Why Quantitative Probability Assessments Are Empirically Justifiable in Foreign Policy Analysis

Jeffrey A. Friedman, *Assistant Professor of Government, Dartmouth College*[1]
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*

This draft: December 11, 2015 (11,970 words)
All comments welcome

Abstract. Aristotle counseled us to seek precision insofar as the nature of the subject permits. But how much is too much? This article provides the first systematic test of long-standing debates about how precisely foreign policy analysts can estimate probabilities. Using a data set of 888,328 forecasts drawn from a series of geopolitical forecasting tournaments, we demonstrate that qualitative probability assessments, including seven-step scales employed by U.S. intelligence analysts, systematically sacrifice accuracy. Respondents' capacities to extract these "returns to precision" correlate more with forecasting skill, effort, and training in probabilistic reasoning than with numeracy, education, or cognitive style. Our results indicate that foreign policy analysts can parse their judgments more precisely than conventional wisdom supposes and that this ability can be cultivated. We argue that it is possible to improve the value of intelligence reports and other political commentary by supplementing qualitative descriptions of uncertainty with quantitative estimates of subjective probability.

Introduction

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. The Chiefs believed that the plan's chances of success were roughly one-in-three. But when they conveyed this view to the president in writing, they stated only that "This plan has a fair chance of success." The report's author, Brigadier General David Gray, claimed that "We thought other people would think that 'a fair chance' would mean 'not too good.'" President Kennedy, by contrast, reportedly interpreted "a fair chance" to suggest support for the invasion. After the fact, Gray believed that his imprecise

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

language enabled a strategic blunder, while Kennedy resented what he saw as the Chiefs' failure to offer a clearer expression of doubt (Wyden 1979, 88-90).

Yet many scholars and practitioners of foreign policy are also skeptical of expressing probability assessments explicitly, especially if doing so involves using numbers. 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. As the group struggled to defend and reconcile their differing judgments, President Obama himself questioned whether these judgments "disguised uncertainty as opposed to actually providing you with more useful information" (Bowden 2012, 160).

This paper demonstrates that numeric probability assessments actually *do* provide more information than "estimative verbs," "confidence levels," "words of estimative probability" and other qualitative terminologies commonly used by foreign policy analysts. We draw on a data set of 888,328 geopolitical forecasts collected by the Good Judgment Project in collaboration with the U.S. Intelligence Community. These data show that rounding numeric probability estimates to different degrees of (im)precision systematically sacrifices predictive accuracy. These findings do not depend on extreme probability estimates, short time horizons, particular scoring rules, or special question types. Forecasters' abilities to extract these "returns to precision" correlate mainly with forecasting skill, effort, and training in probabilistic reasoning, rather than with numeracy, education, or cognitive style. Our results indicate that foreign policy analysts can estimate probabilities more precisely than conventional wisdom supposes and that this ability can be cultivated. We argue that it is possible to improve the informational value of intelligence reports and other forms of political commentary by supplementing qualitative, natural-language based descriptions of uncertainty with quantitative estimates of subjective probability.

Although our empirical analysis focuses on long-standing debates about foreign policy analysis, this is only one of many fields featuring sharp disagreement about the value of precision in probability assessment. In medicine, law, finance, and other professions, high-stakes decisions depend on assessing uncertainty, and practitioners disagree about the wisdom – even the ethics – of making those judgments explicit (Erev and Cohen 1990; Wallsten and Budescu 1995). The practice of quantifying probability estimates generates deep controversy in areas including political punditry (Gardner 2011), regulation (Sunstein 2014), medicine (Nakao and Axelrod 1983), and climate science (Budescu, Broomell, and Por 2014), just as it has been a focal point for a cultural divide between so-called "mathematicians" and "poets" in the U.S. Intelligence Community for more than fifty years (Kent 1964; Johnston 2005).

Generally speaking, opponents of fine-grained (and especially quantitative) probability assessments argue that these assessments represent unjustifiable detail. In the worst case, overprecision could create an illusion of rigor, making subjective judgments appear sounder than they really are. At the very least, it is harder to justify devoting effort to refining estimates if additional precision merely represents random noise. Can analysts of foreign policy or other subjects reliably calibrate their judgments more finely than the coarse distinctions employed by the U.S. Intelligence Community? Do numeric probability assessments contain more information than common qualitative expressions? These are ultimately empirical questions, which scholars can address with appropriate data and methods. Yet to date, there has been no systematic study identifying how precisely foreign policy analysts can reliably estimate probabilities.

Our main contribution is to demonstrate that quantitative probability assessments are empirically justifiable in foreign policy analysis, despite widespread arguments to the contrary. And though findings in one domain do not transfer directly to others, foreign policy analysis is generally held to be a discipline featuring an unusually high degree of uncertainty, complexity, and subjectivity. If respondents in our study can reliably quantify probability assessments, then other disciplines may also benefit from subjecting skepticism about probabilistic precision to direct empirical testing.

We present this analysis in seven parts. Section 1 describes debates about expressing uncertainty in foreign policy analysis, against the backdrop of broader controversies about subjective probability assessment. Sections 2 and 3 describe our data and empirical approach. Section 4 shows how commonly-used qualitative expressions systematically sacrifice predictive accuracy 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 and other fields.

Section 1. Expressing Probability in Foreign Policy Analysis

Some scholars and practitioners argue that foreign policy analysts should avoid assessing probability altogether. The U.S. Intelligence Community (IC) deliberately eschews long-term predictions in its *Global Trends* reports on the assumption that these predictions would be indefensible. Similarly, Thomas Fingar, formerly Chair of the U.S. 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" (Fingar 2011, 53, 74; MacEachin 1995; Davis 1997).

Such views reflect the assumption that world politics is so complex that prediction is effectively meaningless (Beyerchen 1992/93). Yet predictability is a matter of degree. Karl Popper (1972, 207) 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." International affairs may be more "cloudlike" than many other disciplines, but it is ultimately an empirical question as to how finely foreign policy analysts can parse their probability assessments. Scholars have yet to address this question directly.

Over the past decade, questions about the proper level of precision for expressing probability have become especially important within the U.S. Intelligence Community. 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 uncertainty surrounding key judgments in appropriate detail. The White House-appointed Silberman-Robb Commission (2005, 419, 409) 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." The 2004 Intelligence Reform and Terrorism Prevention Act required analysts to "properly caveat and express uncertainties or confidence in analytic judgments." Yet there is currently no consensus on what "properly

expressing" uncertainty entails, and there are several common proposals and practices to consider:

Estimative verbs. Foreign policy analysts commonly communicate probability with phrases such as "we judge," "we estimate," or "we assess." 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." Although these estimative verbs indicate that judgments are uncertain, such phrasings do not parse uncertainty any further than implying that these conclusions are likely to be true (Lowenthal 2006, 128).

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. For example, the 2007 NIE on *Iran: Nuclear Capabilities and Intentions* includes statements such as "We judge with high confidence that in fall 2003, Tehran halted its nuclear weapons program" and "We assess with moderate confidence Tehran had not restarted its nuclear weapons program as of mid-2007." A statement made with "high confidence" is presumably more likely to be true than a statement made with "moderate" confidence. Dividing the number line into "high," "moderate," and "low" confidence categories thus presents more detail than estimative verbs alone.

Words of estimative probability. Recent NIEs include front matter defining "words of estimative probability" (WEPs). These terms allow analysts to express probability qualitatively, yet more finely than what confidence levels allow (Wheaton 2012). Figure 1 presents three such spectrums. Over time, this guidance has become increasingly specific. Note that while these spectrums mitigate the kinds of extreme miscommunication that appeared in the Bay of Pigs episode, the difference is one of degree and not kind. To use the spectrums in Figure 1 correctly, analysts must specify their probability assessments precisely enough to select the appropriate term, and then coarsen those estimates with qualitative language.

[Figure 1]

Numeric expressions. Though numeric probability assessments are rare in published intelligence analysis,[2] the debate over bin Laden's location shows how quantitative expressions of subjective probability appear in important settings. Many observers advocate the broader use of quantitative probability expressions, including numerical percentages, bettor's odds such as "5-to-1", and frequency representations such as "1-in-10" (Nye 1994; Schrage 2005; Marchio 2014; Barnes 2015).

In principle, "Words of Estimative Probability" spectrums have been recommended for use in 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. Some judgments use estimative verbs ("Tehran's decision to halt its nuclear weapons program *suggests*

[2] A review of declassified National Intelligence Estimates from 1964-94 found that 96 percent of key judgments expressed probability in ways that lacked clear quantitative equivalents (Friedman and Zeckhauser 2012).

it is less determined to develop nuclear weapons than we have been judging since 2005"). Other judgments use confidence levels ("We assess with *high confidence* that until fall 2003, Iranian military entities were working under government direction to develop nuclear weapons"). Some employ words of estimative probability ("We assess centrifuge enrichment is how Iran *probably* could first produce enough fissile material for a weapon"). Some statements include 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"). Other statements offer 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").[3] Moreover, the guidelines in Figure 1 do not apply to analysis produced outside the IC.[4]

There are several reasons to favor greater consistency in expressing estimative probability. Consistent standards facilitate accountability, evaluation, and improvement (Tetlock and Mellers 2011; Dhami et al. 2015). Consistency also facilitates clear communication: even if analysts define "words of estimative probability" in structured ways, individuals vary widely in how they intuitively process that language (Beyth-Merom 1982; Mosteller and Youtz 1990; Wark 1964; Johnson 1973). Even when respondents receive explicit lexicons, they often still interpret those terms in ways that authors did not intend (Budescu et al. 2014; Ho et al. in press). Decision makers have political incentives (Rovner 2011) and natural tendencies (Piercey 2009) to interpret ambiguous analysis in self-serving ways. And most importantly for the purposes of this paper, if vague expressions systematically sacrifice information in foreign policy analysis, consistent guidance can help to ensure that this information is not lost. Given that uncertainty surrounds virtually all important intelligence estimates, military plans, and foreign policy debates, even modest gains could bring major benefits.

Returns to precision

We define *returns to precision* as the degree to which probability assessments are more informative when evaluated at higher degrees of precision. All else being equal, we expect precision to have diminishing marginal returns. Yet there is no clear theoretical basis for predicting where these returns should become negligible. Without empirical analysis, assumptions about foreign policy analysts' ability to parse probabilities are essentially speculation.

Of course, there are other issues to consider in debates about the expression of probability in foreign policy analysis. Making judgments more precise might make it more difficult for analysts to agree on contentious issues. Yet airing disagreements can reveal discrepancies among analysts' views and encourage careful reasoning (Kent 1964; Morell 2015, 156-61).[5] More

[3] Emphasis added throughout.
[4] For instance, U.S. Army Field Manual 5-19, *Composite Risk Management* (par. 1-23) guides planners to assess risks using "five levels of probability – frequent, likely, occasional, seldom, and unlikely." The definitions of those terms are noticeably vague. For example, the word "frequent" is defined as "Occurs very often, known to happen regularly. In illustration, given 500 exposures to the hazard, expect that it will definitely happen to someone."
[5] Kent's seminal essay recalled an episode where analysts agreed to use the term "a serious possibility" in a National Intelligence Estimate, and later learned that individual analysts' beliefs ranged from 20 percent to 80 percent. Morell's recounting of debates about bin Laden's location in 2011 contains a vivid anecdote about how parsing the

importantly for the purposes of the present analysis, one cannot say whether any additional effort required to parse probabilities is "worth it" without knowing whether analysts have the ability to parse those probabilities effectively. If vague expressions of probability consistently sacrifice meaningful information, it would be difficult to justify less precision 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.

Debates about communicating probability must also consider the way that decision makers interpret foreign policy analyses. One common argument is that numbers create false impressions of scientific rigor, and that analysts should thus take care not to make their judgments more precise than what they can defend (Budescu and Wallsten 1987; Ho et al. 2014). This "congruence principle" is normatively appealing. Yet one cannot say what level of precision is "too precise" without empirically evaluating analysts' ability to parse probabilities. Once again, advancing debates about handling and mishandling estimative probability requires empirically investigating returns to precision.

Section 2. Forecasting Data from the Good Judgment Project

To our knowledge, this paper provides the first systematic analysis of returns to precision in geopolitical forecasting or in any other field. Our study employs 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 1,832 unique individuals[6] registered 888,328 forecasts in response to 380 questions administered between 2011 and 2015.[7]

IARPA and GJP collaborated in writing forecasting questions to ensure their relevance to intelligence analysis.[8] Questions covered issues 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 another 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.

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

---

term "a strong possibility" revealed that top officials held very different prior assumptions about their ability to interpret circumstantial evidence in supporting major national security decisions.

[6] For an overview of GJP and its findings, see Mellers et al. 2014, Mellers et al. 2015a, Mellers et al. 2015b, and Tetlock and Gardner 2015. For policy implications, see Tetlock et al. 2014. On GJP's statistical method, see Satopää, Baron, et al. 2014, and Satopää, Jensen, et al. 2014.

[7] GJP also administered a prediction market. 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 price.

[8] 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.

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. At the end of each year, GJP identified the top two percent of performers as "superforecasters." Superforecasters generally remained superior to other GJP respondents in all subsequent tournament years (Mellers et al. 2014).

GJP's data are uniquely well-suited to evaluating empirical claims about returns to precision 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 that forecasters employed. Thus GJP does not provide a true experimental comparison of numerical percentages versus WEPs, confidence levels, or estimative verbs. Nonetheless, 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, all of the proposals discussed in Section 1 require approximate numerical reasoning if they are 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 percent when predicting an outcome that occurred. 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 this method 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 76 days (standard deviation, 80 days) before it could be evaluated. By contrast, some foreign policy analyses, such as the U.S. Intelligence Community's *Global Trends* series mentioned above, consider much more distant futures. 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 robust across time horizons within GJP data. Our general findings on returns to precision are not driven by short-term forecasts that critics might argue are easier to address than the questions that intelligence analysts generally face.

Third, GJP only asked respondents to make forecasts, but foreign policy analysis also often requires making probabilistic statements about current or past states of the world, as illustrated by 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 accounting for additional uncertainty about how states of the world may change in the future. If predicting the future is harder than assessing uncertainty about the present and past, then our findings should be

conservative in identifying returns to precision when estimating probabilities in international affairs. Without data necessary to substantiate this claim directly, however, we emphasize that our empirical results pertain directly to *predictive* accuracy in geopolitical *forecasting*, which is a subset of foreign policy analysis.

Section 3. Measuring Estimative Precision and Predictive Accuracy

Methods for scoring predictive accuracy are the subject of a large literature. The key to scoring predictive accuracy is to evaluate a large volume of data. For example, suppose an analyst says it is 80 percent likely that French President Francois Hollande will be reelected, but Hollande instead loses the election. It is not possible to say how much we should attribute this discrepancy to bad assessment versus bad luck. However, if we examine a large volume of estimates, we can examine whether events this analyst says are 80 percent likely actually occur roughly 80 percent of the time. Similarly, we can examine whether outcomes to which forecasters assigned a 54 percent percent probability actually occur more often than outcomes to which they assigned a 50 percent probability. In this section, we describe our method for evaluating the extent to which such distinctions are meaningful.

A *strictly proper scoring rule* evaluates probability assessments in a manner that gives respondents an incentive to report their true beliefs. Our main metric for measuring predictive accuracy in this paper is the commonly-used Brier Score. (We also explore an alternative, logarithmic scoring rule.[9]) 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. The formula is $(1/N) \cdot \sum_1^N (Y_n - p_n)^2$, where N is the number of predicted outcomes.[10]

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: 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, the forecaster's score would be $[(1-0.60)^2 + (0-0.40)^2]/2 = 0.16$. If Assad remains, 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 forecasts, indicating that respondents assign higher probabilities to events that occur and lower probabilities to events that do not occur.

To translate numerical forecasts into corresponding verbal expressions, we round probability assessments to the midpoint of the "bin" that each verbal expression represents. For example, if analysts use the five-step "words of estimative probability" spectrum in Figure 1, the phrase "even chance" implies a predicted probability between 40 and 60 percent. Absent additional information, the expected value of a probability estimate falling within this range is 50 percent.

[9] We believe the Brier Score is a more appropriate measure for our purposes because of the severe penalties which the logarithmic scoring rule assigns to misplaced extreme estimates. Logarithmic scoring requires changing estimates of 0.00 and 1.00 (comprising 19 percent of our data points), since an error on these estimates imposes an infinite penalty. See Section 4.

[10] We only use data from questions with binary outcomes ($N = 2$).

In practice, a decision maker may combine this estimate with other information and prior assumptions to justify a prediction that is higher or lower than 50 percent.[11] 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 (Savage 1954).[12] We generalize this approach by dividing the number line into B bins, then rounding each forecast to the midpoint of its associated bin. When forecasts fall on boundaries between bins (e.g., a forecast of 50 percent when $B = 2$), we randomize the direction of rounding.[13]

Using forecasting questions as the unit of analysis

Though our data comprise 888,328 forecasts, these forecasts are correlated within questions and within individuals who updated forecasts before questions closed for evaluation.[14] It would be inappropriate to treat all forecasts in our data set as representing independent observations.

We thus take the forecasting question as our unit of analysis. We identify a subset of forecasters to evaluate (all forecasters, superforecasters, etc.). 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 γ is a subset of GJP forecasters; j is a forecasting question; $mean(\cdot)$ is the mean of a vector; i is a forecaster; 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;[15] $and\ Brier_{ijk}$ is the Brier Score for an estimate made by a given forecaster on a given question on a given day. Thus, $x_{\gamma j}$ 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 888,328 forecasts to 380 forecasting questions. Evaluating individual forecasts returns similar estimates of returns to precision, albeit with inappropriately small p-values when making comparisons.[16]

[11] Thus, while the word "remote" applies to estimates from 0 to 20 percent under the five-bin system of WEPs, decision makers might anticipate that analysts using this term are attempting to convey a probability closer to zero. (This is presumably one of the problems that the DNI attempted to solve with the 2015 WEP guidelines shown in Figure 1.) We examined this issue by rounding estimates to the empirical expected probability of forecasts falling within each bin. This alternative reduces rounding errors for most groups of forecasters, but still leaves statistically significant losses of accuracy consistent with our other findings.

[12] Ellsberg 1961 showed that many decision makers are "ambiguity averse": in practice they *do* treat an estimate of "50 percent" differently from an estimate of "40 to 60 percent." Ellsberg also explained why ambiguity aversion is irrational.

[13] Though the WEP spectrum defined by the DNI in 2015 defines "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.

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

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

[16] 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. In Section 4, we show that these "lay-up" forecasts do not drive our results.

Calculating "rounding errors"

The simplest way to evaluate returns to precision is to calculate how often forecasts become less accurate when we round them off.[17] We report these data below: for example, rounding superforecasters' estimates into seven equally-spaced "words of estimative probability" reduces their accuracy on 94 percent of questions. One drawback of this approach, however, is that it potentially rewards overconfidence. For example, imagine that forecasters generally assign estimates of 80 percent to outcomes that occur 70 percent of the time. Rounding those estimates down to 70 percent might make a majority of forecasts seem worse, but it would actually improve overall calibration. This is why it is important to measure predictive accuracy using proper scoring rules like Brier Scores.

We calculate *rounding errors* on forecasting questions by measuring proportional changes in Brier Scores when we round individual forecasts into bins of different widths. Thus, we define $\tilde{x}_{\gamma jB}$ as a question-level point-estimate of forecasting accuracy among forecasters in group γ on question j, having rounded individual respondents' forecasts to the midpoints of B bins. For example, we estimate the rounding error associated with transforming probabilities into three-step "confidence levels" as $(\tilde{x}_{\gamma jB=3} - x_{\gamma j})/x_{\gamma j}$. Our findings also hold when we round predictions to the empirical expected value (that is, the frequency-weighted mean) of forecasts falling within each bin.

We calculate proportional changes in predictive accuracy to alleviate the asymmetrical penalties imposed by rounding in different regions of the probability scale. We also describe both mean and median rounding errors, and show that our results are similar if we use a logarithmic scoring rule. These analyses help to ensure that when we estimate the degree to which rounding probabilistic assessments influences their predictive value, our findings represent consistent losses of information as opposed to the impact of extreme data points.[18]

Predictive accuracy and decision quality

Enhancing predictive accuracy will not always improve decision quality. Yet this is no reason not to seek gains wherever possible. In fact, the difficulty of anticipating where changes in informational quality are most likely to impact decision making is exactly why it is important to seek broad improvements in foreign policy analysis.

When considering drone strikes or special forces missions, for example, decision makers continually wrestle with whether the 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.

[17] Thus, rather than calculating aggregate Brier Scores, we estimate an average forecast for any group of forecasters on any forecasting question, and we examine how often that aggregate forecast draws closer to or further from the realized outcome when we round off the raw data.

[18] As described below, we measure statistical significance using traditional comparisons of means and Wilcoxon signed-rank tests, which reduce the sensitivity of empirical tests to changes in scoring rule.

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

Section 4. Rounding Errors across Modes of Expression

Table 1 shows how rounding forecasts to different degrees of (im)precision reduces their predictive accuracy. Rounding numeric assessments to "confidence levels" or "estimative verbs" substantially removes information from GJP forecasts. On average, GJP forecasts become 31.4 percent worse when rounded into two bins. This change is not driven by outliers, as the median rounding error is 22.1 percent. Even the worst-performing group of forecasters, untrained individuals, incurs an average rounding error of 15 percent when we rounded their forecasts 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. Even though we adopted an intentionally conservative approach to estimating statistical significance, every subgroup in our analysis encounters consistent ($p<0.001$) losses of predictive accuracy when we round forecasts according to the lexicon currently recommended by the U.S. Director of National Intelligence. The IC's previous "words of estimative probability" guidelines, which divide the unit interval into seven equal bins, induce greater variance: rounding errors here tend to be larger and less consistent.[19] Superforecasters continue to suffer the largest losses under both "words of estimative probability" rounding systems. Coarsening probability assessments thus prevents the best forecasters from reaching their full potential, sacrificing information disproportionately from the sources that produce the most reliable assessments.

These comparisons are especially meaningful in relation to the challenges which scholars generally face when evaluating methods of intelligence estimation. Mark Lowenthal (2008, 314), a 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]." Thomas Fingar (2011, 34, 130), formerly the IC's top analyst, writes that "By and large, analysts do not have an empirical basis for using or eschewing particular methods." By contrast, our results *do* provide an empirical basis for expressing probabilities more precisely than what current IC practice allows. Geopolitical forecasting may

[19] The new DNI 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.

be subjective, but our data indicate that when GJP participants responded to questions posed by the IC, their views were systematically more informative at higher degrees of 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 important to determine whether returns to precision appear primarily when making "easy" forecasts. Two main indicators of forecasting ease are the forecast's size (more extreme probabilities reflect greater certainty, which should correlate with easier questions) and its time horizon (nearer-term events should be easier to predict). We address these two subjects in turn. Our results show that GJP respondents extract returns to precision across a broad range of forecasts.

[Figure 2]

Figure 2 presents a histogram of GJP forecast values.[20] As a general rule, GJP forecasters assigned estimates at intervals of five percentage points.[21] This pattern alone is important, because it indicates that when left without restrictions on how granular (that is, how fine-grained) their forecasts should be, GJP respondents prefer to express probabilities more finely than common qualitative expressions allow.

Figure 2 also demonstrates that GJP forecasters were especially willing to make fine-grained forecasts when predicting probabilities close to zero or one. Since low-probability, high-consequence events represent some of the most crucial 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."[22] We then calculated rounding errors for each question using only forecasts that fell into a particular bin. This allows us to examine how much information forecasters lose by employing each individual term on this spectrum.

[Table 2]

Table 2 shows that GJP analysts, on the whole, demonstrate returns to precision across the number line. We find that superforecasters can reliably employ numeric precision within each individual "word of estimative probability" category. Non-superforecasters achieve mixed results from rounding their most extreme estimates: here, Table 2 shows a loss of information on average but a gain of information at the median. This pattern clearly indicates that our aggregate rounding errors are not driven by the most extreme forecasts in our data set.

[20] The histogram is symmetric because predicting that an outcome will occur with probability p implies that the outcome will *not* occur with probability $1 - p$. The histogram does not reflect how long those estimates remained active before respondents revised them or before questions closed.
[21] Forty-nine percent of forecasts in the data set are multiples of 0.10 and 25 percent of forecasts are additional multiples of 0.05. Table 4 shows that this is not being driven by a small set of highly-granular forecasters.
[22] Thus we replicate our previous analysis while limiting the data set only to forecasts within a given range.

Table 2 also shows that these results also do not hinge on Brier scoring's nonlinear properties. When we recalculate rounding errors using a logarithmic scoring rule,[23] we again find that superforecasters exhibit reliable returns to precision in every category, and that rounding sacrifices predictive accuracy for non-superforecasters in every category besides the extremes.[24] These findings reinforce the proposition that foreign policy analysts can extract returns to precision on a wide range of forecasts, and that our general findings are not driven by responses to "easy" questions that respondents could address with certainty.

Returns to precision across time horizons

Next, we explore whether returns to precision depend on short-term forecasts. 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. In our data set, the mean time horizon was 76 days (standard deviation 80 days, median 48). 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 were 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 equal numbers of observations: *Short-Term* forecasts were made less than 36 days before questions closed (excluding Lay-Ups); *Medium-Term* forecasts were made from 36 to 96 days prior to closing; *Long-Term* forecasts were made more than 96 days prior to questions closing.[25]

[Table 3]

Table 3 presents results. Removing Lay-Up forecasts from the analysis has limited impact on aggregate rounding errors. Lay-Ups are clearly not driving our overall results. Moreover, while rounding errors decline as we limit our analysis to long-term forecasts alone, the same basic pattern persists in this subset of the data: rounding probability assessments into confidence levels and estimative verbs sacrifices sizable and statistically significant amounts of information; "words of estimative probability" recoup some but not all of these losses; and returns to precision remain particularly large for high-quality forecasters.[26]

---

[23] This rule scores analysts' predictions according to the natural logarithm of the probability they assigned to the observed outcome. Thus if an analyst predicts a 60 percent chance that Hollande is reelected and this happens, then the analyst's score is $\ln(0.60)$, whereas if Hollande is not reelected, then the analyst's score is $\ln(0.40)$. Higher logarithmic scores are better. In order to prevent scores of $-\infty$, we convert estimates of 1.00 and 0.00 to 0.99 and 0.01, respectively.

[24] The benefits to rounding non-superforecasters extreme estimates increase under logarithmic scoring because of the way that this function imposes severe penalties on erroneous estimates made with near-certainty. If non-superforecasters have a tendency towards overconfidence in their most extreme estimates, we would expect this to result in a larger penalty under logarithmic scoring relative to Brier scoring.

[25] There were 109,240 "lay-ups" in our data, and 259,696 forecasts in each of the short-, medium-, and long-term categories.

[26] Median rounding errors for short- and medium-term forecasts are ordered as expected: the shorter the time horizon the greater the rounding error (though Table 3 shows that rounding errors are still statistically significant on

Section 5. Variation across Individuals

In this section, we analyze which attributes predict individual differences in returns to precision. We examine variables capturing skill, effort, experience, preparation, and cognitive style. We chose these variables not just because they plausibly explain individual variation in returns to precision, but also because they shed light 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.[27] If these factors predict individual-level returns to precision, this finding would be hopeful for 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 determinants of returns to precision, then organizations might seek to capture these skills mainly through selecting personnel. In the 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. We expect that higher-quality forecasters would incur greater penalties from having their forecasts rounded. Four additional variables capture effort, training, and experience. *Number of 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 parse probabilities more effectively.

Average Revisions per Question captures how often respondents updated their beliefs on each forecasting question. This variable proxies for effort and engagement with GJP; 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. We expect that respondents who are comfortable expressing their views precisely would incur larger rounding penalties than forecasters who provided coarser judgments.[28] *Probabilistic Training* takes a value of 1 if the forecaster received training in probabilistic reasoning from GJP. 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.

Two variables capture respondents' education prior to participating in the Good Judgment Project. *Education Level* is a four-category variable capturing respondents' highest academic degree (1: no bachelor's; 2: bachelor's; 3: master's; 4: doctorate).[29] Advanced education could enhance respondents' abilities to analyze complex questions and to parse probabilities reliably.

---

the long-term forecasts). *Average* rounding errors on medium-term forecasts 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.

[27] The prospect for improving forecasting skill, even with relatively limited training, is well-established in the decision science literature. See Alpert and Raiffa 1982 and Dhami et al. 2015, among others.

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

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

Numeracy represents respondents' scores on a series of word problems designed to capture mathematical fluency (Lipkus et al. 2001; Peters et al. 2006). If respondents are better able to reason numerically, they might be able to parse probabilities more effectively.[30] 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 session, for instance, lasted just one hour.)

GJP data include several indices of "cognitive style," including: *Raven's Progressive Matrices*, where higher scores indicate better reasoning ability (Arthur et al. 1999); an expanded Cognitive Reflection Test (*Expanded CRT*), where higher scores indicate an increased propensity to suppress misleading intuitive reactions in favor of more accurate, deliberative answers (Baron et al. in press); *Fox-Hedgehog*, a variable where higher scores capture respondents' self-assessed tendency to rely on ad hoc reasoning versus simplifying frameworks (Mellers et al. 2015a); and *Need for Cognition*, an index of respondents' self-assessed preference for addressing complex problems (Cacioppo and Petty 1982).[31] 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 had worse Brier Scores than respondents' original predictions. We define each individual's *threshold of estimative precision* (B^*) as the smallest number of bins where rounding errors are not statistically distinct from zero ($p < 0.05$). Thus any level of (im)precision lower than B^* systematically sacrifices predictive accuracy.

Analysis

Table 4 shows summary statistics for each variable, including these variables' correlation with individual B^* thresholds across 1,832 forecasters in our sample.[32] All bivariate correlations are in the expected direction. Generally speaking, variables capturing skill, effort, training, and experience are more closely correlated with individual-level returns to precision than variables capturing education and cognitive style.

[Table 4]

Table 5 presents ordinary least squares regression analyses predicting individual B^* thresholds. We standardized non-binary independent variables. Each coefficient in Table 5 thus reflects the extent to which B^* thresholds increase, on average, when each predictor increases by one standard deviation, or when the Training variable changes from 0 to 1. We include Brier

[30] GJP changed numeracy tests between years 2 and 3 of the competition. We standardized numeracy test results so that they represent comparable indices. If a respondent participated in multiple years of the forecasting competition, we averaged Numeracy values across years.
[31] If a respondent participated in multiple competition years, we averaged values across years. GJP changed CRT tests after Year 2 of the competition, so we standardized each test's results in order to provide comparable measures.
[32] We exclude forecasters who made less than 25 forecasts in a given competition year.

Score in all models. The purpose of Table 5 is to examine how different combinations of variables describe returns to precision across respondents.

Model 1 demonstrates that forecasting skill alone predicts substantial variation in individual-level returns to precision (R^2=0.21). Model 2 shows that adding variables for effort and training substantially improves model fit (R^2=0.34). In particular, the variables for Number of Questions, Average Revisions per Question and Probabilistic Training are statistically significant predictors of individual-level returns to precision.[33] By contrast, Model 3 shows that our education and cognitive style variables predict almost no individual-level variation in returns to precision. None of the targets for selection variables is statistically significant in Model 3, although Raven's Progressive Matrices (p=0.07) and Expanded CRT (p=.10) approach the standard threshold.

[Table 5]

When we examine all predictors together in Model 4, Need for Cognition (p=0.04) is the only "Target for Selection" that retains statistical significance, while the Average Revisions per Question (p=0.07) variable falls just outside the usual statistical significance threshold. Model 5 then replicates our analysis with the "Targets for Cultivation" variables using only the 1,307 observations for which we have data on all ten variables. We find that the "Targets for Selection" variables increase R^2 by just 0.0035 over a regression including the "Targets for Cultivation" variables alone.[34] Once again, the Average Revisions per Question (p=0.07) variable falls slightly short of the standard statistical significance threshold in Model 5.[35]

Two practical implications emerge from these results. 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 demonstrated substantially higher returns to precision than their peers. (The magnitude of this correlation is about as large as what we would predict by increasing our cognitive style attributes by three standard deviations each.) Especially since this training was randomly-assigned, our findings suggest that the IC and other organizations could replicate and presumably exceed this benefit by training their own personnel.

We also found that respondents' experience making forecasts and their willingness to revise those forecasts consistently predicted higher returns to precision. These findings provide additional grounds for optimism that professional forecasters could replicate and potentially exceed the returns to precision shown in GJP's data. Foreign policy analysts are full-time professionals who 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 twice per question).

It is not surprising that Number of Questions predicts B^* thresholds. Forecasters who registered more predictions were not only more experienced and more engaged in the competition, but they also provided more forecasts for calculating B^* thresholds such that

[33] 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.
[34] A likelihood ratio test cannot reject the hypothesis that Models 4 and 5 have identical model fit (p=0.23).
[35] Estimating Model 1 in a sample with those same 1,307 observations only returns a coefficient for Brier Scores of -2.11 (.27)*** and a constant term of 4.10 (.10)***, with R^2 and AIC scores of 0.22 and 7,165, respectively.

smaller rounding errors would register as being statistically significant. Our analyses cannot distinguish the extent to which this correlation results from sample size versus experience. Yet either interpretation has the same practical implication: the more forecasts an analyst makes, the more likely it becomes that rounding off her estimates will systematically sacrifice information. Given the vast quantity of uncertain judgments that the IC produces, the relationship we observe between Number of Questions and returns to precision further emphasizes that GJP data may understate the degree to which professional forecasters could achieve meaningful returns to precision by quantifying probability assessments.

Second, and no less important, our findings reject the notion that returns to precision correlate with innate individual-level attributes. While the intelligence literature frequently distinguishes between "mathematicians" and "poets" (Kent 1964; Johnston 2005), 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. Rather, our data suggest that when skilled forecasters of all kinds take the time and effort to make precise forecasts, this adds information to foreign policy analysis.

Section 6. Variation across Questions

We also coded B^* thresholds for each forecasting question that GJP posed.[36] This variable had a mean of 6.1 bins (standard deviation 4.4). B^* thresholds were greater than 7 bins for 42 percent of questions. This finding reinforces the argument that foreign policy analysts can achieve returns to precision on a wide range of questions. Our results do not simply hinge on a few questions where forecasters happened to make particularly informative estimates.

Nevertheless, there might still be clusters of questions that are particularly amenable (or resistant) to precise estimation. For example, questions relating to economics and finance could lend themselves to quantitative analysis in a way that analyzing diplomacy and armed conflict does not. Similarly, GJP respondents might have found it much easier to answer questions relating to North American or European issues as opposed to African or South Asia issues, with which they may have been less familiar. If these kinds of distinctions predict substantial variation in question-level B^* thresholds, this would suggest that even if returns to precision are real, they are also limited to particular subsets of foreign policy analysis.

Addressing this issue calls for inductive analysis of whether returns to precision correlate with particular question types. To conduct such an analysis, we employed data from Horowitz et al. (2015), who classify the content of each GJP question with respect to 11 "region" tags and 15 "function" tags. The region tags corresponded to Sub-Saharan Africa, Central/South America, North America, South/Central Asia, East/Southeast Asia, Eastern Europe, Western Europe, Middle East/North Africa, Oceania, Global, and the Arctic. The function 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. Tags

[36] The data analyzed in this section span 375 questions; key data were missing for the remaining five questions in our data set.

were not mutually exclusive. Table 6 describes the incidence of each regional and functional topic across GJP questions.

We examined these variables in ordinary least squares regressions predicting question-level B^* thresholds based on dummy variables for question type. Table 6 presents results. Model 1 combines all tags within a single regression. Models 2 and 3 examine region and function tags, respectively. Model 4 optimizes model fit, as measured by AIC score.

This purely inductive analysis is not intended to advance a theoretical framework for explaining question-level returns to precision. Rather, Table 6 indicates the extent to which returns to precision belong to identifiable subsets of question types. And while all four models identify statistically significant patterns, none produces a particular high model fit. When we examine all 26 question tags simultaneously – in a regression that clearly entails overfitting – the model's R^2 is just 0.16. The constant term remains stable across models, indicating that baseline returns to precision are relatively unaffected when controlling for up to 26 question types.

Thus even in a purely inductive effort to identify how returns to precision vary across questions, we see little indication that forecasters' ability to specify their estimates is confined to particular topics. These findings reinforce our broader argument that foreign policy analysts can consistently parse probabilities more finely than what common systems of qualitative expression allow. The final section connects these findings to practical debates about foreign policy analysis and probability assessment in other fields.

Section 7. Discussion

This paper demonstrates that foreign policy analysts can consistently estimate probabilities more precisely than what conventional wisdom supposes and what standard practices allow. Coarsening probability estimates into "estimative verbs," "confidence levels" or "words of estimative probability" systematically sacrifices predictive accuracy. These findings do not depend on extreme forecasts, short time horizons, particular scoring rules, or special question types. Qualitative expression is most harmful to the highest-quality forecasters, but the ability to parse probabilities is not unique to analysts with special educational backgrounds or quantitative skills. Our findings therefore suggest that it is possible to improve the informational value of intelligence estimates and other foreign policy analyses by supplementing qualitative descriptions of uncertainty with quantitative estimates of subjective probability.

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

First, foreign policy analysts should express probabilities more precisely than "confidence levels" and "estimative verbs." Our data indicate that these crude expressions sacrifice substantial information.

Second, our data suggest that quantitative precision is empirically justifiable when assessing probabilities in foreign policy analysis. Existing "words of estimative probability" spectrums, including those currently recommended for use by U.S. intelligence analysts, do not allow foreign policy analysts to express distinctions that they can consistently employ.

This finding does not end the debate about the proper means for communicating uncertainty in intelligence and foreign policy decision making. Yet one of the most common arguments for using "confidence levels" or "words of estimative probability" is that additional precision simply represents random noise. Our data refute this argument. We thus believe that the burden of proof rests with opponents of numerical probabilities to justify why it is worth sacrificing the informational gains that quantitative precision provides. This argument applies not only to intelligence analysts and foreign policy officials, but also to scholars and pundits who participate in public debates about international affairs. Although these discussions shape voters' opinions and public policy, they are often just as vague as intelligence reporting, and sometimes even more so (Tetlock 2009; Gardner 2011; Silver 2012; Tetlock and Gardner 2015).

Third, our data show that there are practical benefits to cultivating skills in probability assessment. GJP forecasters who were randomly assigned to just one hour of training in probabilistic reasoning achieved significantly higher returns to precision than did their peers. More generally, as discussed in Section 5, our findings suggest that if the IC and other organizations prioritized this subject, they could achieve even better performance than what was observed in our data.

Most broadly, while our study is motivated by long-standing debates about foreign policy analysis, our approach applies to any field where scholars and practitioners debate proper means of probability assessment. 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 (Nakao and Axelrod 1983; Braddock et al. 1999; Politi, Han, and Col 2007).

While our empirical findings do not apply directly to other domains, our results suggest that other disciplines should also revisit basic skepticism about probabilistic precision. Foreign policy analysis is widely considered to be an area in which probability assessment is unusually difficult. International politics involves a large number of variables that interact in complex, nonlinear ways within contexts that are frequently unique. Foreign policy analysts generally lack access to broadly-accepted theoretical models or to large, well-behaved data sets on recurring situations for grounding their inferences (Beyerchen 1992/93; Fingar 2011). By comparison, analysts in professions such as law, medicine, and finance often have much stronger bases for defining reference classes, for estimating base rates, and for employing analytic tools to assist with assessing uncertainty. At the very least, this paper demonstrates that the value of precision in probability assessment is ultimately an empirical question, while advancing a method for examining the level of specificity that is achievable in many fields.

References

Arthur, W., Jr., T. C. Tubre, D. S. Paul, and M. L. Sanchez-Hu. 1999. College-Sample Psychometric and Normative Data on a Short Form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment* 17 (4): 354-361.

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.

Baron, J., S. E. Scott, K. Fincher, and S. E. Metz. 2015. Why Does the Cognitive Reflection Test (Sometimes) Predict Utilitarian Moral Judgment (and Other Things)? *Journal of Applied Research in Memory and Cognition*. In press.

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*. New York: Columbia University Press.

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*. New York: 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., 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.

Budescu, David V., Han-Hui Por, Stephen B. Broomell, and Michael Smithson. 2014. The Interpretation of IPCC Probabilistic Statements Around the World. *Nature Climate Change* 4: 508-512.

Budescu, David V. and Thomas Wallsten. 1987. Subjective Estimation Based on Precise and Vague Uncertainties. In G. Wright and P. Ayton (eds.), *Judgmental Forecasting*. New York: Wiley.

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

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*. Washington, D.C.: U.S. Government Printing Office.

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

Dhami, Mandeep K., David R. Mandel, Barbara A. Mellers, and Philip E. Tetlock. 2015. Improving Intelligence Analysis with Decision Science. *Perspectives on Psychological Science* 10 (6): 743-757.

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

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

Friedman, Jeffrey A. 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*. New York: Plume.

Ho, Emily H., David V. Budescu, Mandeep K. Dhami, and David R. Mandel. In Press. On the Effective Communication of Uncertainty: Lessons from the Climate Change and Intelligence Analysis Domains. *Behavioral Science and Policy*.

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.

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

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

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

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

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

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* 111 (30): 10984-10989.

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*. New York: Twelve.

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

Nakao, Michael A. and Seymour Axelrod. 1983. Numbers Are Better Than Words: Verbal Specifications of Frequency Have No Place in Medicine. *American Journal of Medicine* 74: 1061-65.

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.

Piercey, M. David. 2009. Motivated Reasoning and Verbal vs. Numerical Probability Assessment: Evidence from an Accounting Context. *Organizational Behavior and Human Decision Processes* 108 (2): 330-341.

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

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.

Rovner, Joshua. 2011. *Fixing the Facts: National Security and the Politics of Intelligence*. Ithaca, N.Y.: Cornell University Press.

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.

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley.

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*. New York: 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.

Sunstein, Cass R. 2014. *Valuing Life: Humanizing the Regulatory State*. Chicago, IL: University of Chicago Press.

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

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

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): 542-554.

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.

Wark, David L. 1964. The Definition of Some Estimative Expressions. *Studies in Intelligence* 8 (4).

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

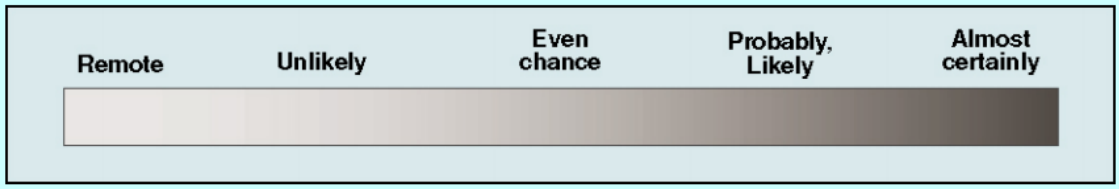
Wheaton, Kristan J. 2012. The Revolution Begins on Page Five: The Changing Nature of NIEs. *International Journal of Intelligence and CounterIntelligence*. 25 (2): 330-349.

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

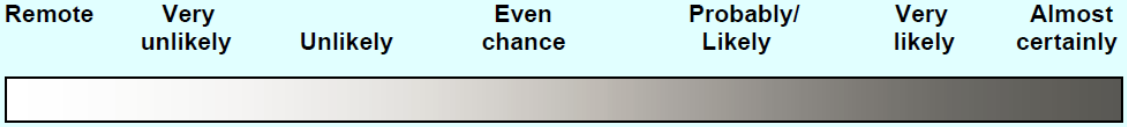
Figure 1. "Words of Estimative Probability"

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%

Figure 2. Histogram of forecasts in GJP data

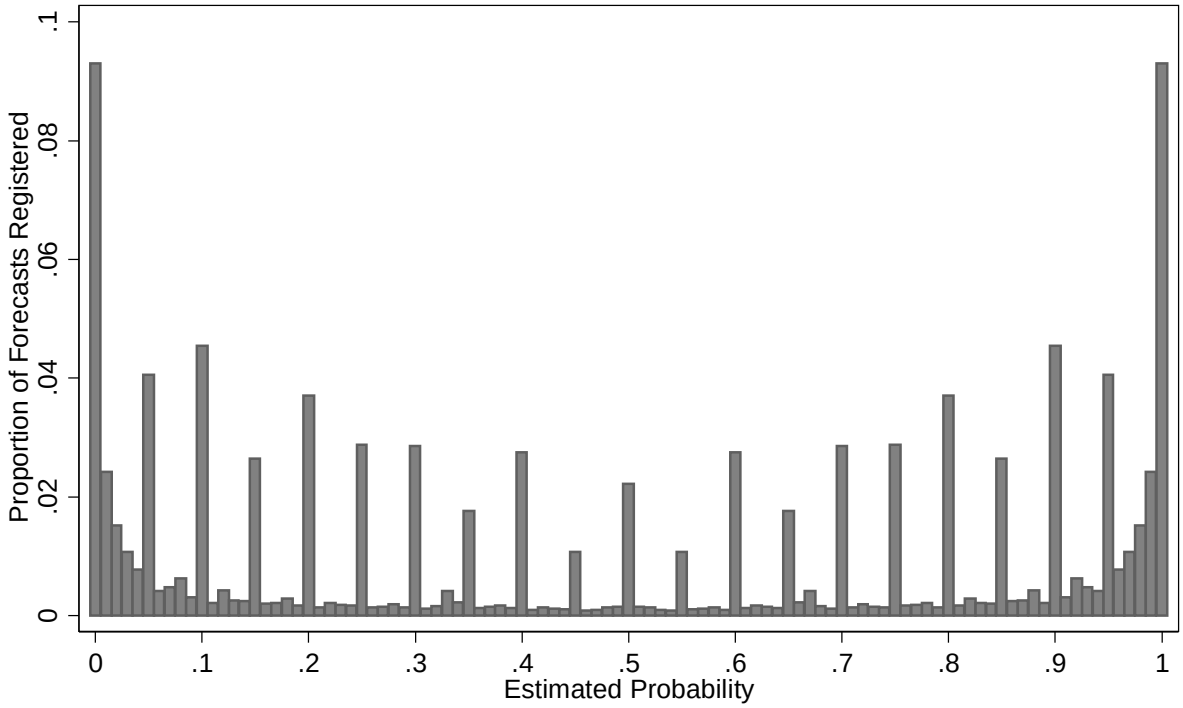


Figure 3. Distribution of forecasts by time horizon

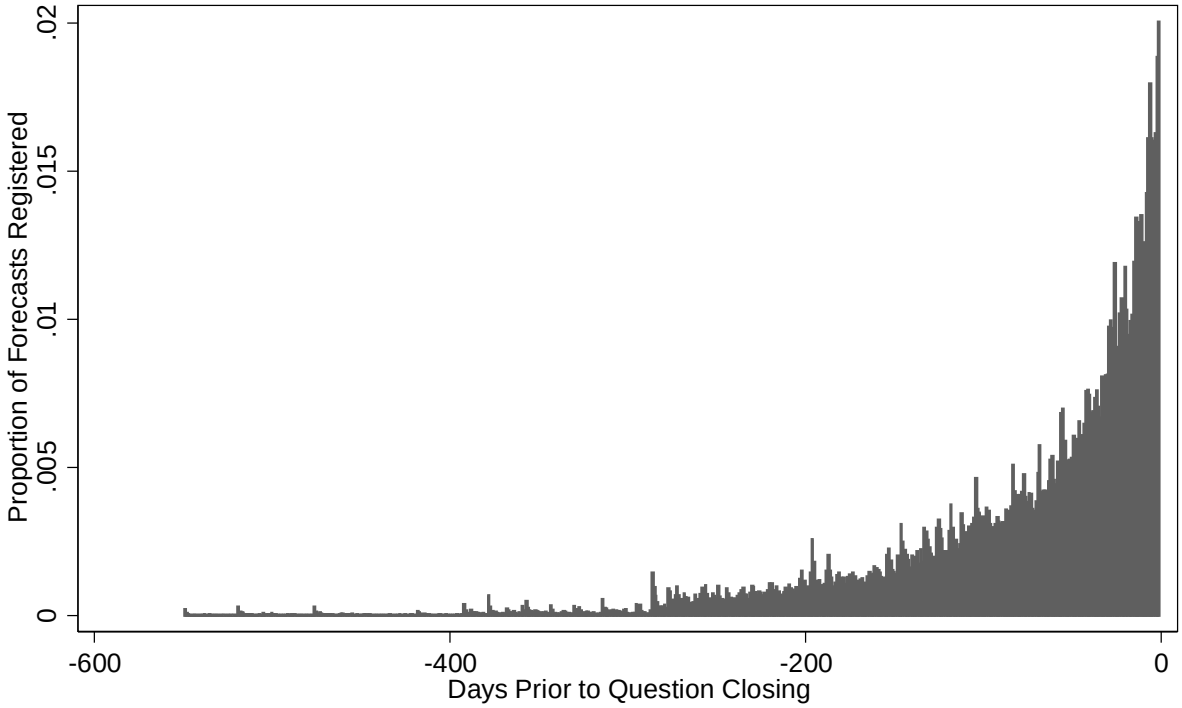


Table 1. Estimative Precision and Predictive Accuracy – Aggregated Results

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:*	0.153	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%***
			Proportion of Aggregate Forecasts Degraded By Rounding			
All forecasters			72%	92%	93%	91%
Super-forecasters			84%	94%	93%	93%

* p<0.05, ** p<0.01, *** p<0.001.
†Currently recommended by the Office of the Director of National Intelligence (see Figure 1).

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

		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)
Rounding Errors via Brier Scoring								
All Forecasters	*Mean:*	3.4%***	4.3%***	2.3%***	1.3%***	2.3%***	4.3%***	3.4%***
	Median	-0.5%***	3.7%***	2.2%***	1.1%***	2.2%	3.7%***	-0.5%***
Super-Forecasters	*Mean:*	85.8%***	16.2%***	7.0%***	1.8%***	7.0%***	16.2%***	85.8%***
	Median	32.2%***	12.1%***	4.1%***	1.0%***	4.1%***	12.1%***	32.2%***
Rounding Errors via Logarithmic Scoring								
All Forecasters	*Mean:*	-1.1%***	3.5%***	1.6%***	0.9%***	1.6%***	3.5%***	-1.1%***
	Median	-7.5%***	3.7%***	1.7%***	0.8%***	1.7%***	3.7%***	-7.5%***
Super-Forecasters	*Mean:*	70.4%	9.9%***	4.4%***	1.2%***	4.4%***	9.9%***	70.4%
	Median	55.0%***	9.4%***	3.1%***	0.7%***	3.1%***	9.4%***	55.0%***
Proportion of Aggregate Forecasts Degraded By Rounding								
All Forecasters		95%	91%	87%	69%	87%	91%	95%
Super-Forecasters		94%	90%	80%	58%	80%	90%	94%

* p<0.05, ** p<0.01, *** p<0.001.

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						
72% of all aggregate forecasts degraded by rounding to 7 bins 92% of superforecasters' aggregate forecasts degraded by rounding to 7 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%***
Super-forecasters	*Mean:*	0.093	6.1%***	40.4%**	236.1%***	562.0%***
	Median:	0.032	1.7%***	10.2%***	54.7%***	141.7%***
All forecasts (excluding "lay-ups")						
88% of all aggregate forecasts degraded by rounding to 7 bins 90% of superforecasters' aggregate forecasts degraded by rounding to 7 bins						
All forecasters	*Mean:*	0.165	0.8***	1.6%*	8.7%***	24.1%***
	Median:	0.134	1.0***	1.2%***	5.4%***	15.5%***
Super-forecasters	*Mean:*	0.102	4.4***	29.1%***	174.0%***	422.7%***
	Median:	0.039	1.6***	7.8%***	37.4%***	108.0%***
Long-term forecasts: ≥97 days						
83% of all aggregate forecasts degraded by rounding to 7 bins 85% of superforecasters' aggregate forecasts degraded by rounding to 7 bins						
All forecasters	*Mean:*	0.187	0.7%***	1.2%	6.6%***	18.3%***
	Median:	0.155	0.9%***	0.8%*	3.7%***	11.8%***
Super-forecasters	*Mean:*	0.119	0.9%***	13.5%**	78.2%***	204.4%***
	Median:	0.047	0.7%**	6.3%***	22.6%***	72.9%***

$^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$.

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^*)	1,832	4.36	8.03	1.00	101.00	-
Forecasting Skill						
Median Brier Score	1,832	0.24	0.13	0.00	1.00	-0.46
Effort, Training, Experience						
Number of Questions	1,832	85.13	62.92	25.00	375.00	0.31
Average Revisions per Question	1,832	2.49	5.03	1.00	101.4	0.20
Granularity	1,832	0.50	0.19	0.00	1.00	0.02
Probabilistic Training	1,832	0.65	0.48	0.00	1.00	0.16
Education						
Education Level	1,818	1.87	0.80	0.00	3.00	0.02
Numeracy	1,810	-0.03	0.93	-4.80	0.87	0.00
Cognitive Style						
Raven's Progressive Matrices	1,813	7.71	2.67	0.00	12.00	0.07
Cognitive Reflection Test	1,617	-0.05	0.99	-3.71	1.11	0.13
Fox-Hedgehog	1,640	2.28	1.01	1.00	5.00	0.03
Need for Cognition	1,648	5.73	0.65	3.33	7.00	0.09

Twenty 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	-1.84 (.17)***	-1.74 (.16)***	-2.07 (.28)***	-1.99 (.28)***	-2.02 (.27)***
Number of Questions		1.31 (.10)***		1.31 (.10)***	1.32 (.10)***
Average Revisions per Question		0.48 (.21)*		0.61 (.34)	0.62 (.34)
Granularity		-0.08 (.10)		0.09 (.14)	0.11 (.14)
Probabilistic Training (dummy)		1.05 (.16)***		1.17 (.22)***	1.14 (.21)***
Targets for selection					
Numeracy			0.02 (.11)	-0.02 (.10)	
Education Level			0.14 (.11)	0.02 (.10)	
Raven's Progressive Matrices			0.21 (.12)	0.11 (.11)	
Cognitive Reflection Test			0.03 (.11)	0.03 (.11)	
Fox-Hedgehog			0.14 (.11)	0.09 (.09)	
Need for Cognition			0.19 (.11)	0.21 (.10)*	
Constant	3.88 (.08)***	3.20 (.12)***	4.09 (.10)***	3.06 (.15)***	3.09 (.15)***
N	1,832	1,832	1,307	1,307	1,307
R^2	0.21	0.34	0.23	0.37	0.37
AIC	9,885	9,551	7,164	6,901	6,898

Ordinary least squares regression predicting B^* thresholds for individual respondents. Non-binary independent variables standardized. Robust standard errors. * p<0.05 ** p<0.01 ***p<0.001.
[†]Model 5 only retains observations available in Models 3-4.

Table 6. Predicting Question-Level Returns to Precision

	Model 1: *All Tags*	Model 2: *Region Tags*	Model 3: *Function Tags*	Model 4: *Optimal AIC*
Africa (*N*=34)	-0.88 (.99)	-0.91 (1.01)		
Central/South America (18)	1.02 (1.46)	0.81 (1.54)		1.31 (1.05)
North America (23)	2.21 (1.20)	1.64 (1.09)		2.45 (1.00)*
South/Central Asia (18)	0.08 (1.30)	0.18 (1.31)		
East/Southeast Asia (80)	-0.38 (.85)	-0.30 (.86)		
Eastern Europe (57)	-0.18 (.72)	-0.19 (.70)		
Western Europe (56)	-1.36 (.75)	-1.86 (.76)*		-1.02 (.63)
Global (6)	-2.02 (1.22)	-1.81 (1.40)		
Mid. East/North Africa (125)	-0.59 (.81)	-0.17 (.77)		
Oceania (5)	4.89 (1.77)**	3.92 (2.77)		4.71 (1.95)*
Arctic (1)	-0.49 (1.63)	-2.07 (1.11)		
Commodities (9)	-0.42 (1.24)		-0.94 (1.17)	
Currencies (11)	-2.75 (1.21)*		-1.84 (1.02)	-2.84 (1.37)*
Diplomatic Relations (60)	1.26 (.82)		1.41 (.83)	1.36 (.61)*
Domestic Conflict (85)	-0.38 (.74)		-0.56 (.72)	
Economic Growth (39)	-1.68 (.89)		-1.98 (.81)*	-1.41 (.75)
Elections (49)	-2.21 (.73)**		-2.44 (.71)***	-2.27 (.67)***
Int'l Organizations (33)	0.81 (.97)		0.50 (.90)	
Int'l Security (75)	0.41 (.70)		0.31 (.70)	
Leaders (47)	1.39 (.96)		1.44 (.95)	1.45 (.69)*
Public Health (3)	-2.57 (1.01)*		-3.87 (1.02)***	-3.34 (2.39)
Resources (17)	0.39 (.98)		-0.06 (.89)	
Technology (4)	2.80 (1.11)*		2.16 (.61)***	
Trade (22)	-3.41 (.83)***		-2.01 (.77)**	-3.20 (.98)***
Treaties (41)	-0.05 (.73)		-0.07 (.69)	
Weapons (28)	-0.76 (.84)		-0.71 (.81)	
Constant	6.84 (1.15)***	6.48 (.74)***	6.53 (.70)***	6.35 (.31)***
Observations	375	375	375	375
R^2	0.16	0.05	0.11	0.15
AIC	2,153	2,169	2,156	2,132

Robust standard errors. * p<0.05 ** p<0.01 ***p<0.001.