ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2015, Vol. 10(3) 267–281
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691615577794
pps.sagepub.com
SAGE

Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions

Barbara Mellers[1], Eric Stone[1], Terry Murray[2], Angela Minster[3], Nick Rohrbaugh[1], Michael Bishop[1], Eva Chen[1], Joshua Baker[1], Yuan Hou[1], Michael Horowitz[1], Lyle Ungar[1], and Philip Tetlock[1]

[1]Department of Psychology, University of Pennsylvania; [2]Haas School of Business, University of California, Berkeley; and [3]Department of Statistics, Temple University

Abstract

Across a wide range of tasks, research has shown that people make poor probabilistic predictions of future events. Recently, the U.S. Intelligence Community sponsored a series of forecasting tournaments designed to explore the best strategies for generating accurate subjective probability estimates of geopolitical events. In this article, we describe the winning strategy: culling off top performers each year and assigning them into elite teams of *superforecasters*. Defying expectations of regression toward the mean 2 years in a row, superforecasters maintained high accuracy across hundreds of questions and a wide array of topics. We find support for four mutually reinforcing explanations of superforecaster performance: (a) cognitive abilities and styles, (b) task-specific skills, (c) motivation and commitment, and (d) enriched environments. These findings suggest that superforecasters are partly discovered and partly created—and that the high-performance incentives of tournaments highlight aspects of human judgment that would not come to light in laboratory paradigms focused on typical performance.

Keywords

predictions, forecasts, expertise, teams, probability training

Accurate probability estimates are critical for good decision making in fields as diverse as medical diagnosis, portfolio management, and intelligence analysis (Baron, 2000). People depend on meteorologists to provide accurate predictions of the weather, central banks to generate accurate forecasts of economic trends, and intelligence agencies to anticipate threats to national security. However, the psychological literature, in which the accuracy of subjective probability judgments has been examined, raises serious doubts about people's competence as intuitive forecasters. People often misuse Bayes's theorem when updating their beliefs (Birnbaum & Mellers, 1983; Kahneman & Tversky, 1985). They test hypotheses in suboptimal ways (Ofir, 1988; Wason, 1968). They are susceptible to the hindsight bias or the "I-knew-it-all-along" effect (Christensen-Szalanski & Wilhelm, 1991). Furthermore, people are often overconfident about what they know, consistent with self-serving biases and positive illusions (Baron, 2000; Gilovich, Griffin, & Kahneman, 2002; Lichtenstein, Fischhoff, & Phillips, 1982). Entrepreneurs believe their personal chances of creating a successful business are much higher than statistics suggest (Cooper, Woo, & Dunkelberg, 1988), and automobile drivers believe they are safer and more skilled behind the wheel than their peers (Svenson, 1981). Even experts, such as skilled basketball players, are overly optimistic about their chances of future successes (Jagacinski, Isaac, & Burke, 1977).

The focus of this research literature, however, has been on *typical performance*—on how relatively unmotivated people perform on relatively unfamiliar laboratory

Corresponding Author:
Barbara Mellers, Department of Psychology, University of Pennsylvania, Solomon Labs, 3720 Walnut St., Philadelphia, PA 19104
E-mail: mellers@wharton.upenn.edu

tasks—and not on *optimal performance*—on how well people could perform. Researchers also have rarely given participants training on how to handle the tasks or opportunities to learn from performance. In this article, we reverse this emphasis and focus on optimal performance: How well do people perform as intuitive forecasters when we select exceptional talent, offer debiasing training, reward top performers, and let the forecasters wrestle with real-world problems that they find intrinsically interesting (not laboratory problems that have been chosen, a priori, to make one or another theoretical point about human rationality)? The opportunity to answer these questions arose in three forecasting tournaments sponsored by the Intelligence Advanced Research Projects Activity (IARPA) between 2011 and 2014.

Forecasting Tournaments

Although governments invest massive sums to predict the actions of political actors and the consequences of those actions, they invest astonishingly little to evaluate the effectiveness of their favorite methods of generating predictions. IARPA attempted to address these issues by sponsoring a series of geopolitical forecasting tournaments designed to test the best strategies of making intuitive probability judgments. Individuals from five university-based research programs competed to develop innovative methods of assigning probability estimates to high-impact events around the globe. Illustrative questions included the following: Who will be the president of Russia in 2012? Will North Korea detonate another nuclear weapon in the next 3 months? How many refugees will flee Syria next year? How fast will China's economy grow next quarter? Questions were not chosen by researchers but by IARPA, and they were ecologically representative of the challenges of intelligence analysis, with one key exception—each question had to pass the clairvoyance test (written clearly enough as to leave no room for ex post facto quarrels over what really happened—and who was right).

Every day, individuals from the competing research programs submitted aggregate forecasts for a large set of geopolitical events to IARPA. To generate aggregates, facilitators from each program were responsible for recruiting participants, devising methods to elicit uncertainty, and creating algorithms for combining opinions (Satopää, Baron, et al., 2014; Satopää, Jensen, Mellers, Tetlock, & Ungar, 2014). To ensure a level playing field, we scored all research groups on the official accuracy metric—the well-known Brier scoring rule, described later (Brier, 1950).

Each tournament spanned approximately 9 months of the year. Participants from around the world used websites developed by each program to submit probability estimates of the likelihood of geopolitical events. Slightly more than 100 questions were posed in the first two tournaments, and there were approximately 150 questions in the third tournament. Questions were released in small bundles over the course of the tournament and remained open for an average of 102 days (ranging from 3 to 418 days).

The Good Judgment Project (GJP)

In this article, we describe the best-performing strategy of the winning research program: the GJP. To preempt confusion, we should clarify at the outset the limits on the conclusions we can draw. First, the mere fact that a given strategy "won" a tournament does not mean that it was optimal or even close to optimal. Second, each project had to rely on considerable guesswork in selecting a set of strategies—any one component or combination of components of which could be the cause of competitive success or failure. Different research literatures—ranging from cognitive debiasing to group dynamics to prediction markets and statistical aggregation techniques—offered contradictory guidance on how to proceed.

Prior to the forecasting tournaments, we had no basis for assuming that people were even capable of making consistent and reliable geopolitical forecasts. Past research has demonstrated that intuitive predictions (probabilistic and nonprobabilistic) are worse than statistical predictions and, sometimes, worse than chance (Dawes, Faust, & Meehl, 1989). For these reasons, the research project we describe here is best viewed as a mixed hypothesis-generation-and-testing exercise in a high-stakes setting that was not intended to favor one side or the other in the great rationality debate (Tetlock & Mellers, 2002).

We recruited forecasters from professional societies, research centers, alumni associations, science blogs, and word of mouth. Each year, we began with a large number of forecasters (ranging from 2,200 to 3,900), and attrition ranged from 3% to 7%. Participation required a bachelor's degree or higher and completion of a battery of psychological and political knowledge tests that took about 2 hr. Participants tended to be men (83%) and U.S. citizens (74%), with an average age of 40 years. Almost two thirds (64%) had some postgraduate training.

An illustrative forecasting question was as follows: "Will Italy's Silvio Berlusconi resign, lose reelection/confidence vote, or otherwise vacate office before 1 January 2012?" Forecasters predicted the chance that the event would occur (in which 0% = *certain it will not occur*, and 100% = *certain it will occur*), and they were encouraged to update their beliefs as often as they wished before the close of each question.

Forecasters received status rewards according to their accuracy or Brier score during the tournament. Brier scores are used to assess the accuracy of probabilistic forecasts and to encourage the reporting of true beliefs (no blurring probability and value judgments by ignoring the dangers of under- or overpredicting an outcome). A Brier score is the sum of squared deviations between forecasts and reality (in which reality is coded as 1 for the event and 0 otherwise), ranging from 0 (*best*) to 2 (*worst*). Suppose a question has two possible outcomes, and a forecaster predicted a probability of 0.75 for the outcome that did occur and 0.25 for the one that did not. The Brier score would be $(1 - 0.75)^2 + (0 - 0.25)^2 = 0.125$.

Leaderboards were displayed that showed the top 10% of individual Brier scores in each condition, member Brier scores for those on a given team, and team Brier scores across all teams. The Brier score for any question in a team was the median of individual scores. The median score reduced the negative effects that any particular individual could have had with a mean score. In addition to status rewards, forecasters were paid for participation but not for accuracy. Those who met the minimum requirement of making at least 25 forecasts received a $150 Amazon gift certificate in the first tournament and a $250 certificate in the second and third tournaments. Furthermore, those who returned to participate in Years 2 or 3 were given an additional $100 gift certificate.

We knew we needed to conduct randomized control experiments to determine how best to design effective training and to create supportive work environments (i.e., collaborative vs. independent forecasts). That way, we could learn what worked "best." Experiments were conducted with both surveys and prediction markets and are discussed in detail elsewhere (Atanasov et al., 2014; Mellers, Ungar, et al., 2014).

Superforecasters

In this article, we focus solely on the superforecaster phenomenon. Superforecaster recruitment began at the end of Year 1, when a new strategy was added to the GJP's portfolio of performance boosters: tracking, that is, stratifying forecasters into groups on the basis of performance. All forecasters were ranked on the basis of performance. From this list, we selected 60 forecasters (the top 5 forecasters from each of 12 experimental conditions) who were assigned randomly to 5 elite teams of 12 members each and were given the title of superforecasters. All superforecasters knew they were working with other superforecasters.

Note that the working conditions for superforecasters were informed by our previously reported experiments (Mellers, Ungar, et al., 2014). Forecasters in Year 1 performed better when they received cognitive-debiasing training and when they worked in collaborative teams, not individually. Accordingly, we decided that superforecasters would also work in teams that were trained in cognitive debiasing.

The outcome of this tracking strategy for the following year was nonobvious. One could imagine the full spectrum of predictions. On the pessimistic side, one might say that superforecasters would regress toward the mean of regular teams because chance is the dominant driver of Year 1 performance (in the spirit of Hartzmark, 1991). The labeling of "super" would make them overconfident or arrogant (in the spirit of Levitt & March, 1988). On the optimistic side, one might think that superforecasters would get a boost from the stimulation of working with other high performers (in the spirit of Betts & Shkolnik, 2000). They would benefit from self-fulfilling-prophecies (Rosenthal, 1966), and they would be more intelligent; intelligence is the best all-purpose predictor of job performance (in the spirit of Hunter & Schmidt, 1996).

If we had derived our predictions by averaging these clashing views, we would have predicted a rather weak effect somewhere in the middle. Furthermore, we would have been wrong. In Year 2, we learned that funneling top performers into elite teams boosted accuracy far more than either of our two experimental manipulations—cognitive debiasing and collaborative teaming—administered alone or together. Superforecaster performance exceeded all other groups by a wide margin. Forecasting appeared to be somewhat skill-based, and the acquisition of that skill was accelerated when the best performers worked with each other in elite teams (Mellers, Ungar, et al., 2014). Data from Year 2 were unequivocal.

Does the Superforecaster Edge Hold Up Against Stronger Tests of Accuracy?

Although the Year 2 questions were highly heterogeneous (ranging from finance and commodity markets to naval strategy to electoral politics to pandemics), it is still possible that superforecasters were just lucky. The best way to find out was to raise the bar by tracking forecasters again in Year 3 and by assessing superforecaster performance relative to comparison groups on a wide array of accuracy measures.

To that end, we used the same approach at the end of Year 2 as we had at the end of Year 1. We selected 60 new superforecasters solely on the basis of performance accuracy. If returning forecasters from Year 1 and newly anointed superforecasters from Year 2 continued their outstanding performance through Year 3, the scales of plausibility would tip further against the just-luck hypothesis.

A key element of tracking is that the best performers interact in elite teams with other top performers. To better understand the effects of superforecaster performance in *elite versus regular teams*, we created a comparison

Table 1. Measures of Performance (Average of Years 2 and 3), Motivation and Commitment, and Enriched Environments

Performance measure	Superforecasters	Top-team individuals	All others
Standard Brier scores	−0.34	**−0.14**	**0.04**
Resolution	0.40	**0.35**	**0.32**
Average calibration	0.01	**0.03**	**0.04**
AUC (0%–100%)	96.00	**84.00**	**75.00**
Learning rates	−0.26	**−0.18**	**0.00**

Note: Bold values of top-team individuals and all others are significantly different from superforecasters at the .01 level. Brier scores are the average sum of squared deviations between outcomes and probability predictions over days and questions (lower scores are better). Resolution measures the ability to distinguish signals from noises. Calibration measures the difference between average forecasts and proportion of correct responses. Area under the curve (AUC) reflects the probability of detecting a true positive response over a false positive response. Learning is represented as the slope of the Brier scores over the course of each year.

group of the next most accurate forecasters (after superforecasters) serving on regular teams. These forecasters also were high performers who fell just short of the superforecaster cutoff in Years 1 and 2 and who worked in teams with others who were also high performers. This superforecaster comparison group of high performers in regular teams is called *top-team individuals*. Although the comparison between superforecasters and top-team individuals is not a pure measure of the effect of elite versus regular teams, both groups were among the best performers from a large pool of already talented individual forecasters.

We also created a second comparison group consisting of the other 1,498 forecasters. This group, called *all others*, provided a broader baseline for assessing superforecasters. To increase the reliability of performance estimates, we only included forecasters in both comparison groups if they had made forecasts for at least 25 questions in each tournament.

Tests with Brier scores

In the tournament, forecasters were allowed to select their own questions. The element of self-selection raises the possibility that superforecasters were skilled at predicting unpredictability (a nontrivial achievement in itself) and simply cherry-picked easier-to-predict questions. To minimize the effects of question difficulty and also to highlight the relative rather than the absolute performance of forecasters, we standardized Brier scores within each question and averaged across questions. Table 1 shows five measures of performance for superforecasters, top-team individuals, and all others. Bold values for the comparison group indicate that the comparison group had significantly worse performance than superforecasters. Brier scores for superforecasters were significantly more accurate than those of both top-team individuals and all others.

In Figure 1, we provide visual displays of standardized Brier scores for the three groups over the three tournaments.[1] Superforecasters (blue bar) selected at the end of the Year 1 tournament were slightly more accurate than top-team individuals (red bar)—a result predetermined by our selection criteria. All others (green bar) were much worse. If superforecasters had performed worse and regressed to the mean in Years 2 and 3, blue bars would have risen (worse performance). However, superforecasters performed even better. Blue bars for Years 2 and 3 show even greater accuracy. By contrast, top-team individuals and all others showed regression effects (worse performance), as seen by the higher red and green bars. In short, the superforecaster intervention from Year 2 was cleanly replicated in Year 3, and the solidity of the replication convinced us that tracking was extremely unlikely to be a lucky accident.

Another test of superforecasters is to ask whether they continue to win under tough conditions with little time and information. We examined forecasts that (a) were made on the first day a question was released (when clues were least plentiful) and (b) were submitted during a 4-min period between the first time the participant sees the forecasting question and the submission of the forecast. During this period, forecasters read the question and the resolution criteria, gathered and synthesized information, and made a prediction.

Figure 2 shows average Brier scores over Years 2 and 3 as box plots. Even with relatively little effort and deliberation, superforecasters outperformed top-team individuals and all others. Tight restrictions on time and information did not erode the superforecaster advantage.

Tests with resolution and calibration

Brier scores can be decomposed into three variance components: variability, resolution, and calibration (Murphy & Winkler, 1987). Variability is independent of skill and

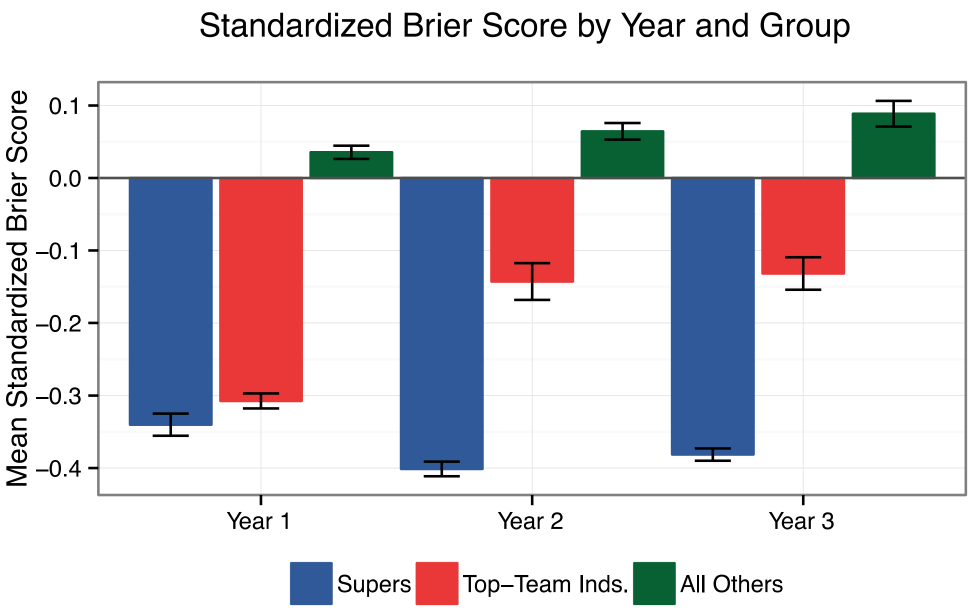

Fig. 1. Mean standardized Brier scores for superforecasters (Supers, who were the top performers) and the two comparison groups. Top-Team Inds. = top-team individuals who were the next best forecasters and who served in regular teams; All Others = remaining forecasters who served in both years. Error bars are plus and minus one standard error, and lower scores indicate greater accuracy.

simply reflects the base rate for events in the environment. It is harder to predict the weather in St. Louis than in San Diego because of the greater variability in St. Louis weather. However, this distinction has nothing to do with skill, so variability is not discussed further.

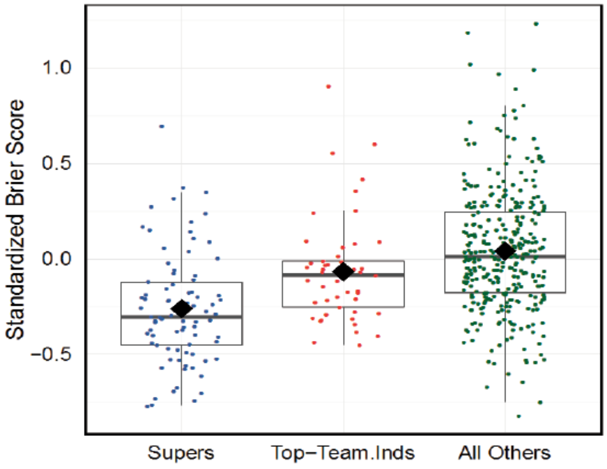

Fig. 2. Mean standardized Brier scores for superforecasters (Supers) and the two comparison groups in Years 2 and 3 for forecasts made within the first 24 hr of the release of a question and within 4 min of logging onto the computer. Even when forecasts were made quickly without research, Supers were more accurate. Diamonds are means, and the horizontal lines associated with each box are the 25th, 50th, and 75th percentiles of the distributions. Vertical lines extend to the most extreme points, excluding outliers. Top-Team Inds. = top-team individuals.

Resolution refers to appropriate decisiveness or the ability to accurately distinguish signals from noise. Superforecasters had greater resolution than both top team individuals and all others (see Table 1). *Calibration* refers to appropriate humility or the ability to make subjective forecasts that, in the long run, coincide with the objective base rates of events. People are said to be "well-calibrated" overall if their average confidence equals their percentage correct; that is, they are neither overconfident nor underconfident. Superforecasters were significantly better calibrated than the other two groups (see Table 1).

Tests with area under the curve (AUC)

Another measure of accuracy—AUC or discrimination ability—comes from signal detection theory (Swets, Dawes, & Monahan, 2000). AUC has the advantage of being independent of an individual's response threshold or tendency to call a signal "Signal" and to call a noise "Noise." AUCs are easy to evaluate visually by constructing a receiver operating characteristic (ROC) curve that plots the probability of a hit (true positive) against the probability of a false alarm (false positive). Curves show constant levels of ability across different response tendencies. Imagine the curve is a line in a box with area 1.0. AUC is the proportion of area in the box that falls

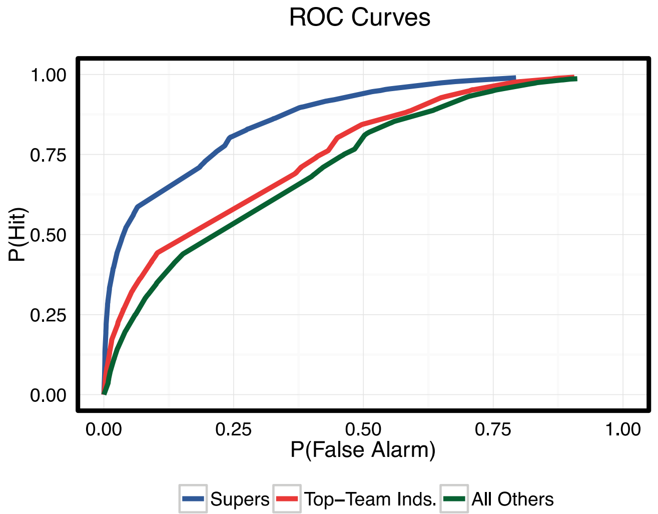

Fig. 3. Curves show the probability of a hit (true positives) plotted against the probability of a false alarm (false positives). Curves that lie closer to the upper left corner indicate better performance. ROC = receiver operating characteristic; Supers = superforecasters; Top-Team Inds. = top-team individuals.

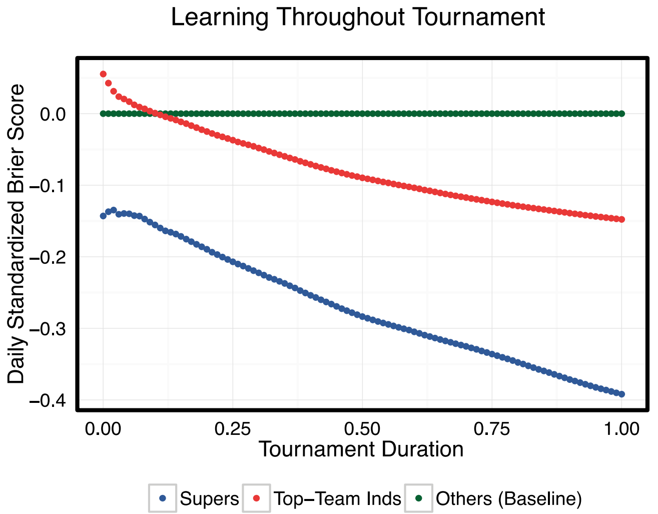

Fig. 4. Daily standardized Brier scores for superforecasters (Supers) and top-team individuals (Top-Team Inds.) relative to all others (flat green curve) averaged over Years 2 and 3 of the tournament. Steeper negative slopes indicate faster learning. Solid lines are linear regressions.

beneath the curve. An area of .50 (or 50%) would be equivalent to guessing. It is rare for decision makers or algorithms to exceed .84 (or 84%).

In Figure 3, we present AUCs using binary questions for superforecasters and the two comparison groups. Curves closer to the upper left-hand corner of the graph (greater AUC) indicate better performance. Computed values of AUC were 96%, 84%, and 75% for superforecasters, top-team individuals, and all others, respectively (see Table 1). Superforecaster superiority held up strongly on this metric as well. Indeed, in one analysis, the ROC for superforecasters was as accurate 300-plus days into the future, when the ROC for regular forecasters was only 60 days out.

Tests of learning

Our final measure of skill is the reduction of forecasting errors within a tournament year. In Figure 4, we plot mean daily standardized Brier scores averaged over Years 2 and 3 for superforecasters (blue line) and top-team individuals (red line) relative to all others (flat green horizontal line) as a function of time within the tournament season. Steeper slopes reflect faster learning rates. Superforecasters were more accurate at the outset—and they improved faster as new information became available (see Table 1).

Thus, the evidence of superforecasters' superiority is robust: They had better overall Brier scores (see Figure 1) and better "quick-response" Brier scores (see Figure 2) in Years 2 and 3; they had better resolution and calibration scores (see Table 1); they scored better on the AUC metric of skill at distinguishing signals from noise (see Figure 3); and they were the fastest learners within the tournaments (see Figure 4).

What Makes Superforecasters So Super?

Next, we examine the evidence supporting four mutually reinforcing drivers of superforecaster performance: (a) cognitive abilities and styles, (b) task-specific skills, (c) motivation and commitment, and (d) enriched environments.

Cognitive abilities and styles

Overall, the Year 1 forecasters were above average in fluid intelligence relative to general-population samples (Mellers, Stone, et al., 2014), and superforecasters were the most accurate forecasters within that large group. In another study, we showed that fluid intelligence was the strongest dispositional predictor of accuracy (Mellers, Stone, et al., 2014). Therefore, an obvious notion is that superforecasters were simply more intelligent than the non-superforecasters: faster information processors and more reliable pattern detectors (Cattell & Horn, 1978).

To test this idea—as well as all other dispositional ideas in this section—we asked all forecasters to take a battery of tests at the start of each tournament. Superforecasters scored higher than top-team individuals and all others on all measures of fluid intelligence, including the Raven's Advanced Progressive Matrices (Arthur, Tubre, Paul, & Sanchez-Ku, 1999; Balboni, Naglieri, & Cubelli, 2010), the Shipley–2 Abstraction Test (Shipley, 2009), the Cognitive Reflection Test (Frederick, 2005), an extended version of the Cognitive Reflection Test (Baron, Scott, Fincher, & Metz, 2014), and portions of two Numeracy scales (Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006). Scores on the Shipley–2 indicated that superforecasters were at least one standard deviation higher than the general population on fluid intelligence.

Superforecasters may also have greater crystallized intelligence and, in particular, more domain-relevant crystallized intelligence, which means they know about world politics: who the key players are, what they want, and the economic and institutional constraints they face. We found that superforecasters did indeed have higher scores on tests of both the Shipley–2 Vocabulary (Shipley, 2009) and on tests of domestic political knowledge and international affairs. Superforecasters scored at least one standard deviation higher than the general population on crystallized intelligence and even higher on the political knowledge questions.

Cognitive styles may also play a role in performance, especially those that complemented their abilities: a competitive streak, greater appetites for intellectual challenges, and willingness to change their minds in response to new evidence. We asked all forecasters, "Why did you choose to participate in the tournament?" Superforecasters stood out most in their endorsements of "wanting to be among the top forecasters." Superforecasters also enjoyed solving problems more than top-team individuals and all others and scored higher on the Need for Cognition scale (Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984). Finally, superforecasters had higher scores than all others on actively open-minded thinking, a scale that has already been shown to predict both persistence in information search followed by greater accuracy in estimates of uncertain quantities (Haran, Ritov, & Mellers, 2013).

Superforecasters might also be more inclined to embrace a secular, agnostic/atheistic worldview that treats everything as subject to deterministic scientific laws and the mathematical laws of probability. This worldview predisposes superforecasters to treat their beliefs more as testable hypotheses and less as sacred possessions—and to be warier of overinterpreting coincidences by attributing them to supernatural mechanisms such as fate. Consistent with this hypothesis, superforecasters agreed more with items such as "Randomness is often a factor in our personal lives," and they disagreed more with items such as "Events unfold according to God's plan." Superforecasters also reacted differently to close-call counterfactuals. A common reaction, when learning that something came very close to happening/not happening (such as a couple who met in only one peculiar way as opposed to a couple who could have met in many possible ways), is to attribute the event to fate (it was meant to be; Kray et al., 2010). In a fully factorial between-subjects design (in which the three groups were assigned to read either close-call or a not-close-call stories), superforecasters were likeliest to attribute the close calls to coincidence (flip enough coins enough times, and unlikely things are bound to happen now and then), and they were least likely to invoke destiny or providence.

In Table 2, we summarize the comparison statistics bearing on the cognitive abilities and styles hypotheses discussed here, with possible ranges of variables in parentheses. As underscored in Table 2, superforecasters have distinctive dispositional profiles, scoring higher on several measures of fluid intelligence and crystallized intelligence, higher on the desire to be the best, the need for cognition, open-minded thinking, and endorsements of a scientific worldview with little tolerance for supernaturalism. Table 3 shows that these same variables correlate with forecasting accuracy.

Task-specific skills

Shifting from broad dispositional variables to much narrower task-specific skill sets, superforecasters again stand out. They have a deeper understanding of what needs to done to do well in a tournament that requires balancing outside and inside perspectives on events and translating vague qualitative impressions into precise probability metrics. We divide these skill sets into two categories: sensitivity to scope and sensitivity to more fine-grained (granular) probabilistic distinctions.

Scope insensitivity. Kahneman and Knetsch (1992) introduced the phrase *scope insensitivity* when exploring prices that emerged from contingent valuation methods. These are survey methods of assigning monetary values to public goods (goods that, by definition, cannot be traded directly in markets, such as clean air or pristine Alaskan shorelines). Extreme scope insensitivity occurs when measures of worth fail to budge across vast variations in the scope of the good. For example, the amount people said they would pay to save 2,000 migratory birds from an Alaskan oil spill was virtually identical to what they said they would pay to save 200,000 birds from the same spill (Desvouges et al., 1993; Schkade & Payne, 1994).

Scope insensitivity can take diverse forms in forecasting tournaments—and we ran a series of experimental comparisons of relative susceptibility using the five pairs of questions in Table 4. Each individual was randomly assigned to one version of each pair in a between-subjects design. For four of the pairs, the wider-scope-of-events class should be judged as likelier than the subset

Table 2. Measures Relevant to the Four Hypotheses About Why Superforecasters Are So Super

Measure	Superforecasters	Top-team individuals	All others	Controls
Cognitive abilities				
Raven's APM-SF (0–12)	9.13	**8.26**	**7.75**	
Shipley–2 Abstraction Test (0–25)	20.09	**18.58**	**18.49**	
CRT (0–3)	2.78	**2.46**	**2.26**	
Extended CRT (0–18)	16.64	**15.39**	**14.56**	
Numeracy (0–4)	3.67	**3.33**	**3.19**	
Shipley–2 Vocabulary (0–40)	37.50	**36.89**	**36.79**	
Political knowledge (Year 1; 0–35)	29.59	29.37	**28.66**	
Political knowledge (Year 2; 0–50)	38.45	**37.30**	**36.29**	
Political knowledge (Year 3; 0–55)	32.20	31.23	**31.12**	
Cognitive styles				
Why? Be at the top (1–7)	5.60	**4.86**	**4.81**	
Need for cognition (1–7)	5.97			
Open-minded (1–7)	6.01	5.97	**5.89**	
Belief in fate (0–7)	2.65			**3.17**
Close call (effect size)	.16			**.40**
Task-appropriate skills				
Scope sensitivity	.22			**0.12**
Granularity (unique numbers)	57	**29**	**30**	
Motivation: Skill cultivation				
Number of forecasting questions (Year 1)	76	**65**	**57**	
Number of forecasting questions (Year 2)	116	**84**	**82**	
Number of forecasting questions (Year 3)	81	**52**	**60**	
Motivation: Belief updating				
Average number of forecasts per question (Year 1)	2.77	**1.51**	**1.43**	
Average number of forecasts per question (Year 2)	5.64	**2.15**	**1.79**	
Average number of forecasts per question (Year 3)	6.70	**5.14**	**2.92**	
Motivation: Information gathering				
Average number of news clicks (Year 2)	187.31	**45.72**	**24.89**	
Average number of news clicks (Year 3)	344.73	**63.12**	**89.80**	
Enriched versus regular Environments				
Average number of comments (Year 2)	262.23	**51.88**		
Average number of comments (Year 3)	622.89	**112.26**		
Average number of words per comment (Year 2)	36.62	**28.49**		
Average number of words per comment (Year 3)	31.66	**24.80**		
Average number of forum posts (Year 2)	36.13	**2.25**		
Average number of forum posts (Year 3)	43.64	**4.94**		
Average number of news shared (Year 2)	91.57	**9.16**		
Average number of news shared (Year 3)	181.93	**28.61**		
Average % of comments with questions (Year 2)	0.47	**0.19**		
Average % of comments with questions (Year 3)	0.32	0.27		
Average % of replies (Year 2)	7.29	**2.59**		
Average % of replies (Year 3)	6.54	**2.99**		
Consensus rate within questions	–0.06	**0.05**	**0.06**	

Note: Bold values of top-team individuals and all others indicate a significant difference when compared with superforecasters at the .01 level. Values in parentheses beside each variable represent possible ranges of scores. Raven's APM-SF = Short Form of the Raven's Advanced Progressive Matrices; CRT = Cognitive Reflection Test.

class on the basis of elementary logic. For example, a dictator cannot be more likely to fall in 3 months than in 6 months. Superforecasters were more sensitive to scope—and they tended to report forecasts for the wider class that were greater than those for the subclass, even though they only saw one question and thus could not make direct comparisons of the sort possible in a within-subject design (which would have made the task much

Table 3. Correlates With Measures With Accuracy

Measure	Correlation	*t*(1774)	*p*
Raven's Advanced Progressive Matrices	−.18	−7.70	<.001
Shipley–2 Abstraction Test	−.22	−9.49	<.001
Shipley–2 Vocabulary	−.09	−3.80	<.001
CRT	−.16	−6.82	<.001
Extended CRT	−.23	−9.95	<.001
Numeracy	−.16	−6.82	<.001
Political knowledge (Year 1)	−.12	−5.09	<.001
Political knowledge (Year 2)	−.18	−7.70	<.001
Political knowledge (Year 3)	−.14	−5.95	<.001
Motivate—Be at the top	−.11	−4.66	<.001
Need for cognition	−.07	−2.95	<.002
Active open-mindedness	−.12	−5.09	<.001
Average number of articles checked	−.18	−7.70	<.001
Average number of articles shared	−.20	−8.53	<.001
Average number of comments with questions	−.18	−7.68	<.001
Average number of replies to questions	−.18	−7.70	<.001

Note: CRT = Cognitive Reflection Test.

easier and off-target). Bolded forecasts in Table 4 indicate significant differences between the set/subset (more likely/less likely) versions of a question. For instance, the first question measured sensitivity to quantity. When the official Euro to U.S. dollar exchange rate was 1.37, forecasters were asked whether the exchange rate would exceed 1.38 (or 1.40) before December 31, 2014. Superforecasters' responses of 73% and 48% for exchange rates of 1.38 and 1.40 differed significantly, but those of controls did not.

Questions 2, 3, and 4 had the same format as Question 1, but these questions focused on regional and temporal sensitivity. Both groups showed regional and temporal sensitivity in Questions 2 and 3, respectively, and neither showed temporal sensitivity in Question 4.

To analyze these four questions as a whole, we conducted a linear mixed effects model with forecasts as the dependent variable and group, scope, question (to account for heteroskedasticity in questions), and the interaction between group and scope as independent variables. Table 2 shows that superforecasters had a greater overall estimated difference between responses for more and less likely questions than that of controls (0.22 vs. 0.12). Of course, it is one thing to show that superforecasters are more scope sensitive than lower performing forecasters and quite another to show that superforecasters are as sensitive to scope as they should be—a complex problem to be tackled in future research.

Using the fifth pair of questions in Table 4, we explored sensitivity to an arbitrary anchor that forecasters should have ignored (as opposed to the scope sensitivity tests that focused forecasters on a cue they should have attended to). We asked participants to estimate the World Economic Output (WEO) for 2014. Prior to the estimation question, we introduced either a low or high question about whether the WEO would grow beyond 2.8% or 3.3% in 2014, respectively. Then we asked them to estimate the WEO in 2014. In this case, we hypothesized that superforecasters would be less sensitive to the anchor, a potential source of bias. Superforecasters' estimates were appropriately insensitive to the anchor, whereas controls' estimates covaried with anchors. Superforecasters were thus more resistant to the anchoring bias.

Forecasting granularity. Translating case-specific assessments of causal propensities (e.g., the balance of forces keeping Greece in vs. pushing Greece out of the Eurozone) into a probability assessment of Greece remaining in the Eurozone is a nontrivial challenge. Past work suggests that, not surprisingly, people tend to assign higher probabilities to outcomes when the net balance of causal forces favors that outcome. However, what can be said about the ability to provide nuanced predictions that actually reflect gradual shifts in the balance of causal forces?

One possibility is that superforecasters are better at translating complex qualitative causal judgments into probability judgments because they are more knowledgeable, more scope sensitive, and capable of finer grained distinctions along the probability scales. Probability scales can be "sliced" into segments that reflect different perceptual gradations of uncertainty, a process known as *granularity* (Yaniv & Foster, 1995). The decision weighting function in prospect theory

Table 4. Scope Sensitivity

Question	Numerical	Superforecasters	Controls
Question 1			
Less likely	Will the official Euro to U.S. dollar exchange rate exceed 1.40 before December 31, 2014?[a]	**48%**	42%
More likely	Will the official Euro to U.S. dollar exchange rate exceed 1.38 before December 31, 2014?	**73%**	51%
Question 2: Regional			
Less likely	Will Israel deploy at least one unmanned aerial vehicle (UAV) over the territory of Iran before December 1, 2014?	**28%**	**47%**
More likely	Will Israel deploy at least one UAV over the territory of another country before December 1, 2014?	**60%**	**59%**
Question 3: Temporal			
Less likely	Will Turkey ratify a new constitution by February 1, 2014?	**7%**	**30%**
More likely	Will Turkey ratify a new constitution by February 1, 2016?	**35%**	**49%**
Question 4: Temporal			
Less likely	Will there be a significant lethal confrontation in the Middle East between Syria and Turkey before July 1, 2014?	16%	26%
More likely	Will there be a significant lethal confrontation in the Middle East between Syria and Turkey before July 1, 2015?	20%	34%
Question 5: Anchoring			
Low	Do you think the World Economic Output (WEO) will grow beyond 2.8% in 2014?[b] What is your best estimate of the WEO in 2014?	3.2%	**2.7%**
High	Do you think the WEO will grow beyond 3.3% in 2014? What is your best estimate of the WEO in 2014?	3.4%	**3.1%**

[a]This question was asked when the exchange rate was 1.37. [b]This question was asked when the WEO projection for 2013 was 2.9%.
Note: Bold values are significantly different at the .01 level.

suggests that, in their decisions, people give greater weight to changes in probabilities at the extremes than to changes in the middle region between 0.2 and 0.8 (Kahneman & Tversky, 1979). Superforecasters might make more probabilistic distinctions than others, and perhaps more distinctions at the extremes, reflecting their finer grained appreciation of uncertainty.

We explored the translation of information into probability judgments in two ways. First, we examined the total number of unique probability numbers (i.e., 0–100) made by individuals across all questions they attempted. Table 2 shows that these averages were 57 for superforecasters, 29 for top-team individuals, and 30 for all others, respectively. Superforecasters submitted almost twice as many unique numerical predictions overall as the other groups.

Superforecasters also selected more numbers that reflected granularity. For each participant, we examined the percentage of forecasts that were multiples of 10%, multiples of 5% (that were not also of 10%), and multiples of 1% (that were not also of 10% or 5%). Top-team individuals and all others were most likely to make forecasts divisible by 10% (10%, 20%, 30%, etc.). By contrast, superforecasters were most likely to make forecasts divisible by 1% and only 1% (e.g., 17%, 28%, and 83%, and excluding all multiples of 5% and 10%). Superforecasters made more granular forecasts.

Greater granularity does not necessarily imply greater accuracy. To explore this relationship, we rounded forecasts to the nearest 0.05, 0.10, or 0.33 to see whether Brier scores became less accurate on the basis of rounded forecasts rather than unrounded forecasts. Less accuracy after rounding to the nearest 0.05, 0.10, or 0.33 implies that forecasters used at least 21, 11, or 4 distinctions along the probability scale, respectively. If Brier scores were significantly worse after moving from no rounding to 0.05, we would conclude that the more granular probabilities contained more information at their initial level than at 21 categories, and those forecasters were using at least 21 categories. If Brier scores were significantly worse after rounding from 0.10 to 0.33, we would conclude that information was lost if forecasters went from 11 to 4 categories, so forecasters were using at least 4 categories.

For superforecasters, rounding to the nearest 0.10 produced significantly worse Brier scores. However, for the other two groups, rounding to the nearest 0.10 had no influence. It was not until rounding was done to the nearest 0.33 that accuracy declined. In short, information was lost when superforecasters went from 10 to 4 categories. Superforecasters were using at least 4 categories but quite likely more. Probabilistic distinctions made by the two comparison groups were markedly cruder.

Motivation and commitment

Everyone knows the old joke about how to get to Carnegie Hall: practice, practice, practice. Highly skilled performance depends on intense, focused, and long-term commitment. Ericsson, Krampe, and Romer (1993) have argued that expert performance is the end result of prolonged, deliberate practice, and even among elite performers, performance skill correlates with the amount of deliberate practice. Deliberate practice or "grit" predicts grade point averages of Ivy League undergraduate students, student retention at West Point, and rankings in the National Spelling Bee (Duckworth, Peterson, Matthew, & Kelly, 2007).

In a parallel research program, Dweck (2006) found that better performance is associated with the belief that skills are learned rather than innate. Successful forecasters in our tournaments may be likelier to view the task as a cultivatable skill rather than a God-given or DNA-given endowment. Forecasters with a growth-mindset orientation should presumably try more questions and update their predictions more often.

Table 2 shows that, even in Year 1 (before being anointed as a super), superforecasters engaged in patterns of behavior suggestive of the view that forecasting skill can be cultivated via deep deliberative practice. One measure of motivation and commitment is the number of questions attempted by each forecaster. In all 3 years of the tournament, superforecasters attempted more questions than top-team individuals or all others. In Year 1, they attempted 25% more questions than the other groups—and in later years, the effort gap grew, with superforecasters attempting approximately 40% more questions than the other groups. Another measure of motivation and commitment is the frequency with which forecasters updated their beliefs. Table 2 shows that superforecasters updated their beliefs more often than either top-team individuals or all others. Even in Year 1, superforecasters made an average of 2.77 forecasts per question, whereas the comparison groups made an average of 1.47 forecasts per question. Differences continued in Years 2 and 3. Frequency of belief updating was important; it turned out to be the strongest single behavioral predictor of accuracy (Mellers, Stone, et al., 2014).

In Years 2 and 3, the GJP website provided a news reader that used Google search queries with keywords to help forecasters collect articles from reputable and relevant sources. We counted how often forecasters clicked on the news reader. During Years 2 and 3, superforecasters clicked on an average of 255 stories, significantly more than top-team individuals and all others, who clicked on 55 and 58 stories, respectively. In sum, several variables suggest that superforecasters were more committed to cultivating skills than were top-team individuals or all others.

Enriched environments

A large literature on peer effects in the classroom suggests that students benefit from working in cohorts of similar ability levels (see Epple & Romano, 2011, for a review). Grouping students on the basis of prior performance can accelerate learning, especially among high achievers, by motivating each other and making the challenges more enjoyable. Of course, grouping is also controversial: It increases inequality (just as our superforecaster manipulation did).

Table 2 shows that elite superforecaster teams engaged with each other more frequently than regular teams. Their engagement could occur in two ways. Forecasters could post specific comments about a question beside the question or make general comments on a forum about broader topics, such as cognitive triage (deciding how to allocate effort across questions). Table 2 shows that in Years 2 and 3, superforecasters communicated more on both variables. They made roughly 5 times more specific-question comments than top-team individuals, and superforecasters' comments were roughly one third longer than those of top-team individuals. Furthermore, in the forum, superforecasters posted an average of 36 and 44 general comments in Years 2 and 3, respectively, whereas top-team individuals posted an average of only 2 and 5 comments, respectively.

Previously, we argued that superforecasters were likelier than other groups to gather news and opinion pieces. They were also likelier to share those news stories and opinion pieces with teammates, consistent with the enriched-environment hypothesis. Table 2 shows that superforecasters shared more than 10 times as many news links as top-team individuals in Year 2 and more than 6 times as many in Year 3.[2]

Superforecasters were more willing to probe the knowledge of their teammates. The proportion of sentences containing a question mark relative to total words was 0.47% (or 1 out of every 213 words) for superforecasters and was 0.19% (1 out of every 526 words) for top-team individuals. In addition to asking more questions, superforecasters provided more answers to posted inquiries. On average, 11% of superforecasters' comments were replies, whereas only 2% of top-team individuals' comments were replies to previous questions.

This difference in responsiveness became even starker when we compared the percentage of comments containing question marks that received replies relative to the total number of comments with question marks. On average, superforecasters got replies to 23% of their questions. Average top-team individuals received replies to only 4% of their questions. Superforecasters felt accountable to each other in ways that top-team individuals did not. Again, these variables were predictors of accuracy (see Table 4).

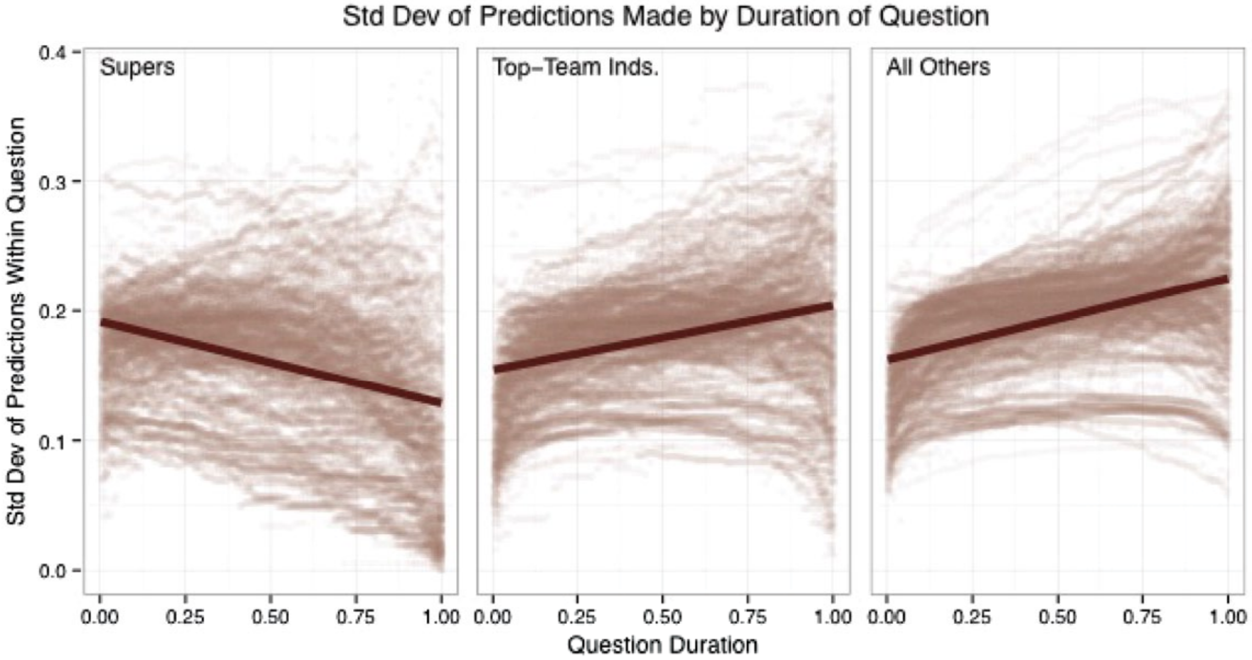

Fig. 5. Average daily standard deviations (Std Devs) in forecasts over participants plotted against the percentage duration of a question with separate linear regressions for superforecasters (Supers), top-team individuals (Top-Team Inds.), and all others. The steeper negative slope for Supers means that they were quicker to reach consensus (greater area under the curve).

Well-functioning teams might disagree initially on a question but still reach consensus as evidence mounts and time passes. We examined daily variability in forecasts as a function of the number of days that questions were open (measured as a percentage of total life span of the question—ranging from 0% [*start*] to 100% [*finish*]). Figure 5 shows the standard deviation in daily forecasts against the percentage of the elapsed life span of the question, averaged over questions. Dark red lines are regression lines, and values appear in Table 2. Rising red lines indicate that top-team individuals and all others had more, not less, forecast variability with the passage of time. However, for superforecasters, the regression line points downward, indicating greater consensus over time. Sharing more news, posing more questions, and providing more answers helped superforecasters reach consensual closure faster and more efficiently than the other comparison groups—and these actions appear to do so at little or no risk of groupthink or premature closure.

Discussion

Are superforecasters different kinds of people—or simply people who do different kinds of things? Our data suggest that the answer is a blend of both. The superforecaster phenomenon has implications for debates about the development of expertise, the role of intelligence and other individual differences in success, and the robustness of cognitive biases. We highlight four sets of findings that various theoretical camps are likely to find either dissonant or congenial:

1. Psychologists who believe that deliberative practice in skill development has been underplayed (Duckworth et al., 2007; Ericsson et al., 1993) might understandably find it tempting to treat superforecasters as simply diligent workers—and to posit that virtually anyone could do what these forecasters did if they practiced as deeply and deliberatively as superforecasters. This interpretation fails, however, to take into account the many pretournament ways in which superforecasters differed from others forecasters and, even more so, from the general population. Superforecasters had cognitive abilities and styles that may have enabled them both to start the tournament with more relevant skills, to perform better from the outset, and then to fine tune those skills more rapidly via deliberate practice.
2. Psychologists who believe that human intelligence has been underplayed in skill development can point to the consistency with which fluid intelligence (operationalized in diverse ways) emerges as a reliable predictor of performance across a heterogeneous array of forecasting problems

(Hunter & Schmidt, 1996). In this view, the door to superforecaster status is not open to all (Gottfredson, 1997). Above-average fluid intelligence (one standard deviation plus) is almost a defining feature of superforecasters. However, so too is a strong behavioral commitment to cultivating the skill of assigning realistic probability estimates. Superforecasters try a lot harder, attempting many more questions and updating their beliefs more often as new evidence arises.

3. Psychologists who believe that cognitive style is not just reducible to cognitive ability can point to the predictive roles played by the need for cognition and actively open-minded thinking. Superforecasters derive more enjoyment from problem solving and are more willing to treat their belief as testable propositions, not sacred possessions to be defended to the death (Baron & Spranca, 1997; Fiske & Tetlock, 1997). The cognitive style effects are, however, markedly weaker than those of both fluid intelligence and of behavioral commitment (e.g., frequency of updating).
4. Psychologists who believe that people can overcome even biases rooted in basic perceptual-cognitive processes, such as anchoring effects and scope insensitivity (Kahneman, 2011), may be encouraged by some results. When people either have the right dispositional profile (fluid intelligence plus open-minded thinking) or have been properly motivated by tournaments to monitor their thought processes and to execute System 2 overrides of System 1, they can reduce biases. However, the biases are not eliminated entirely. Even superforecasters show some scope insensitivity, and even superforecasters may not have used all the numbers that were meaningfully different on the probability scale. There may well be perceptual-cognitive constraints on even the best human probability judgments.

Our analysis leaves many theoretical questions dangling: Is it possible—as advocates of self-fulfilling prophecy arguments might propose—to transform ordinary forecasters into superforecasters by simply labeling them "high-potential late bloomers"? The answer to this question requires experimental manipulation of the labeling of moderate-to-high-functioning forecasters. Is it possible—as advocates of the enriched-environment hypothesis might propose—to transform ordinary forecasters into high performers by just giving them access to the conversations, debates, and deliberations of superteams? Furthermore, why have we not seen interactions among independent variables that many theorists might expect? Does not fluid intelligence matter more when one is higher in crystallized intelligence—and has more information to bring to bear in problem solving? Or does not fluid intelligence confer more advantages when it is joined to an open-minded cognitive style or to a deliberative practice work ethic? So far, these interactions have not materialized as reliable predictors of forecaster accuracy. Are these nonoccurrences a sign that our underlying psychological assumptions are wrong—or a sign that the tournament lacks the massive statistical power needed to test the hypotheses? We suspect the answer is the latter. Real-world accuracy data are noisy. Some questions are resolved as very close calls that could easily have gone the other way. Furthermore, forecaster performance is also affected by changes in forecasters' personal and professional lives outside the tournament, such as the birth of a child or a work schedule that suddenly tightens up.

To summarize, the study of forecasting ability in the real world can create odd theoretical bedfellows. How often in the mainstream research literatures do references to fluid intelligence appear alongside references to scope sensitivity—or references to granularity appear alongside references to behavioral commitments and growth mindsets or self-fulfilling prophecies? As we ourselves well know, these real-world environments motivate investigators to be more eclectic than they would be if they were playing the hypothesis testing game under normal laboratory conditions.

Critics of tournaments might dismiss our experiments and interventions as a crude throw-everything-we-know-at-the-problem approach. However, the Zimbardo prison "experiment" (Zimbardo, 1973) and the original Milgram obedience study (Milgram, 1963) were also heavy-handed, multidimensional interventions. These researchers wanted to make a larger societal point, not nail down specific hypotheses. This approach seems fully justified when we are talking about reducing the likelihood of multi-trillion-dollar mistakes that claim thousands of lives. The U.S. Intelligence Community had been ferociously criticized for missing the rise of the Al-Qaeda threat prior to 9/11 and for claiming that there were weapons of mass destruction in Iraq when, in fact, there were none.

There is, of course, no guarantee that improving the Brier scores of intelligence analysts by the amounts that we achieved (i.e., 50% or 60%) would have prevented any particular intelligence failure (Jervis, 2010). However, there is a compelling case that by improving Brier scores, analysts bring down the overall likelihood of the two canonical prediction errors that bedevil any forecasting system: false-positives (e.g., telling the President that something is a slam dunk, and that thing does not happen) and false-negatives (e.g., telling the President something is impossible, and it happens).

Werner Heisenberg (1958/1999) famously remarked that what we observe is not nature itself but nature exposed to our mode of questioning. This observation applies as much to psychology as to quantum physics. Researchers' understanding of flaws in intuitive forecasting is largely grounded in laboratory studies of typical performance; their understanding of what people are capable of doing is grounded in real-world tournaments that encourage researchers and forecasters to work together to generate the most accurate possible probabilities using the best possible methods. Superforecasters have achieved a surprising degree of accuracy—and this may be just the beginning of those surprises.

Acknowledgments

The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation thereon. 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 the Intelligence Advanced Research Projects Activity, the Department of Interior National Business Center, or the U.S. Government.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by Research Contract D11PC20061 to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity via the Department of Interior National Business Center.

Notes

1. Figure 1 shows the Year 1 Brier scores for only 45 of the 60 superforecasters. The others served in the prediction market in Year 1, so we were unable to compute Brier scores for them individually.
2. Participants who did not make at least one comment during the year with at least 50 total words were excluded from the analyses.

References

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*, 354–361.

Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2014). *Distilling the wisdom of crowds: Prediction markets versus prediction polls*. Manuscript under review.

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*, 222–235.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). New York, NY: Cambridge University Press.

Baron, J., Scott, S. E., Fincher, K., & Metz, S. E. (2014). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, Special Issue on Modeling and Aiding Intuitions in Organizational Decision Making*. Advance online publication. doi:10.1016/j.jarmac.2014.09.003

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes*, *70*, 1–16.

Betts, J. R., & Shkolnik, J. L. (2000). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review*, *19*, 1–15.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology*, *45*, 792–804.

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

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131.

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

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement*, *15*, 139–164.

Christensen-Szalanski, J., & Wilhelm, C. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *48*, 147–168.

Cooper, A., Woo, C., & Dunkelberg, W. (1988). Entrepreneurs' perceived chances for success. *Journal of Business Venturing*, *3*, 97–108.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1673.

Desvouges, W. H., Johnson, F., Dunford, R., Hudson, S., Wilson, K., & Boyle, K. (1993). Measuring resource damages with contingent valuation: Tests of validity and reliability. In J. A. Hausman (Ed.), *Contingent valuation, A critical assessment* (pp. 91–164). Amsterdam, The Netherlands: North Holland.

Duckworth, A. L., Peterson, C., Matthew, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Ballantine Books.

Epple, D., & Romano, R. (2011). Peer effects in education: A survey of the theory and evidence. *Handbook of Social Economics*, *1*, 1053–1163.

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

Fiske, A., & Tetlock, P. E. (1997). Taboo tradeoffs: Reactions to transgressions that transgress spheres of influence. *Political Psychology*, *18*, 255–297.

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

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

Gottfredson, L. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79–132.

Haran, U., Ritov, I., & Mellers, B. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Journal of Decision Making*, *8*, 188–201.

Hartzmark, M. L. (1991). Luck versus forecast ability: Determinants of trader performance in futures markets. *The Journal of Business*, *64*, 49–74.

Heisenberg, W. (1999). *Physics and philosophy: The revolution in modern science*. Amherst, NY: Prometheus Books. (Original work published 1958)

Hunter, J. E., & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and the Law*, *2*, 447–472.

Jagacinski, R. J., Isaac, P. D., & Burke, M. W. (1977). Application of signal detection theory to perceptual-motor skills: Decision processes in basketball shooting. *Journal of Motor Behavior*, *9*, 225–234.

Jervis, R. (2010). *Why intelligence fails*. Ithaca, NY: Cornell University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Cambridge University Press.

Kahneman, D., & Knetsch, J. (1992). Valuing public goods: The purchase of moral satisfaction. *Journal of Environmental Economics and Management*, *22*, 57–70.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–292.

Kahneman, D., & Tversky, A. (1985). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). New York, NY: Cambridge University Press.

Kray, L. J., George, L. G., Galinsky, A. D., Roese, N. J., Liljenquist, K. A., & Tetlock, P. E. (2010). From what might have been to what must have been: Counterfactual thinking creates meaning. *Journal of Personality and Social Psychology*, *98*, 106–118.

Levitt, B., & March, J. G. (1988). Organizational learning. *Annual Review of Sociology*, *14*, 319–340.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

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

Mellers, B. A., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, E. S., Ungar, L., . . . Tetlock, P. (2014). *The psychology of intelligence analysis: Drivers of prediction accuracy in world politics*. Manuscript under review.

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. (2014). Psychological strategies for winning a geopolitical tournament. *Psychological Science*, *25*, 1106–1115.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*, 1330–1338.

Ofir, C. (1988). Pseudodiagnosticity in judgment under uncertainty. *Organizational Behavior and Human Decision Making*, *42*, 343–363.

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

Rosenthal, R. (1966). *Experimenter effects in behavioral research*. East Norwalk, CT: Appleton-Century-Crofts.

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

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

Schkade, D., & Payne, J. (1994). How people respond to contingent valuation questions: A verbal protocol analysis of willingness to pay for an environmental regulation. *Journal of Environmental Economics and Management*, *26*, 88–109.

Shipley, W. C. (2009). *Shipley–2 manual*. Torrance, CA: Western Psychological Services.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*, 143–148.

Swets, J. A., Dawes, R., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Tetlock, P. E., & Mellers, B. A. (2002). The great rationality debate. *Psychological Science*, *13*, 94–99.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, *20*, 273–281.

Yaniv, I., & Foster, D. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness tradeoff. *Journal of Experimental Psychology: General*, *4*, 424–432.

Zimbardo, P. (1973). On the ethics of intervention in human psychological research: With special reference to the Stanford Prison Experiment. *Cognition*, *2*, 243–256.