Will Bashar al-Assad be ousted from Syria's presidency by the end of 2016?" There are two possible outcomes here: either Assad is ousted, or he remains. Say our forecaster predicts a 60 percent chance that Assad is ousted and a 40 percent chance he remains. If Assad is ousted, then the forecaster's score for this prediction would be $[(1-0.60)^2+(0-0.40)^2]/2=0.16$. If Assad remains, then the forecaster's score for this prediction would be $[(0-0.60)^2+(1-0.40)^2]/2=0.36$. Lower Brier Scores reflect better forecasting, indicating that respondents assign higher probabilities to events that occur and lower probabilities to events that do not occur.

We translate numerical forecasts into corresponding verbal expressions by rounding to the midpoint of the "bin" that each verbal expression comprises. For example, if analysts use the five-step "words of estimative probability" spectrum in Figure 1, then stating that some outcome has an "even chance" of occurring implies that its predicted probability could fall anywhere between 40 and 60 percent. Absent additional information, the expected value of a probability estimate falling inside this range is 50 percent.

In practice, a decision maker may combine this estimate with other information and prior assumptions to justify a prediction higher or lower than this.[28] Logically speaking, 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. We generalize this approach by dividing the number line into B bins, then rounding forecasts to the midpoint of the bin in which they fall. When forecasts fall on boundaries between bins (e.g., a forecast of 20 percent when $B=5$), we randomize the direction of rounding.[29]

Using forecasting questions as the unit of analysis

The most straightforward way to estimate returns to precision would be to calculate Brier Scores for a sample of forecasts, to round those forecasts to different degrees of (im)precision, and then to recalculate Brier Scores. Yet this approach treats each forecast in our data set as an independent observation. This is inappropriate, as forecasts are highly correlated within questions posed by GJP.

We thus take the forecasting question to be our unit of analysis.[30] We define a subset of forecasters to evaluate, such as all forecasters, superforecasters, or trained forecasters working in groups. We then calculate an *aggregate Brier Score* for that group on each forecasting question using the formula $x_{\gamma j}=mean_{i\in\gamma}[mean_{k\in K_{ij}}(Brier_{ijk})]$, where $mean(\cdot)$ is the mean of a

[28] For example, even though the word "remote" can in principle mean anything from 0 to 20 percent under the five-bin system of WEPs, decision makers might anticipate that when analysts use this term, they are usually attempting to convey a probability closer to zero. (This is presumably one of the problems that the DNI's new WEP spectrum intends to solve). To address this issue, we re-examined our findings by rounding estimates to the expected probability of forecasts fall within each bin. This alternative reduces rounding errors for most groups of forecasters, but still leaves statistically significant losses of accuracy in a manner consistent with the findings we present below.
[29] Though the current WEP spectrum defined in ICD-203 defines "remote" and "almost certain" as comprising assessments of 0.01-0.05 and 0.95-0.99, respectively, we also included GJP forecasts of 0.0 and 1.0 in these categories for the purposes of our analysis.
[30] In supplementary material, we demonstrate that evaluating individual forecasts leads to similar estimates of returns to precision, albeit with inappropriately small *p*-values when making comparisons.

vector; γ is a subset of GJP forecasters; i is a forecaster; j is a forecasting question; k is a day in the forecasting tournament; K_{ij} is the set of all forecasts made by forecaster i on question j while the question remained open;[31] $Brier_{ijk}$ is the Brier Score for an estimate made by a given forecaster on a given question on a given day; and $x_{\gamma j}$ is a question-level point-estimate of forecast accuracy among forecasters γ on question j.

This method represents a deliberately conservative approach to assessing statistical significance, because it reduces our maximum sample size from 764,448 forecasts to 288 forecasting questions.[32]

Calculating "rounding errors"

We calculate *rounding errors* on forecasting questions by measuring the proportional change in Brier Score that occurs when we round forecasts into bins of different widths. Thus, we define $\tilde{x}_{\gamma jB}$ as a question-level point-estimate of forecasting accuracy among forecasters γ on question j using forecasts rounded to the midpoints of B bins. To estimate the error with rounding γ's estimates on question j into confidence levels, for example, we would thus calculate $(\tilde{x}_{\gamma jB=3} - x_{\gamma j})/x_{\gamma j}$.

Calculating proportional changes in predictive accuracy helps to alleviate the asymmetrical penalties imposed by rounding in different regions of the probability scale. Moreover, in the analysis below, we describe both the mean and the median rounding errors that correspond to shifting forecasts. This helps to ensure that when we estimate the degree to which rounding probabilistic assessments impacts their predictive value, our findings represent consistent losses of information and not just high-leverage observations.[33]

Predictive accuracy and decision quality

Our empirical analysis focuses on improving predictive accuracy. Though increases in predictive accuracy will not always drive improvements in decision quality, this is no reason not to seek gains wherever possible, especially since the costs of increasing estimative precision are far lower than the costs of other attempted intelligence reforms. The U.S. government has repeatedly ordered large-scale organizational overhauls of its Intelligence Community despite ambiguous theoretical and empirical justifications for doing so.[34] If such costly measures are

[31] With a maximum of one forecast per day, recorded as a forecaster's most recent estimate prior to midnight, U.S. Eastern Time.

[32] An additional advantage of our aggregation method is that, by averaging across days during which a question remained open, we reduce the influence of forecasts made just before a closing date. In Section 4 we present additional evidence that these "lay-up" forecasts are not driving our results.

[33] As described below, we also measure statistical significance using both traditional comparisons of means and Wilcoxon signed-rank tests. Wilcoxon tests are insensitive to accuracy-score rescaling up to a positive affine transformation. Thus, while the magnitudes of the presented rounding errors are liable to change under alternative strictly proper scoring regimes, the substantive conclusions of our nonparametric analyses will remain the same..

[34] For skepticism about organizational reform, see Posner 2005, Betts 2007, and Bar-Joseph and McDermott 2008, Pillar 2011.

justified on such a contested basis, then it should also be desirable to implement guidelines for expressing estimative probabilities more precisely if this improves predictive accuracy.

We are aware of no systematic analysis explaining where changes in predictive accuracy are most likely to influence decision outcomes. Yet this is another reason to seek broad improvement wherever possible. When considering drone strikes or hostage rescue missions, for example, decision makers continually wrestle with whether the intelligence is sufficiently certain to move forward. In many cases, shifting a probability estimate from, say, 55 to 60 percent might not matter. But when policymakers encounter such decisions many times over, there are bound to be instances where small shifts in probability are critical. The fact that we cannot always know *ex ante* where small shifts in those probabilities will be most important is a strong justification for ensuring that analysts avoid unnecessarily discarding information across the board. And in the next section, we demonstrate that rounding GJP forecasts according to common modes of expression in intelligence analysis systematically impairs their predictive accuracy.

Section 4. Rounding Errors Across Modes of Expression

Table 1 shows how rounding forecasts to different degrees of (im)precision reduces their predictive accuracy. These data reveal that rounding numerical forecasts to "confidence levels" or "estimative verbs" substantially reduces the value of GJP forecasts. On average, GJP forecasts became 29.2 percent worse when we round them into two bins.[35] This change is not driven by outliers, as the median rounding error is 20.1 percent. Even the worst-performing group of forecasters, untrained individuals, incurs an average rounding penalty of more than 15 percent when rounding to "estimative verbs." For superforecasters, this penalty is far worse, with an average rounding error of over 500 percent. We also see large rounding penalties when we shift GJP respondents' forecasts to "confidence levels": on average, this level of imprecision degrades forecast accuracy by more than 10 percent, and substantially more for high-performing forecasters.

[Table 1]

Using "words of estimative probability" recovers some, but not all, of these losses. Across all GJP forecasts, we see rounding errors of roughly 1-2% when rounding to either of the seven-bin spectrums recently employed by the U.S. Intelligence Community. Superforecasters continue to suffer substantially greater losses in both formats. The DNI's newest doctrine is an improvement over its predecessor in that it imposes lower rounding errors than the previous system of seven, equally-spaced bins. Nevertheless, all subgroups in our analysis suffer losses of accuracy that are statistically significant at the $p<.001$ threshold when we round their forecasts according to the DNI's current recommendation.[36]

[35] This is the average rounding error across all 288 questions in our data set, when we round all forecasts in our data set from numerical estimates into two bins.

[36] Table 1 shows that the DNI's new guidance imposes rounding errors that are smaller, but in many cases more statistically significant than rounding errors under the previous WEP spectrum with seven equal bins. This is because, as we demonstrate below, the new DNI spectrum must expand bins in the middle of the spectrum in order to achieve additional precision at the extremes. This makes a majority of forecasts worse, even if the average rounding error declines.

We estimate the statistical significance of these patterns in two ways: paired-sample t-tests when assessing mean changes in Brier Scores, and paired-sample Wilcoxon signed-rank tests when assessing the median change in Brier Score. The signed-rank analysis is especially informative in this context, as it relies on only the ordinal properties of forecast accuracy, and is thus insensitive to the nonlinear properties of strictly proper scoring rules. No matter what loss function we assign to calculating "rounding errors" (e.g., whether we use Brier Scores, logarithmic scores, or any other formula) the signed-rank test will report the same result, based solely on the proportion of forecasts that lose value due to rounding. Altogether, rounding forecasts to seven bins according to current DNI guidance reduces performance on 72 percent of questions in our data set.[37] This finding is highly statistically significant ($p<.001$) and indicates how qualitative expressions of probability consistently prevent foreign policy analysts from reaching their full potential.

Such comparisons are especially meaningful in relation to the difficulty that the IC generally faces in evaluating methods of intelligence estimation. Mark Lowenthal, an intelligence scholar with three decades' experience in the IC, observes that "No one has yet come up with any methodologies, machines or thought processes that will appreciably raise the Intelligence Community's [performance]."[38] Thomas Fingar, formerly the IC's top analyst, writes that "By and large, analysts do not have an empirical basis for using or eschewing particular methods."[39] By contrast, our results *do* provide an empirical basis for expressing probabilities more precisely than what current IC practice entails. Geopolitical forecasting may be subjective, but our data indicate that when GJP participants responded to questions posed by the IC, their views are systematically more informative when evaluated at higher degrees of estimative precision.

Returns to precision across the number line

We now examine whether there are specific kinds of forecasts where respondents consistently extract larger (or smaller) returns to precision. It is particularly important to determine whether returns to precision appear primarily when making "easy" forecasts. Two main indicators of forecasting ease are the size of the forecast (more extreme probabilities reflect greater certainty, which should correlate with easier questions) and time horizons (as respondents should find it easier to predict events in the nearer-term). We address these issues in turn. Our results show that GJP respondents extract returns to precision on a broad range of forecasts.

[Figure 2]

Figure 2 presents a histogram of GJP forecast values. The diamonds represent probability estimates made by superforecasters. The bars represent estimates made by all other respondents.[40] In general, GJP forecasters appear to be most comfortable assigning estimates at intervals of five percentage points. This alone is an important result, because it indicates that

[37] Using the older WEP spectrum with 7 evenly-spaced bins reduces predictive accuracy on 68 percent of questions. Confidence levels and estimative verbs each reduce predictive accuracy on 86 percent of questions in our data.
[38] Lowenthal 2008, 314.
[39] Fingar 2011, 34, 130. Cf. Tetlock and Mellers 2011, 8.
[40] The histogram does not reflect how long those estimates remained active before respondents revised them or before questions closed. This is relevant for scoring performance, but not for evaluating the degree of granularity which forecasters tended to employ when registering their beliefs.

when left without restrictions on how granular to make their forecasts, GJP respondents appear to prefer expressing probabilities much more finely than common qualitative expressions allow.[41]

Figure 2 also demonstrates that GJP forecasters were especially willing to make granular forecasts when predicting probabilities close to zero. Since low-probability, high-consequence events are some of the most important issues in intelligence analysis, it is important to know whether GJP forecasters actually extract meaningful returns in this context.

To see how returns to precision vary across the probability spectrum, we divided forecasts into seven equal bins according to the IC's 2007 definition of "words of estimative probability." We then calculated rounding errors for each question using only forecasts which fell into a particular bin. This allows us to examine how much information forecasters lose by employing each individual term of this spectrum.[42]

[Tables 2a & 2b]

Table 2a shows that GJP forecasters extract larger returns to precision when making their highest and lowest probability forecasts. But we see consistent and substantial returns to precision in other parts of the spectrum as well. In almost every category, rounding off numerical forecasts costs multiple percentage points of predictive accuracy. This reinforces the proposition that geopolitical forecasters can consistently extract returns to precision, and that this is not simply a property of particular, "easy" questions.

Table 2b then presents rounding errors within each category of the new "words of estimative probability" spectrum which the DNI developed in 2015.[43] We find that the new, tighter bins for "remote" and "almost certain" forecasts not only eliminate rounding errors, but even improve many estimates by guarding against overconfidence. Yet because analysts achieve returns to precision on such a wide range forecasts, we also see that the new DNI guidance exacerbates rounding errors elsewhere on the spectrum, where bins must become wider in order to compensate for narrowing other categories. There are no free lunches here: it may be possible to mitigate or redistribute rounding errors by changing categories' definitions, but qualitative expressions still prevent analysts from reaching their full potential. Moreover, as systems of qualitative expression become more complex, these adjustments undermine the notion that using such systems is simpler and more intuitive than communicating assessments numerically.

Returns to precision 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 time when the forecasting question was resolved.[44] In our data

[41] This pattern is not driven by a subset of highly active respondents. The median respondent in our data set registered probability estimates that were not multiples of 0.10 on 49 percent of forecasts. The 25th percentile on this measure of how often forecasters used "non-round numbers" was 36 percent, and the 10th percentile was 21 percent.
[42] Graphing median rounding errors returns substantively similar results.
[43] In fact, the 2015 WEP spectrum induces rounding errors on slightly *more* questions overall than does the 2007 version (72 percent versus 68 percent).
[44] 8,509 forecasts in GJP's data were logged after a relevant event was judged to take place. For example, if a question asked whether country X would conduct a missile test by a certain date, and X conducted a missile test

set, the mean of this time lag was 72 days (standard deviation 83 days, median 41). Figure 3 shows this variable's distribution.

[Figure 3]

We identified forecasts as *Lay-Ups* if they were made with no more than five percent probability, and registered within two weeks of a question's closing time. Since these should be the easiest forecasts in the data set, we expect to see special returns to precision within this category. We divided all other forecasts into three categories with nearly equal numbers of observations: *Short-Term* forecasts were made less than 29 days before questions closed; *Medium-Term* forecasts were made from 29 to 83 days prior to closing; *Long-Term* forecasts were made more than 83 days prior to questions closing.

[Table 3]

Table 3 presents results. As expected, the costs to rounding "lay-up" forecasts are extremely large.[45] Yet removing these data from the analysis has a limited impact on aggregate rounding errors. Moreover, even though returns to precision decline when we examine longer-term forecasts, the same basic pattern holds here as in our original analysis: again, we see that rounding forecasts into confidence levels and estimative verbs sacrifices sizable (and statistically significant) amounts of information; again, we see that "words of estimative probability" recoup some (but not all) of these losses, which remain especially meaningful for high-quality forecasters.[46]

Section 5. Variation Across Individuals

In this section, we identify attributes that predict individual differences in returns to precision. We examine a battery of variables capturing skill, effort, experience, preparation, and cognitive style. We chose these variables not just because they plausibly explain variation in returns to precision, but also because they shed light on the important question on how to maximize returns to precision in practice.

Forecasting skill, effort, experience, and training can all be cultivated in a wide range of personnel. If these factors predict the degree to which foreign policy analysts can parse probabilities, then this would be grounds for optimism in thinking that the IC and other organizations can replicate and potentially exceed the performance shown in our data. By comparison, attributes like numeracy, education, and need for cognition are expensive to change. If these are the primary predictors of returns to precision in our data set, this would suggest that improving returns to precision in practice depends on selecting specific kinds of personnel. In the

---

before that date, this information might come to light with a time lag during which GJP questions would remain open. We exclude such forecasts from the analysis presented below.

[45] For calculating average rounding errors for superforecasters on lay-up forecasts, we dropped data on four questions where all forecasts were at 0.00 – otherwise, the average rounding error would be infinite.

[46] We omitted "medium-term" forecasts from Table 2 for clarity. The median rounding errors on medium-term forecasts adhere to the anticipated pattern, lying between their counterparts for long- and short-term forecasts. *Average* rounding errors on these medium-term forecasts are generally higher than for short-term forecasts, but this is because short-term forecast scores are suppressed by having "lay-ups" removed, and thus medium-term forecasts contain many more (accurate) low-probability forecasts.

analysis below, we reflect this distinction by dividing variables into *Targets for Cultivation* and *Targets for Selection*.

We measure forecasting skill using each respondent's median *Brier Score* across forecasts. All else being equal, higher-quality forecasters should receive greater penalties from having their forecasts manipulated. We use median Brier Score instead of the mean to reduce the impact of high-leverage observations.

Four additional variables capture effort, training, and experience. *Total Forecasting Questions* counts the number of distinct questions to which an individual responded throughout all years of the competition. All else being equal, we expect that respondents who have more experience making probability assessments (or who are simply more engaged in the competition) will be able to parse those probabilities in more informative ways. *Average Revisions per Question* captures how often respondents changed their answers to each forecasting question. This variable also proxies for effort and engagement with GJP; all else being equal, we expect that respondents who update forecasts more often will capture additional returns to precision. *Granularity* measures the proportion of a respondent's forecasts that were not recorded in multiples of 10 percentage points. All else being equal, we expect that those respondents who took care to express their views more precisely would incur larger rounding penalties than forecasters who rounded off their judgments to begin with.[47] *Probabilistic Training* takes a value of 1 if the forecaster received training in probability assessment from GJP.

We code two variables capturing respondents' "education" prior to participating in the Good Judgment Project. *Education Level* is a four-tiered variable capturing respondent's most advanced degree (0: no bachelor's; 1: bachelor's; 2: master's; 3: doctorate). More advanced education could also give respondents better ability to analyze complex questions and to parse probabilities reliably. *Numeracy* represents each respondent's score on a series of word problems designed to capture mathematical fluency. If respondents are better able to reason numerically, then this could translate into a better ability to parse probabilities.[48] In principle, organizations can cultivate both of these attributes. Indeed, the U.S. Intelligence Community pays for many employees' advanced degrees. However, numeracy and education levels are substantially more expensive to increase than the "effort and training" variables described above. GJP's training sessions lasted roughly one hour, for instance, while earning a graduate degree can take several years.

GJP data include several indices of "cognitive style," including: *Raven's Progressive Matrices*, where higher scores indicate better reasoning ability; an expanded Cognitive Reflection Test (*Expanded CRT*), where higher scores indicate an increased propensity for respondents to suppress misleading intuitive reactions in favor of more accurate, deliberative answers;[49] *Fox-Hedgehog*, a variable where higher scores capture respondents' self-assessed tendency to rely on one big simplifying framework versus more general, ad hoc reasoning; and

[47] A similar index of granularity representing the proportion of forecasts that were not multiples of 0.05 yields similar results.
[48] On numeracy and decision making, see Peters et al. 2006. GJP changed numeracy tests between years 2 and 3 of the competition. We thus normalized test results so that they represent the number of standard deviations above or below the mean that each respondent scored relative to all other forecasters who took that year's test.
[49] As with our numeracy variable, we normalized test results in order to combine different test versions across tournament years.

Need for Cognition, an index of respondents' self-assessed preference for addressing complex problems. Table 3 presents summary statistics for these variables.

We measured variation in returns to precision across individuals by examining each respondent's forecasts. We estimate Brier Scores after rounding each forecast into progressively larger numbers of bins, starting at $B = 2$. For each value of B, we conduct a one-sided paired-sample Wilcoxon signed rank test to determine whether forecasts rounded to B bins have worse Brier Scores than respondents' original predictions. We define the *threshold of estimative precision* (B^*) as the fewest number of bins where rounding errors are not statistically distinct from zero ($p < .05$). Thus any level of precision lower than B^* systematically sacrifices predictive accuracy.

Analysis

Table 4 shows summary statistics for each variable described here, as well as how individual-level attributes correlate with respondents' B^* thresholds. We drop forecasters from the analysis who made fewer than 25 forecasts across all three years of the competition. This leaves 1,714 individuals in our sample.

[Table 4]

Table 4 shows that all bivariate correlations are in the expected direction, though their magnitudes range widely. By and large, variables capturing skill, effort, training, and experience are more closely related to individual-level returns to precision than variables capturing education and cognitive style. Table 5 then presents ordinary least squares regression analyses predicting individual B^* thresholds. We centered and standardized continuous variables so that each coefficient in Table 5 reflects the proportion of a standard deviation that B^* thresholds increase when we raise each continuous predictor by a standard deviation, or when we change the *Probabilistic Training* variable from 0 to 1. Model 1 demonstrates that forecasting skill alone predicts substantial variation in individual-level returns to precision (R^2=0.23).

[Table 5]

Model 2 shows that adding variables for effort and training substantially increase model fit (R^2=0.37). In particular, our variables for *Total Forecasting Questions*, *Average Revisions per Question* and *Probabilistic Training* robustly predict returns to precision at the p<.001 level, and have substantial coefficients.[50] By contrast, Model 3 shows that education and cognitive style variables – which organizations should find hardest to cultivate – add much less information beyond our original, spare regression (R^2=0.28 vs. 0.23). Moreover, none of these attributes retains statistical significance when we examine all control variables together in Model 4. Model 5 then confirms that education and cognitive style variables provide little marginal value in predicting returns to precision by replicating Model 2 on the 898 observations for which we have

[50] Adding a squared term for *Total Forecasting Questions* is statistically significant at the p<.01 level, but improves R^2 by less than 2 points.

data on all ten variables of interest: dropping the education and cognitive style variables from the analysis leaves the model's R^2 unchanged.[51]

These results have two main, practical implications. First, returns to precision correlate with factors that foreign policy analysts and organizations can feasibly cultivate. GJP forecasters who received brief training sessions in probabilistic reasoning were roughly one-sixth of a standard deviation higher on returns to precision than their peers. Especially since this training was randomly-assigned, our findings suggest that the IC and other organizations could replicate (and presumably exceed) this benefit of training their own personnel.

We also found that respondents' experience making forecasts, their willingness to revise those forecasts, and their willingness to make forecasts precisely all predicted variation in how well those respondents could parse probabilities (though the coefficient on respondent *Granularity* is consistently smaller than the others). Though these attributes were not randomly-assigned, these findings again provide grounds for optimism in thinking that professional forecasters could replicate and potentially exceed the returns to precision shown in GJP data. Foreign policy analysts 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 GJP respondents (who revised their forecasts, on average, less than once per question).

It is unsurprising to see that *Total Forecasting Questions* predicts B^* thresholds. Forecasters who registered more predictions were not only more experienced and more engaged in the competition, but they also provided larger sample sizes for calculating B^* thresholds. Smaller rounding errors would register as being statistically significant. Our analyses cannot distinguish how much of the correlation between forecast volume and B^* thresholds results from sample size as opposed to individual attributes. Yet both interpretations have the same implications for drawing practical implications: the more forecasts individuals make, the more likely it becomes that rounding off their estimates will cause a systematic loss of information. Our analysis shows that 400 additional forecasts predicts one quarter of a standard deviation's improvement in estimative precision. In the U.S. Intelligence Community, it is plausible to expect that analysts cross this threshold many times over, every day. Thus the relationship we observe between *Total Forecasts* and returns to precision in our data further emphasizes how GJP data presumably understate the degree to which professional forecasters could achieve consistent returns to precision.

Second, and no less important, our findings reject the notion that returns to precision correlate with innate individual-level attributes that foreign policy analysts and organizations cannot cultivate. While notional divisions between "mathematicians" and "poets" are common in the intelligence literature,[52] we see no evidence that returns to precision belong primarily to forecasters who are especially skilled in quantitative reasoning, or who have special educational backgrounds, or who possess particular cognitive styles. Instead, our data suggest that when skilled forecasters of all kinds take the time and effort to make precise forecasts, this consistently adds information to foreign policy analysis.

[51] Similarly, our findings hold if we replicate Models 1 and 2 on this limited sample, with R^2 values of 0.26 and 0.51, respectively.
[52] For example, Kent 1964, Johnston 2005.

Section 6. Variation Across Question Types

We also code B^* thresholds for each forecasting question that GJP posed.[53] Figure 4 presents a histogram of these thresholds. On 40 percent of questions, GJP respondents could effectively parse probabilities into at least 8 bins (that is, more finely than "words of estimative probability").

Explaining variation in returns to precision across questions is largely a matter of coding question difficulty. Two ex post measures of question difficulty are the daily variance in respondents' estimates, and how far the daily average estimate on each question departed from 0.50. These two factors alone explain substantial variation in question-level B^* thresholds: an ordinary least squares regression featuring these variables alone returns an R^2 of 0.61.[54]

Next, we examine whether there are specific topics where GJP forecasters appeared better-able to parse probabilities. GJP coded question types on several dimensions. Each forecasting problem received a primary region tag (10 potential values) and a primary functional tag (14 potential values).[55] When we combine all 24 tags in a regression predicting question-level B^* thresholds, none is statistically significant. Only five tags are statistically significant predictors of B^* thresholds in univariate regressions.[56] When all five of these tags are added to the regression model described above based on question-level forecast mean and variance, these variables increase R^2 by less than 0.02 (and only the region dummy for Western Europe retains statistical significance). Adding all 24 content tags are added to the regression model based on forecast mean and variance improves R^2 by just 0.04.

These analyses reinforce the broader argument developed throughout this paper: that foreign policy analysts can consistently parse probabilities more finely than what common systems of qualitative expression allow, and that their ability to do so does not systematically depend on idiosyncratic features of the prediction at hand. The concluding section of this paper connects these findings to practical debates about intelligence analysis, foreign policy, and probability assessment in other fields.

Section 7. Discussion

This paper has demonstrated that when foreign policy analysts do not express probabilities numerically, they are systematically hindered from reaching their full potential. These findings do not depend on extreme forecasts, near time horizons, or particular question types. Returns to precision do not primarily belong to forecasters with special educational backgrounds or quantitative skills. These findings suggest that it is possible to improve the informational value

[53] The data analyzed in this section span 282 questions. While the majority of our analyses include estimates drawn from 288 questions, we exclude 6 in the present analyses due to missing data.
[54] See supplementary material for regression output.
[55] Horowitz et al. 2015. The regional tags were Sub-Saharan Africa, Central/South America, North America, South/Central Asia, East/Southeast Asia, Eastern Europe, Western Europe, Middle East/North Africa, Oceania, and Global. The functional tags were Commodities, Currencies, Diplomatic Relations, Domestic Conflict, Economic Growth/Policy, Elections, International Organizations, International Security/Conflict, Leader Entry/Exit, Public Health, Resources/Environment, Technology, Trade, Treaties/Agreements, and Weapons.
[56] Region tags for North America and Western Europe; Functional tags for Currencies, Diplomatic Relations, and Domestic Conflict.

of intelligence estimates and other forms of foreign policy analysis simply by using clearer language.

As mentioned in Section 1, returns to precision are one of many factors to consider when developing guidelines for probability assessment. Nevertheless, our analysis has four main practical implications.

First, *foreign policy analysts should always express probabilities more precisely than "confidence levels" and "estimative verbs."* Our data indicate that these crude modes of expressing probability sacrifice substantial information. Even if some analysts resist using numbers to express probability, our results indicate that they can consistently improve the informational value of their estimates with careful word choice (including more consistent implementation of official guidelines for using "words of estimative probability"). Across all forecasters in our data set, the difference between using "estimative verbs" and seven-step WEPs is worth twenty to thirty percent on the typical forecast. For high-quality forecasters, these penalties are far higher.

Second, our data suggest that scholars and practitioners should *revisit debates about the use of numerical probabilities in intelligence* and other areas of foreign policy analysis. As we discussed in Section 1, many scholars and practitioners oppose the use of numerical probabilities in foreign policy analysis on the grounds that this subject is so difficult that estimative precision is essentially random noise. Our findings show that such cynicism is unfounded. Another common objection is that the use of numerical probabilities would impose costs on the production or interpretation of foreign policy analysis. This hypothesis may be correct, but we are not aware that it has ever been rigorously substantiated. In light of the findings presented here, we believe the burden of proof lies with opponents of numerical probabilities to justify why it is not worth capturing the gains demonstrated here.

Third, our data suggest that *there are substantial benefits to cultivating skills in probability assessment*. GJP forecasters who were randomly assigned to just one hour of training in this subject were substantially better-able to parse their forecasts. More generally, as discussed in Section 5, our findings provide grounds for optimism in thinking that if the IC and other organizations prioritized this subject, then they could achieve even greater returns to precision than what we observed in GJP's data.

Fourth, *internal use of numeric probabilities is valuable even if analysts ultimately choose to report qualitative assessments to decision makers*. As shown in Table 4, one of the most important predictors of individual-level returns to precision in our data was the frequency with which respondents re-evaluated forecasts in light of new information. Similarly, other GJP studies have shown that forecasters who shift their assessments more frequently also systematically achieve better Brier Scores.[57]

If analysts update their forecasts to incorporate new information, it is far easier to do so when assessing probabilities using numbers rather than words. For example, consider an analyst who believes it is 30 percent likely that supplying nonlethal aid to Ukraine would allow Kiev to contain Russian-backed separatists. Say this analyst receives new intelligence that provides a

[57] See, for example, Mellers et al. 2015b.

small signal that this aid would be more useful than she originally thought. Thus our analyst might choose to raise her assessment to 32 percent. If our analyst were expressing uncertainty using WEPs, this piece of information would not nearly be large enough to change her assessment from "likely" to "even chance." Over time, these small changes could aggregate into substantial revisions from our analyst's original assessment. And while it is possible that an analyst making verbal assessments of uncertainty could keep all of these marginal adjustments in mind until they necessitate altering prescribed WEPs, we suspect it is far easier for analysts to make these adjustments incrementally and often.

Broader implications

While our study is motivated by debates about foreign policy analysis, our approach can be applied to any field where scholars and practitioners debate proper means for communicating expert judgment. Medicine is a prime example. One of a physician's most important responsibilities is to communicate clearly with patients about uncertain diagnoses and treatment outcomes. Many medical professionals – like many intelligence analysts – are reluctant to express probabilistic judgments explicitly.[58] Our paper offers a method for examining the level of specificity that is achievable and worthwhile in such circumstances.

Our findings also inform foreign policy discourse in the public domain. Scholars, think-tank analysts, reporters, and pundits regularly assess the kinds of questions that the Good Judgment Project posed. These assessments shape opinions and public policy. Yet they can be just as vague as intelligence reporting, often even more so.[59]

Those concerned with the quality of public foreign policy discourse should also consider what level of specificity is justified and desirable. For example, when a pundit asserts that a military operation is "likely" to succeed, do they mean that the probability is just above even or that it is closer to certainty? When a scholar argues that a crisis is "unlikely" to escalate, is that more like 40 percent or 20 percent? Beyond forcing analysts to be clearer (and thus possibly more careful) in articulating their views, it appears that these distinctions consistently add information to foreign policy analysis. Despite the uncertainty and subjectivity that are inherent to predicting international politics, there are valid grounds for asking foreign policy analysts of all kinds to assess uncertainty more precisely than common practice.

[58] See Braddock et al. 1999 and Politi, Han, and Col 2007. Agrell and Treverton 2015, chs. 5-6 describe overlap between medicine and intelligence analysis in this area.
[59] On forecasting and punditry, see Tetlock 2009, Gardner 2011, and Silver 2012.

References

Agrell, Wilhelm and Gregory F. Treverton. 2015. *National Intelligence and Science: Beyond the Great Divide in Analysis and Policy*. Oxford.

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.

Barnes, Alan. 2015. Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intelligence and National Security*, in press.

Betts, Richard K. 2007. *Enemies of Intelligence: Knowledge and Power in American National Security*. Columbia.

Beyerchen, Alan. 1992/93. Clausewitz, Nonlinearity, and the Unpredictability of War," *International Security* 13 (3): 59-90.

Beyth-Marom, Ruth. 1982. How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting* 1: 257-269.

Bowden, Mark. 2012. *The Finish: The Killing of Osama bin Laden*. Atlantic Monthly.

Braddock, Clarence H., Kelly A. Edwards, Nicole M. Hasenberg, Tracy L. Laidley, and Wendy Levinson. 1999. Informed Decision Making in Outpatient Practices. *Journal of the American Medical Association* 282: 2313-2320.

Budescu, David V., Han-Hui Por, and Stephen B. Broomell. 2012. Effective Communication of Uncertainty in the IPCC Reports. *Climatic Change* 113: 181-200.

------, and Michael Smithson. 2014. The Interpretation of IPCC Probabilistic Statements Around the World. *Nature Climate Change* 4: 508-512.

Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction [Silberman-Robb Commission]. 2005. *Report to the President of the United States*. U.S. Government Printing Office.

Davis, Jack. 1997. *A Compendium of Analytic Tradecraft Notes*. Washington, D.C.: Central Intelligence Agency.

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

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

Friedman, Jeffrey A. and Richard Zeckhauser. 2015. Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden. *Intelligence and National Security* 30 (1): 77-99.

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

Horowitz, Michael C., Philip Rescober, Laura Resnick, Pavel Atanasov, Barbara Mellers, and Philip Tetlock. 2015. Learning and Foreign Policy: Evidence from Crowd-Sourced Geopolitical Forecasts. Paper prepared for International Studies Association annual meeting in New Orleans, La.

Jervis, Robert. 2006. Reports, Politics, and Intelligence Failures: The Case of Iraq. *Journal of Strategic Studies* 29 (1): 3-52.

Johnson, Edgar M. 1973. *Numerical Encoding of Qualitative Expressions of Uncertainty*. Army Research Institute for the Behavioral and Social Sciences.

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

Kent, Sherman. 1964. Words of Estimative Probability. *Studies in Intelligence*.

Keynes, John M. 1937. The General Theory of Employment. *Quarterly Journal of Economics* 51 (2): 209-223.

Lowenthal, Mark M. 2008. Towards a Reasonable Standard for Analysis: How Right, How Often on Which Issues? *Intelligence and National Security* 23/3.

MacEachin, Douglas J. 1995. "Tradecraft of Analysis," in *U.S. Intelligence at the Crossroads: Agendas for Reform*, ed. Roy C. Godson, Ernest May, and Gary Schmitt. Washington, D.C.: Brassey's.

Mandel, David R. and Alan Barnes. 2014. Accuracy of Forecasts in Strategic Intelligence. *Proceedings of the National Academy of Sciences*. EarlyView.

Marchio, James. 2014. "If the Weatherman Can…": The Intelligence Community's Struggle to Express Analytic Uncertainty in the 1970s. *Studies in Intelligence* 58 (4): 31-42

Marrin, Stephen. 2012. Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure? *Intelligence and National Security* 27 (6): 896-912.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, D. Moore, Pavel Atanasov, S. A. Swift, T. Murray, E. Stone, and Philip E. Tetlock. 2014. Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* 25 (5): 1106-15.

Mellers, Barbara, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, Emlen S. Metz, Lyle Ungar, Michael M. Bishop, Michael Horowitz, Ed Merkle, and Philip Tetlock. 2015a. The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics. *Journal of Experimental Psychology: Applied*, in press.

Mellers, Barbara, E. Stone, T. Murray, Angela Minster, Nick Rohrbaugh, M. Bishop, E. Chen, Joshua D. Baker, Y. Hou, Michael Horowitz, Lyle Ungar, and Philip E. Tetlock. 2015b. Improving Probabilistic Predictions by Identifying and Cultivating 'Superforecasters.' *Perspectives in Psychological Science*, in press.

Morell, Michael. 2015. *The Great War of Our Time: An Insider's Account of the CIA's Fight Against Al Qa'ida*. Twelve.

Mosteller, Frederick and Cleo Youtz. 1990. Quantifying Probabilistic Expressions. *Statistical Science* 5 (1): 2-12.

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

Peters, Ellen, Daniel Vastfjall, Paul Slovic, C. K. Mertz, Ketti Mazzocco, and Stephan Dickert. 2006. Numeracy and Decision Making. *Psychological Science* 17: 407-413.

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

Politi, Mary C., Paul K. J. Han, and Nananda F. Col. 2007. Communicating the Uncertainty of Harms and Benefits of Medical Interventions. *Medical Decision Making* 27 (5): 681-695.

Posner, Richard A. 2005. *Preventing Surprise Attacks: Intelligence Reform in the Wake of 9/11*. Rowman & Littlefield.

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.

Satopää, Ville A., Shane T. Jensen, Barbara A. Mellers, Philip E. Tetlock, and Lyle H. Ungar. 2014. Probability Aggregation in Time-Series: Dynamic Hierarchical Modeling of Sparse Expert Beliefs. *Annuals of Applied Statistics* 8: 1256-1280.

Schrage, Michael. 2005. What Percent is 'Slam Dunk'? Give Us Odds on Those Estimates. *Washington Post*, 20 February, B01.

Silver, Nate. 2012. *The Signal and the Noise: Why Most Predictions Fail – But Some Don't*. Penguin.

Steyvers, Mark, Thomas S. Wallsten, Edgar C. Merkle, and Brandon M. Turner. 2014. Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models. *Risk Analysis* 34: 435-452.

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

Tetlock, Philip E. 2009. Reading Tarot on K Street. *The National Interest* 104: 57-67.

Tetlock, Philip E. and Barbara A. Mellers. 2011. Intelligence Management of Intelligence Agencies: Beyond Accountability Ping-Pong. *American Psychologist* 66 (6): 542-554.

Tetlock, Philip E., Barbara Mellers, Nick Rohrbaugh, and Eva Chen. 2014. Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in Psychological Science* 23 (4): 290-295.

Wallsten, Thomas S. and David V. Budescu. 1995. A Review of Human Linguistic Probability Processing: General Principles and Empirical Evidence. *Knowledge Engineering Review* 10: 43-62.

Wallsten, Thomas S., Yaron Shlomi, and Hisuchi Ting. 2008. Exploring Intelligence Analysts' Selection and Interpretation of Probability Terms. University of Maryland Working Paper.

Wark, David L. 1964. The Definition of Some Estimative Expressions. *Studies in Intelligence*.

Weiss, Charles. 2008. Communicating Uncertainty in Intelligence and Other Professions," *International Journal of Intelligence and CounterIntelligence* 21 (1): 57-85.

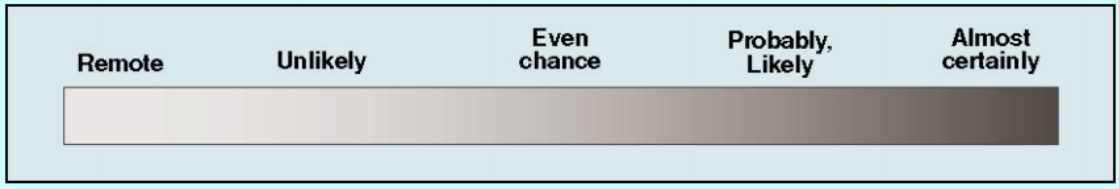
Wheaton, Kristan J. 2012. The Revolution Begins on Page Five: The Changing Nature of NIEs. *International Journal of Intelligence and CounterIntelligence*. 25 (2): 330-349.

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

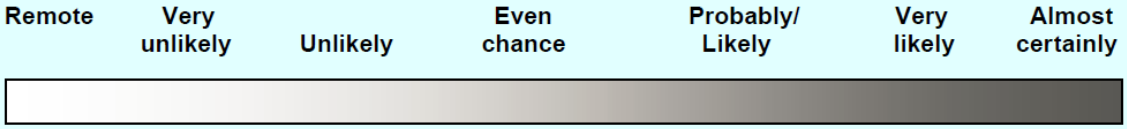
Figure 1. "Words of Estimative Probability"

a. In the January 2007 NIE, *Prospects for Iraq's Stability: A Challenging Road Ahead*

Intelligence judgments pertaining to likelihood are intended to reflect the Community's sense of the probability of a development or event. Assigning precise numerical ratings to such judgments would imply more rigor than we intend. The chart below provides a rough idea of the relationship of terms to each other.



b. In the November 2007 NIE, *Iran: Nuclear Intentions and Capabilities*



c. In the 2015 version of Intelligence Community Directive 203, *Analytic Standards*

(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%

Table 1. Estimative Precision and Predictive Accuracy – Aggregated Results

Reference class		*Brier Scores for Numerical Forecasts*	*Rounding Errors*			
			Words of estimative probability† (2015 version)	Words of estimative probability (7 equal bins)	Confidence levels (3 bins)	Estimative verbs (2 bins)
All forecasters	*Mean:*	0.154	0.7%***	1.8%	+10.8%***	+29.2%***
	Median:	0.122	0.8%***	1.5%***	+6.5%***	+20.1%***
Individuals, No training	*Mean:*	0.189	0.7%***	0.5%	5.5%***	15.2%***
	Median:	0.153	0.7%***	0.2%	3.5%***	12.2%***
Individuals, Trained	*Mean:*	0.173	0.5%***	1.4%	6.7%***	19.1%***
	Median:	0.138	0.9%***	1.0%*	4.4%***	12.9%***
Groups, No training	*Mean:*	0.161	0.6%***	2.2%**	9.4%***	27.9%***
	Median:	0.129	0.6%***	1.4%***	6.1%***	17.1%***
Groups, Trained	*Mean:*	0.137	0.6%***	2.6%	15.2%***	41.0%***
	Median:	0.101	0.7%***	1.7%***	8.9%***	25.8%***
Super-forecasters	*Mean:*	0.090	6.0%***	37.7%***	219.9%***	521.4%***
	Median:	0.033	1.9%***	9.9%***	51.8%***	136.1%***

Table 1 presents rounding errors for different groups of forecasters at different degrees of (im)precision; *p*-values reflect the results of paired-sample t-tests (for mean differences in absolute Brier Score) and paired-sample Wilcoxon signed-rank tests (for median differences in absolute Brier Score). * p<.05, ** p<.001, *** p<.001. †Currently recommended by the Office of the Director of National Intelligence (see Figure 1).

Figure 2. Histogram of forecasts in GJP data

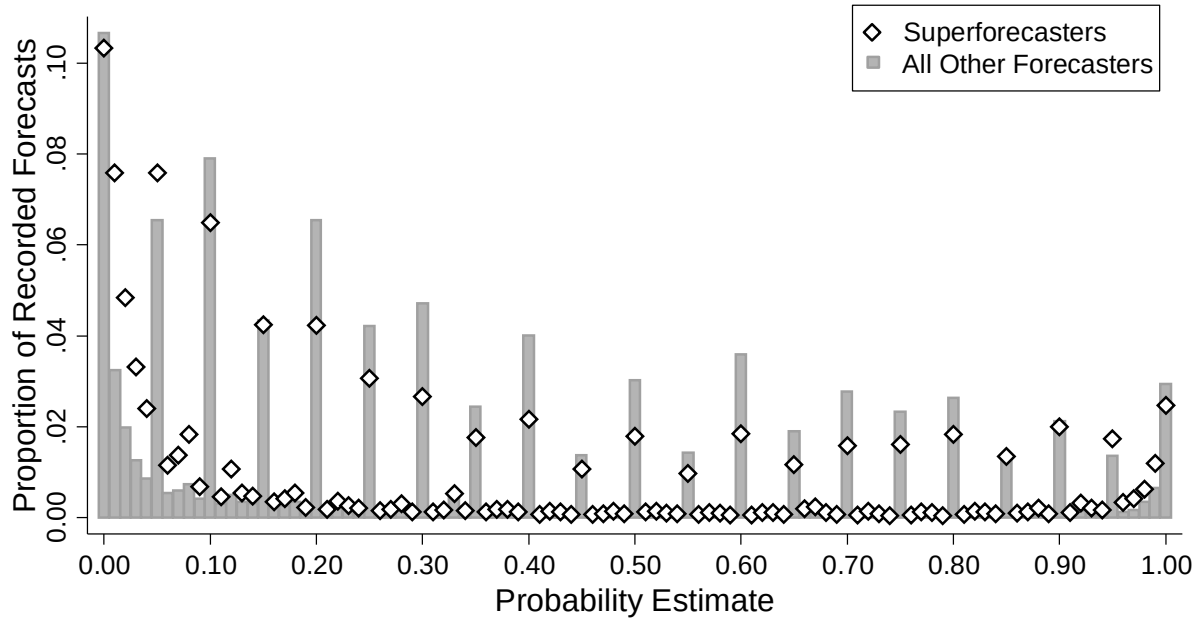


Figure 3. Distribution of forecasts by time horizon

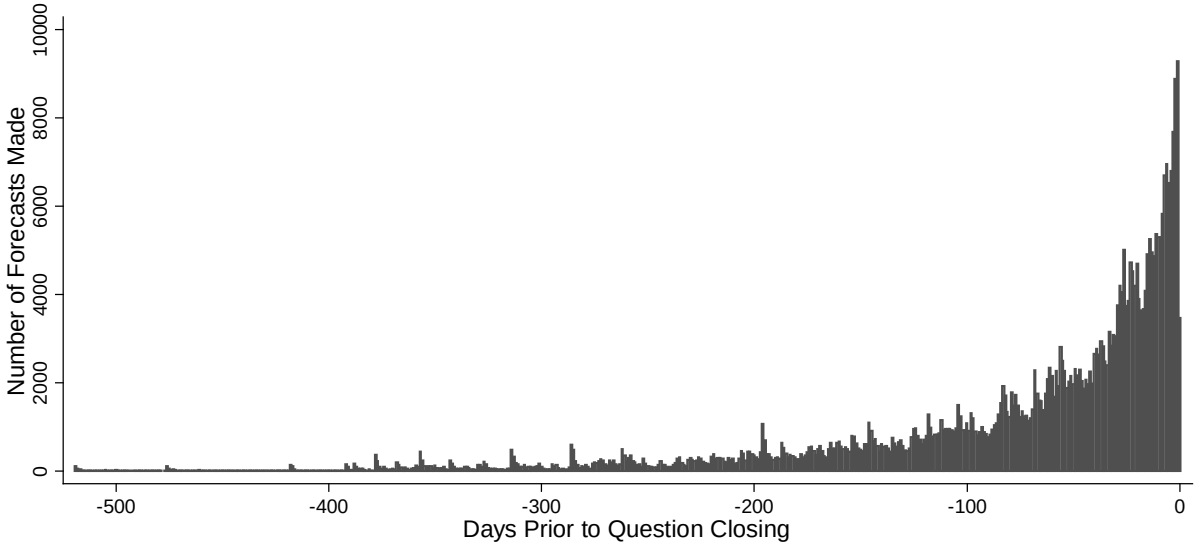


Table 2a. Returns to Precision Across the Number Line, Divided into Seven Equal Bins

Group		All Forecasts (.00-1.0)	Remote (.00-.14)	Very Unlikely (.15-.28)	Unlikely (.29-.42)	Even chance (.43-.56)	Likely (.57-.71)	Very Likely (.72-.85)	Almost Certain (.86-1.0)
All Forecasters	*Mean:*	1.8%	31.9%	7.8%***	3.9%***	1.1%***	1.0%**	1.3%**	12.7%
	Median	1.5%***	25.0%***	8.3%***	4.1%***	0.9%***	0.6%**	0.8%**	3.6%
Super-Forecasters	*Mean:*	37.7%***	146.3%	17.8%***	6.8%***	1.9%*	2.2%	3.9%	78.0%
	Median	9.9%***	78.5%**	15.4%***	4.9%***	0.3%*	1.0%	1.4%	1.6%

Table 2 presents rounding errors when we examine forecasts that fall within different segments of the number line and round those forecasts into seven equal bins. GJP forecasters extract their largest returns to precision when making extreme probability forecasts, but we see substantial rounding errors elsewhere as well. Statistical significance measured using paired sample Wilcoxon signed-rank tests: * p<.05, ** p<.01, *** p<.001.

Table 2b. Returns to Precision Across the Number Line, Binned According 2015 DNI Guidelines

Group		All Forecasts (.00-1.0)	Remote (.00-.05)	Very Unlikely (.05-.20)	Unlikely (.20-.45)	Even chance (.45-.55)	Likely (.55-.80)	Very Likely (.80-.95)	Almost Certain (.95-1.0)
All Forecasters	*Mean:*	0.7%***	-2.2%	-20.1%***	2.7%***	1.3%***	2.5%***	2.3%***	10.3%
	Median	0.8%***	-3.6%	-27.3%	2.4%***	1.0%***	2.9%***	7.7%***	-2.6%
Super-Forecasters	*Mean:*	6.0%***	27.22%†	-7.9%***	9.9%***	2.5%***	5.6%***	2.1%***	-39.21%
	Median	1.9%***	0.4%	-10.3%	7.9%***	0.4%***	4.2%***	2.7%***	-0.7%***

Table 2 presents rounding errors when we examine forecasts that fall within different segments of the number line and round those forecasts according to the DNI's 2015 guidelines. Compared to previous doctrine (see Table 2a), the new guidelines improve performance on extreme forecasts, but worsen performance elsewhere. Statistical significance measured using paired sample Wilcoxon signed-rank tests: * p<.05, ** p<.01, *** p<.001. †Mean does not include one observation where superforecasters unanimously (and accurately) predicted an outcome had a zero percent chance of occurring; rounding this forecast upwards results in an undefined loss of accuracy.

Table 3. Returns to Precision Across Time Horizons

Reference class		*Brier Score across all numerical forecasts*	*Rounding Errors*			
			7 WEPs, 2015 version	7 WEPs, evenly spaced	Confidence levels (3 bins)	Estimative verbs (2 bins)
All forecasts (excluding "lay-ups")						
All forecasters	*Mean:*	0.167	0.7***	1.4%	7.6%***	21.9%***
	Median:	0.134	0.8***	1.1%***	4.4%***	14.5%***
Super-forecasters	*Mean:*	0.102	6.0***	22.3%***	134.9%***	330.0%***
	Median:	0.039	1.9***	7.6%***	37.6%***	107.3%***
Long-term forecasts: ≥83 days						
All forecasters	*Mean:*	0.189	0.8%***	1.2%	6.2%***	17.4%***
	Median:	0.158	0.9%***	0.7%	3.3%***	11.5%***
Super-forecasters	*Mean:*	0.127	1.3%***	10.3%***	61.1%***	161.5%***
	Median:	0.052	1.3%***	5.6%***	21.0%***	62.7%***
Short-term forecasts: ≤29 days (excluding "lay-ups")						
All forecasters	*Mean:*	0.115	1.1%***	4.8%*	29.2%***	80.1%***
	Median:	0.073	1.0%***	2.6%***	11.7%***	42.0%***
Super-forecasters	*Mean:*	0.069	115.3%***	741.5%***	4,176.7%***	9,469.0%***
	Median:	0.007	5.8%***	40.7%***	312.3%**	757.1%**
"Lay-up" forecasts: ≤14 days, ≤5 percent						
All forecasters	*Mean:*	0.196	48.4%	646%***	3,878%***	8,830%***
	Median:	7.0e-4	27.1%***	620%***	3,823%***	8,728%***
Super-forecasters	*Mean:*	0.119	1,076.9%***	6,512%***	35,842%***	80,756%***
	Median:	2.1e-4	325.4%***	2,311%***	13,029%***	29,442%***

p-values calculated using paired-sample t-tests (for mean rounding errors) and paired-sample Wilcoxon signed-rank tests (for median rounding errors). *p<.05, ** p<.01, *** p<.001.

Table 4. Summary Statistics for Individual-Level Attributes

	N	*Mean*	*Std. dev.*	*Min*	*Max*	*Corr. w/B**
Returns to Precision						
Threshold of Estimative Precision (B^*)	1,714	3.81	7.24	1	101	-
Forecasting Skill						
Median Brier Score	1,714	0.14	0.12	0	1.04	-0.47
Effort, Training, Experience						
Total Forecasting Questions	1,714	92.74	64.09	25	288	0.26
Average Revisions per Question	1,714	1.13	3.24	1	89.1	0.35
Granularity	1,714	0.48	0.19	0	1	0.14
Probabilistic Training	1,714	0.65	0.48	0	1	0.24
Education						
Education Level	1,712	1.89	0.79	0	3	0.02
Numeracy	1,711	-0.03	1.02	-4.80	0.72	0.08
Cognitive Style						
Raven's Progressive Matrices	1,708	0.66	0.22	0	1	0.04
Cognitive Reflection Test	1,243	-0.01	1.01	-4.01	-0.97	0.13
Fox-Hedgehog	1,620	2.35	1.02	1	5	0.02
Need for Cognition	1,367	5.82	0.65	3.5	7	0.04

Fourteen respondents' B^* thresholds were Winsorized to 21 bins (i.e., the level of precision corresponding to increments of five percentage points) when estimating bivariate correlations, so as to reduce the impact of outliers.

Table 5. Predicting Individual-Level Returns to Precision

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 5*[†]
Targets for cultivation					
Brier Score	-0.47 (.03)***	-0.42 (.03)***	-0.58 (.04)***	-0.51 (.04)***	-0.51 (.04)***
Total Forecasting Questions		0.23 (.02)***		0.20 (.03)***	0.20 (.03)***
Average Revisions per Question		0.24 (.09)***		0.68 (.07)***	0.70 (.07)**
Granularity		0.03 (.03)		0.07 (.03)*	0.08 (.03)*
Probabilistic Training (dummy)		0.19 (.04)***		0.18 (.04)***	0.17 (.04)***
Targets for selection					
Numeracy			0.06 (.03)*	0.04 (.02)	
Education Level			0.02 (.03)	-0.01 (.03)	
Raven's Progressive Matrices			0.10 (.03)**	0.01 (.03)	
Cognitive Reflection Test			0.04 (.03)	0.01 (.02)	
Fox-Hedgehog			0.05 (.03)	0.04 (.03)	
Need for Cognition			0.02 (.03)	-0.00 (.03)	
Constant	-	-0.12 (.03)***	0.07 (.02)**	-0.15 (.04)***	-0.14 (.04)***
N	1,714	1,714	898	898	898
R^2	0.23	0.37	0.28	0.51	0.51

Ordinary least squares regression predicting B^* thresholds for individual respondents.
Robust standard errors. * p<.05 ** p<.01 ***p<.001.
[†]Model 5 only retains observations used in Model 4

Figure 4. B^* Thresholds Across GJP Questions (N=282)