In *The Future of Risk Management*, Howard Kunreuther, Robert J. Meyer, and Erwann O. Michel-Kerjan, eds. Philadelphia: University of Pennsylvania Press, 2019, pp. 209-226.

CHAPTER 12

Improving the Accuracy of Geopolitical Risk Assessments

Barbara A. Mellers, Philip E. Tetlock, Joshua D. Baker, Jeffrey A. Friedman, and Richard Zeckhauser

The chances of life-threatening events, such as terrorist attacks, hurricanes, floods, or earthquakes, are the lifeblood of risk analysis. Those in the business of measuring risks use a wide array of methods to quantify them. Perhaps the most common approach is statistical: risk is the relative frequency of a bad outcome in a well-defined set of outcomes. For instance, the annual risk of death in the United States due to terrorist attacks, floods, and earthquakes is 1 in 3.5 million (between 1970 and 2007), 1 in 8 million, and 1 in 9 million, respectively.

The Risk of Unique Events

In many cases, risks that matter are unique; they have no reference classes, perhaps because they have never occurred. What is the chance of war breaking out with North Korea or a cyberattack that leaves the United States completely defenseless? When reference classes do not exist, Bayesian methods allow decision-makers to express their beliefs about the chance of an outcome, given the available evidence (e.g., Gill 2015). Bayesian techniques allow people to report their beliefs about Trump being impeached, Greece exiting the Eurozone, or Iran complying with the 2015 nuclear agreement. Such beliefs provide the foundations of policy decisions, such as raising the federal

minimum wage, reducing domestic nuclear stockpiles, or brokering a trade deal with China.

The meaning of probability is a controversial topic. Many people are skeptical that probabilities can ever be assigned to unique events (e.g., Mill 1882; Keynes 1937). Others argue that all events are, in some way, unique; it depends on how one defines the reference class (Berry 1993). This chapter demonstrates how we can improve the accuracy of forecasts about unique geopolitical events using better forecasters, improved psychological interventions, enhanced statistical algorithms, and response scales with a greater number of uncertainty distinctions.

In the U.S. intelligence community, clear communication of risk and uncertainty is essential. Yet quite often, analysts eschew numerical estimates as expressions of their beliefs. They prefer to express their hunches with phrases, such as "liable to happen," "distinct possibility," or "hard to tell." Psychological research on verbal uncertainty phrases shows that such vague verbiage is easily misconstrued. For instance, Wallsten et al. (1986) asked participants to assign numerical values to their interpretations of probability phrases. The resulting numbers differed in meaning across individuals. The word "possible" had an interquartile range as large as 43% (Mosteller and Youtz 1990).

A particularly poignant example of how qualitative expressions can be misunderstood comes from a case study of President John F. Kennedy's decision to invade Cuba at the Bay of Pigs. In 1961, Kennedy asked the joint chiefs of staff to assess the plan's feasibility. The chiefs believed the chances of success were roughly 30%, but they conveyed their views verbally by saying, "This plan has a fair chance of success." The report's author, Brigadier General David Gray, later said, "We thought other people would think 'a fair chance' would mean 'not too good.'" However, President Kennedy allegedly interpreted this statement as something more likely. Gray believed that his imprecise language contributed to what was widely viewed as a serious strategic blunder (Wyden 1979, 88–90), a blunder that quantitative terminology might have avoided. Fifty years later, quantitative estimates are still the exception, not the norm.

Tournaments to the Rescue

In recent years, IARPA (the Intelligence Advanced Research Project Activity), the research wing of the U.S. intelligence community, has funded scientists to

discover how to better predict unique events. From 2011 to 2015, IARPA sponsored a program called ACE (Aggregative Contingent Estimation), comprising four massive geopolitical forecasting tournaments conducted over the span of four years. They supported five university teams to find optimal ways of eliciting beliefs from crowds and aggregating them.

Questions in the tournaments ranged from pandemics and global leadership change to international negotiations and economic shifts. For example, a question released on September 9, 2011, asked, "Who will be inaugurated as President of Russia in 2012?" Our team, The Good Judgment Project, studied over a million forecasts provided by thousands of volunteers who attached numerical probabilities to over 500 such events (Mellers et al. 2014; Tetlock et al. 2014).

In the ACE tournaments, IARPA carefully defined predictive success using a metric called the Brier scoring rule (the sum of squared deviations between forecasts and outcomes, where outcomes were 0 and 1 for the nonoccurrence and occurrence of events, respectively; Brier 1950). Consider the question, "Will Bashar al-Assad be ousted from Syria's presidency by the end of 2016?" For this question, outcomes were binary; Assad either stayed or left. Suppose a forecaster predicted that Assad had a 60% chance of staying and a 40% chance of being ousted. If, at the end of 2016, Assad remained in power, the participant's Brier score would be $[(1–0.60)^2 + (0–0.40)^2] = 0.16$. If Assad was ousted, the forecaster's score would be worse $[(0–0.60)^2 + (1–0.40)^2] = 0.36$. With Brier scores, lower values mean greater accuracy, and zero is a perfect score. The winning university group had the lowest Brier scores, averaged over individuals, days, and questions.

The Good Judgment Project won the ACE tournaments by a wide margin each year by being faster than the competition at finding ways to push probabilities toward 0 for things that did not happen and toward 1 for things that did happen. Five drivers of accuracy accounted for our team's success. They were identifying talent, training forecasters in probabilistic reasoning, putting forecasters in teams, placing the best forecasters in elite teams to work with each other, and aggregating forecasts using new algorithms (Mellers et al. 2014; Mellers et al. 2015a; Mellers et al. 2015b). We will briefly discuss each driver and then add another.

Identifying Talent

Our team investigated the psychological traits, cognitive abilities, and political knowledge of thousands of forecasters to understand who they were and what factors correlated with their performance (Mellers et al. 2015a). Better forecasters had more political knowledge and greater intelligence (both crystalized and fluid), as measured by the Raven's Advanced Progressive Matrices (Arthur et al. 1999; Balboni, Naglieri, and Cubelli 2010); the Shipley-2 abstraction test (Shipley, Gruber, Martin, and Klein 2009); the cognitive reflection test (Frederick 2005); an extended version of the cognitive reflection test (Baron et al. 2015); and questions from two numeracy scales (Lipkus, Samsa, and Rimer 2001; Peters et al. 2006).

Cognitive styles also correlated with performance. Better forecasters had a competitive streak, a greater appetite for intellectual challenges, and a stronger tendency to change their minds in response to new evidence (Mellers et al. 2015a). They scored high on a test of "actively open-minded thinking," which implied that they searched for and took into consideration information that ran counter to their prior beliefs (Haran, Ritov, and Mellers 2013). Finally, they had greater "need for cognition" (Cacioppo and Petty 1982; Cacioppo, Petty, and Kao 1984). They enjoyed analytic problems, complex puzzles, and intellectual challenges.

The most successful forecasters in the Good Judgment Project believed that forecasting ability was not an innate, God-given ability. Everyone knows the old joke about how to get to Carnegie Hall: practice, practice, practice. They viewed highly skilled performance as something they could do only after intense, focused, long-term commitment. Ericsson, Krampe, and Tesch-Romer (1993) argued that expert performance comes from deliberate practice or grit (Duckworth 2016). More successful forecasters were more engaged and showed greater effort and perseverance. Frequency of belief updating turned out to be the strongest single behavioral predictor of accuracy.

The very top performers—a group called "superforecasters"—had many of these characteristics and more (Mellers et al. 2015b; Tetlock and Gardner 2015). They were more inclined to embrace a secular, agnostic/atheistic worldview that treats everything as subject to deterministic laws of science. This worldview predisposed them to treat their beliefs more like testable probabilistic propositions than sacred possessions—and to be more cautious about

over interpreting coincidences by attributing them to supernatural mechanisms such as fate.

Training Forecasters in Probabilistic Reasoning

The Good Judgment Project developed a training module in probabilistic reasoning to help guide participants through the forecasting process. Psychologists have tried for decades to discover methods of improving probability judgments. Promising approaches include statistical training (Fong, Krantz, and Nisbett 1986), feedback (Benson and Önkal 1992), exposure to multiple perspectives (Ariely et al. 2000; Herzog and Hertwig 2009), exposure to historical analogies (Lovallo, Clarke, and Camerer 2012), decomposition of the problem into subsets (Fischhoff, Slovic, and Lichtenstein 1977), and explicit consideration of contradictory evidence (Koriat, Lichtenstein, and Fischhoff 1980).

The Good Judgment training module contained a variety of forecasting recommendations. It gave practical tips about where to find professional and amateur forecasts on the internet. It instructed forecasters to consider multiple reference classes before taking into account information that was specific to the event. It suggested that when forecasters had multiple estimates of the same event from polls, models, or expert opinions, they should average the estimates. Forecasters were also told to imagine possible futures, use decision trees, and avoid judgmental biases such as overconfidence and base-rate neglect (Kahneman, Slovic, and Tversky 1982). The module was interactive, with questions and answers that checked participants' understanding of the concepts.

Placing Forecasters in Teams

Numerous studies have shown that crowd predictions are frequently better than those of a single expert (e.g., Page 2007; Soll and Larrick 2009). But how should the crowd interact in order to generate the most accurate aggregate forecasts? Should they work independently without communicating? Or should they collaborate in teams that promote cooperation within the group and competition across groups? The Good Judgment Project used randomized

control conditions to test the effects of independent forecasts versus forecasts based on team interactions.

The case for working alone is statistical. Independent forecasts will often have uncorrelated errors, and, in the aggregate, they should cancel out (Surowiecki 2004). The case for working collaboratively is that groups can be more accurate than individuals when they are cohesive, engaged and share a mental model of the task (Levine and Moreland 1990; Kerr and Tindale 2004). Social interactions can inspire those who wish to perform well in the presence of others. Team members can share information, answer each other's questions, and encourage those who are less involved. The Good Judgment Project found that teams performed significantly better than individuals working alone (Mellers et al. 2014), and this result was replicated four years in a row.

Placing Top Forecasters in Elite Teams

A large literature on peer effects in the classroom suggests that students benefit from working in cohorts of similar ability levels (see Epple and Romano 2011, for a review). The Good Judgment Project reasoned that superforecasters (the top 2% of forecasters at the end of each year) might also enjoy an advantage if they worked with those of similar skill. But any beneficial effects of tracking would depend on the extent to which geopolitical forecasting was attributable to skill versus luck. If forecasting accuracy was mostly luck, superforecasters should regress to the mean after their initial success. If forecasting accuracy was primarily skill, superforecasters should continue their superior performance and possibly do even better in a richer intellectual environment. Defying expectations of regression toward the mean, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics, year after year (Mellers et al. 2015b). This intervention shows the astonishing potential of dedicated, talented forecasters as they tried to keep getting better.

Aggregation of Forecasts

The Good Judgment Project tried many aggregation rules, but the most successful was a relatively simple one, with a single estimated parameter. Pre-

dictions were combined using the geometric mean of the log odds (Baron et al. 2014; Satopää et al. 2014a; Satopää et al. 2014b). An empirically estimated exponent was applied to the mean. It is well known that individuals are frequently overconfident in their beliefs, but the mean of multiple forecasts may underestimate the total knowledge of the group. If the estimated exponent is greater than 1, the aggregate forecast is shifted toward the nearest end of the probability scale (0 or 1) as if the aggregate has more information than is reflected in the mean. Similarly, when the exponent is less than 1, the aggregate is shifted away from the nearest end of the scale, as if the aggregate has less information than the mean reflects. The exponent "recalibrates" the geometric mean. The algorithm also discounted older forecasts and differentially weighted individuals based on their previous accuracy and/or effort. This aggregation rule outperformed other methods each year for four years (Mellers et al. 2014). Almost a million aggregate forecasts from the Good Judgment Project were on the right side of maybe 86% of the time. This algorithm outperformed the simple mean of forecasts in a control condition by as much as 60%. In short, we discovered that the human forecasts of unique events could not only be predicted, but they could be predicted with a surprising degree of accuracy.

How Do Professional Intelligence Analysts Make Forecasts?

The 2004 Intelligence Reform and Terrorism Prevention Act requires that analysts "properly caveat and express uncertainties or confidence in analytic judgments." Yet there is no consensus on what it means to properly caveat. The intelligence community uses both qualitative and quantitative methods to express doubt. Common qualitative approaches include verbal terms (i.e., "we judge," "we estimate"), confidence levels (expressed as low, medium, or high), or uncertainty phrases (i.e., "unlikely," "possible," and "probable") (Friedman et al. 2017).

This form of expression may seem reasonable at first glance, but the meaning of such phrases is far from clear (Beyth-Marom 1982; Mosteller and Youtz 1990; Wallsten and Budescu 1995). Researchers have found between-subject and within-subject differences in the meaning of uncertainty phrases. When the same uncertainty phrases were associated with different events, the same subjects assigned different levels of probability. "A good chance" of being

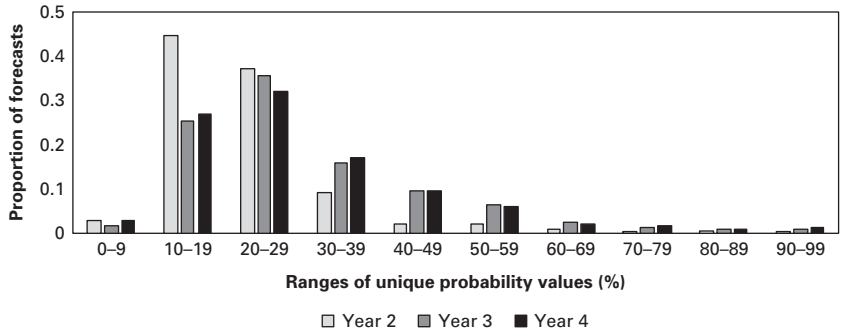

Figure 12.1. The proportion of unique probability values used by forecasters.

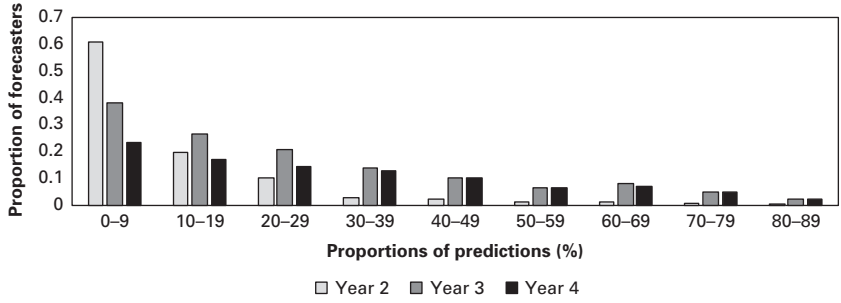

Figure 12.2. The proportion of forecasters using different percentages of more precise predictions (not multiples of 5% or 10%).

assaulted sounded greater to most people than "a good chance" of rain tomorrow. The meaning of probability phrases depended on the desirability of the event (Cohen 1986), the severity of the event (Weber and Hilton 1990), and the base rate of the event (Wallsten, Fillenbaum, and Cox 1986).

Intelligence analysts are currently asked to use a hybrid scale in which they express their beliefs using a numerical rating that is tied to a verbal phrase and a range of probabilities. In November 2015, the National Intelligence Council recommended a seven-point scale, with each phrase anchored to unequal-sized bins of probabilities (Office of the Director of National Intelligence 2015). Numbers were labeled as "Remote," "Very unlikely," "Unlikely," "Even chance," "Likely," "Very likely," and "Almost certainly."

Rating scales such as these represent a compromise between quantitative and qualitative measures. Such measures are typically defended on the basis

of what decision theorists call the "congruence principle," or the idea that probability assessors should use a level of precision that reflects their ability to justify and express their beliefs, but no more (Budescu and Wallsten 1987). In areas of high-risk decision-making, it goes without saying that one should not communicate uncertainty in a manner that suggests greater knowledge than one actually has. But is an uncertainty scale with seven categories enough for analysts to convey their beliefs? Is this the best measure for maximizing predictive accuracy?

Using data from the Good Judgment Project, we could answer these questions by tackling three smaller ones (Friedman et al. 2017). How many distinctions along the probability scale do forecasters typically make? How many categories of uncertainty do forecasters actually *need*? And what predicts the tendency to make more granular probability judgments? Answers will tell us whether current methods for expressing uncertainty used in the intelligence community are sufficiently precise.

Distinctions That Forecasters Make

There are several ways to investigate categories of uncertainty that people are able to use. One way is to count the number of unique probability values forecasters use during the course of a given tournament year. Figure 12.1 shows such data from the last three years of the Good Judgment Project. The first year was not included because additional analyses, discussed later, required superforecasters, and they were not identified until the end of the first year. The proportion of forecasters who used different numbers of unique probabilities (shown as bins of 0 to 9, 10 to 19, 20 to 29, etc.) appear for years 2, 3, and 4. Most forecasters used 10 to 29 unique values of probability for the questions they faced during the year. Average numbers of unique values were 22, 30 and 30, in years 2, 3, and 4 respectively. If forecasters expressed their uncertainties in multiples of 10%, they would be making 11 distinctions. If distinctions were in all multiples of 5%, they would be making 21 distinctions. Values of 22, 30, and 30 suggest that, in year 2, people were using multiples of 5% on average, and in later years, they used more distinctions.

A second way to ask how many levels of uncertainty people can distinguish is to examine which percentages they use. To what extent did forecasters submit beliefs such as "19%" and "53%" versus "20%" and "50%"? Friedman

et al. (2017) examined instances in which forecasters made predictions that were *not* multiples of 5% or 10%. Results are shown in Figure 12.2, with proportions of forecasts that were not multiples of 5% and 10% (in binned intervals). In the second year, more than half of the forecasters made relatively few such predictions (e.g., 23% or 88%, 9% or fewer). As the years went on, forecasters made a greater number of more precise predictions. Both figures show that there are widespread individual differences in the distinctions of uncertainty that forecasters make across the entire probability continuum.

Distinctions That Forecasters Need

On the surface, one might conclude that more granular probabilities convey more information. Yet, in practice, it is unclear whether more precise responses have additional information or whether they simply reflect a desire on the forecaster's part to *appear* more precise. If differences are purely superficial, one might be safe in assuming that people don't need to express as many distinctions as they are actually making. On the other hand, if more precise forecasts are actually more accurate forecasts, then the intelligence community should allow analysts to use more categories to express their beliefs.

To find out whether more precise forecasts were associated with greater accuracy, Mellers et al. (2015b) rounded the predictions of superforecasters to the nearest 0.10, and 0.33 (corresponding to probability scales with 11 distinctions and 4 distinctions, respectively). In both cases, the rounding of superforecasters' predictions significantly decreased accuracy. That is, Brier scores computed on the rounded predictions were less accurate than original Brier scores. Although these analyses don't speak to cognitive processes directly, they are consistent with the hypothesis that superforecasters were capable of reliably making *at least* 12 distinctions on the probability continuum.

Friedman et al. (2017) used a similar, but more extensive method to estimate the number of categories forecasters could reliably use. They drew on more than 750,000 predictions from over 1,700 forecasters, each of whom had made predictions on 25 or more questions in the last three years of the tournament. Each forecaster's predictions were rounded to the midpoint of equal-sized bins (b), where b ranged from 2 to 101 categories. Friedman et al.

Table 12.1. Rounding forecasts to seven categories: Original Brier scores for different groups and percentage of errors added with rounding

Groups of forecasters	*Original Brier scores*	*Seven categories, unequal intervals*	*Seven categories, equal intervals*
Untrained individuals	Mean: 0.1890	0.5%**	0.5%
	Median: 0.162	0.6%**	0.2%**
Trained teams	Mean: 0.1360	0.8%**	3.3%*
	Median: 0.1000	0.9%**	2.4%**
Superforecasters	Mean: 0.093	6.1%**	10.4%**
	Median: 0.032	1.7%**	10.2%**

Note: * $p < 0.05$; ** $p < 0.001$.

recomputed Brier scores after each rounding and compared them to the original Brier scores. Then they looked at the change in the Brier score after rounding.

Table 12.1 shows mean and median percentages of accuracy associated with three groups from the Good Judgment Project (untrained individual forecasters, trained team forecasters, and superforecasters), as well as abbreviated statistical results concerning the impact of rounding of forecasts. After groups, the table presents unrounded Brier scores for each group. Superforecasters had the lowest Brier scores. Columns to the right show the percentage changes in Brier scores due to rounding. Both columns represent uncertainty scales with seven categories, reflecting the current scale used by the intelligence community. We show two different ways of operationalizing the scales. In the first case, the seven categories were associated with equal-sized intervals, and in the second case, they were unequal intervals. Rounding of forecasts to each number of categories was done by using the midpoints of the interval. Means of rounded and original scores were compared with two-sided t tests, and medians were compared with two-sided Wilcoxon signed-rank tests.

Positive percentages shown in the last two columns in Table 12.1 indicate the percentage increase in Brier scores after rounding (less accuracy). Although these changes are not large for untrained individual forecasters or trained team forecasters, differences are statistically significant when compared to the original Brier scores. Decreases in predictive accuracy are much greater for superforecasters than any other groups. Superforecasters tended

to make more granular predictions, and as Friedman et al.'s results demonstrate, their precision often conveyed valuable information.

Table 12.1 tells us that the category rating scales for expressing uncertainty used in the intelligence community do not give forecasters—especially the very best forecasters—enough latitude to convey all of the information they actually have. Forecasters using those scales cannot make the precise predictions that would have maximized their accuracy. Accuracy suffered.

Who Makes Finer Distinctions?

Friedman et al. (2017) developed an index of the granularity, or implicit precision, of each individual's forecasts. They calculated the number of bins, *b*, for which rounded Brier scores were not statistically different from unrounded Brier scores. The minimum number of categories was interpreted as an indirect measure of the fewest distinctions an individual was reliably capable of making. To explore the correlates of this granularity index, Friedman et al. conducted exploratory regressions using predictor variables such as forecasting accuracy, motivation, training, education, cognitive abilities, and cognitive styles.

Forecasting accuracy was the strongest predictor of the estimate of forecasters' precision. Those who made more distinctions along the probability continuum also tended to be more accurate. These individuals also tended to have training in probabilistic reasoning and were more engaged in the tournament (i.e., updated their forecasts more frequently and attempted to address more forecasting questions). Those whose forecasts were more precise also tended to have more experience by participating in the tournaments for a longer period of time.

These results imply that intelligence analysts and other professional forecasters can increase their accuracy by learning to be more precise. Analysts are full-time professionals whose job it is to assess uncertainty on a daily basis over many years. They have more opportunities and incentives to refine and revise their forecasts in light of new information than did forecasters in the Good Judgment Project, who were largely participating for fun.

Variables that did not predict the granularity index, perhaps surprisingly, included education, numeracy, cognitive ability, and cognitive styles (Friedman et al. 2107). These factors represent more innate variables that are harder to change or manipulate, whereas incentives for effort, engagement, and

training in probabilistic reasoning are interventions that organizations could make without huge investments in time and money. In sum, the ability to make more precise predictions appears to be something that forecasters can learn to do with the proper guidance and incentives.

Improving Current Practices

Our results show that standard methods of expressing uncertainty with seven-point rating scales as done in the intelligence community are simply too coarse. As discussed in Friedman et al. (2017), this finding did not depend on the use of extreme probability estimates, questions with shorter time horizons, questions of different types (e.g., military, economic, health-related), or different categories of strictly proper scoring rules. Additionally—and perhaps most importantly—it was the superforecasters who took the greatest hit to predictive accuracy when their response scales were constrained. Given their remarkable accuracy, we suggest that, if anything, superforecasters should have the loudest voices when events are uncertain, information is ambiguous, the stakes are high, or the consequences dire.

Some scholars and practitioners oppose the use of numerical probabilities on grounds that the extra "precision" is essentially noise (e.g., Fingar 2011). The National Intelligence Council (Office of the Director of National Intelligence 2007) also says that "assigning precise numerical ratings to [probabilistic] judgments would imply more rigor than we intend." Others say that numerical probabilities would impose additional mental costs on analysts. Although there may indeed be a learning period, this hypothesis requires empirical testing.

Our message here is simple. We know much more than we did a decade ago about how to accurately estimate the chances of unique events. Some methods are demonstrably better than others. The Good Judgment Project found five ways of improving accuracy. With these drivers, intelligent lay people could make forecasts that were 30% more accurate than those of professional intelligence analysts, even in instances where analysts had access to additional, classified information (Goldstein et al. in press).

Better forecasts require identifying, training, teaming, and tracking forecasters, and optimally aggregating their forecasts. First, getting the right people is essential. Better forecasters tend to take an analytical approach to predictions and to enjoy intellectual challenges. They search for evidence that

runs counter to their favored beliefs and maintain open minds. Second, training helps. People can learn to be better forecasters when they are instructed to use best practices for probabilistic reasoning. Third, geopolitical forecasters are more accurate when working in teams than when working independently. The shared information, encouragement and comradery introduced by the team structure in ACE demonstrably outweighed the potential for herding or groupthink. Fourth, accuracy gets an enormous boost when top performers are allowed to work together in elite teams. The added commitment and desire to not disappoint one's teammates proves to be a stronger incentive than we could have imagined. Finally, simple algorithms that incorporate the discounting of old forecasts, the differential skills of forecasters, and the degree of information overlap among the crowd are far superior to simple averages. These factors show that it is the combination of statistical and psychological insights that improve the predictive accuracy of unique events.

Data from the Good Judgment Project allowed us to test another driver of accuracy—the degree of precision intelligence analysts require in their expressions of uncertainty in order to maximize accuracy. Standard methods are seven-point category rating scales of uncertainty. A comparison of forecasters' original accuracy scores to accuracy scores after rounding to seven categories showed that inaccuracy grows if forecasters are constrained to express their beliefs using only seven categories. Even worse, it is top performers whose accuracy suffers the most when forced to communicate with restricted probability scales. This driver is, by far, the easiest one for the intelligence community to implement.

Unfortunately, it is still the norm that intelligence analysts express uncertainties with vague verbiage. What does it mean when a pundit asserts that a military operation is "likely" to succeed? Is the probability just above 55% or is it closer to 90%? Or, what if an expert says that a crisis is "unlikely" to escalate? Does that mean the probability is 10% or 40%? Despite the uncertainty and subjectivity inherent to policy debates, evidence from the Good Judgment Project suggests that there are valid grounds for asking analysts to assess uncertainty numerically using the entire probability scale; greater accuracy will be the result.

Although findings in one domain may not carry over into others, it is worth considering the notion that other professions might be systematically sacrificing predictive accuracy by using qualitative expressions of probability. Qualitative expression of uncertainty is commonly used in regulatory

policy (Sunstein 2104), medicine (Nakao and Axelrod 1983), and climate science (Budescu et al. 2014). These are all areas where risk and uncertainty play a crucial role in decision-making. Our research provides a methodological template for addressing this question in a principled way, with the first and foremost ingredient being to keep score.

Any organization that strives for the clearest communication of risks should extract as many useful signals from its people and its environment as possible. The Good Judgment Project has tested a variety of methods scientifically and found ways to bolster accuracy. The world is a messy place, and accurate predictions are unquestionably hard. But when the stakes are high—with billions of dollars or thousands of lives on the line—even small increases in predictive accuracy can translate into enormous benefits to society.

References

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., . . . & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*(2), 130.

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (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.

Balboni, G., Naglieri, J. A., & Cubelli, R. (2010). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment, 28*(3), 222–235.

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis, 11*(2), 133–145.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting, 8*(4), 559–573.

Berry, D. A. (1993). A case for Bayesianism in clinical trials. *Statistics in Medicine, 12*(15–16), 1377–1393.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1*(3), 257–269.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.

Budescu, D. V., Por, H. H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change, 4*(6), 508.

Budescu, D. V., & Wallsten, T. (1987). Subjective estimation based on precise and vague uncertainties. In G. Wright and P. Ayton (Eds.), *Judgmental forecasting.* New York: Wiley.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116.

Cacioppo, J. T., Petty, R. E., & Kao, C.-F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306–307.

Cohen, B. L. (1986). *The effect of outcome desirability on comparisons of linguistic and numerical probabilities* (unpublished MA thesis). University of North Carolina at Chapel Hill.

Daniels, R., Ketti, D., and H. Kunreuther (Eds). 2006. *on risk and disaster: Lessons learned from Hurricane Katrina*. Philadelphia: University of Pennsylvania Press.

Duckworth, A. (2016). *Grit: The power of passion and perseverance*. New York: Simon & Schuster.

Epple, D., & Romano, R. E. (2011). Peer effects in education: A survey of the theory and evidence. In *Handbook of social economics* (Vol. 1b, pp. 1053–1163). Amsterdam: North-Holland.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406.

Fingar, T. (2011). *Reducing uncertainty: Intelligence analysis and national security.* Stanford, CA: Stanford University Press.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 552–564.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18*(3), 253–292.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2017). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*, 410–422.

Gill, J. (2015). *Bayesian methods: A social and behavioral sciences approach* (3rd ed.). Boca Raton, FL: CRC Press.

Goldstein, S., Hartman, R., Cornstock, E., & Baumgarten, T.S. (in press). Assessing the accuracy of geopolitical forecasts from the US intelligence community's prediction market. *Journal of Forecasting.*

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*(3), 188.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*(2), 231–237.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology, 55,* 623–655.

Keynes, J. M. (1937). The general theory of employment. *Quarterly Journal of Economics, 51*(2), 209–223.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118.

Levine, J. M., & Moreland, R. L. (1990). Progress in small group research. *Annual Review of Psychology, 41*(1), 585–634.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37–44.

Lovallo, D., Clarke, C., & Camerer, C. (2012). Robust analogizing and the outside view: Two empirical tests of case-based decision making. *Strategic Management Journal, 33*(5), 496–512.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., . . . & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . & Ungar, L. (2015b). Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science, 10*(3), 267–281.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . & Murray, T. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25*(5), 1106–1115.

Mill, J. S. (1882). *A system of logic* (8th ed.). New York: Harper.

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*(1), 2–12.

Nakao, M. A., & Axelrod, S. (1983). Numbers are better than words: Verbal specifications of frequency have no place in medicine. *American Journal of Medicine, 74*(6), 1061–1065.

Office of the Director of National Intelligence. (2007). Intelligence community directive 203: Analytic standards.

Office of the Director of National Intelligence. (2015). Intelligence community directive 203: Analytic standards.

Page, S. (2007). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407–413.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014a). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30*(2), 344–356.

Satopää, V. A., Jensen, S. T., Pemantle, R., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014b). Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs. *Annals of Applied Statistics, 8*, 1256–1280.

Shipley, W., Gruber, C., Martin, T., & Klein, M. (2009). *Shipley-2 manual*. Torrance, CA: Western Psychological Services.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 780.

Sunstein, C. R. (2014). *Valuing life: Humanizing the regulatory state*. Chicago: University of Chicago Press.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York: Doubleday.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. New York: Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science, 23*(4), 290–295.

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review, 10*(1), 43–62.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*, 348–365.

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language, 25*(5), 571–587.

Weber, E., & Hilton, D. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781–789.

Wyden, P. (1979). *Bay of Pigs: The untold story.* New York: Simon & Schuster.