Journal of Experimental Psychology: Applied
2015, Vol. 21, No. 1, 1–14

1076-898X/15/$12.00 http://dx.doi.org/10.1037/xap0000040

The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar, Michael M. Bishop, and Michael Horowitz
University of Pennsylvania

Ed Merkle
University of Missouri

Philip Tetlock
University of Pennsylvania

This article extends psychological methods and concepts into a domain that is as profoundly consequential as it is poorly understood: intelligence analysis. We report findings from a geopolitical forecasting tournament that assessed the accuracy of more than 150,000 forecasts of 743 participants on 199 events occurring over 2 years. Participants were above average in intelligence and political knowledge relative to the general population. Individual differences in performance emerged, and forecasting skills were surprisingly consistent over time. Key predictors were (a) dispositional variables of cognitive ability, political knowledge, and open-mindedness; (b) situational variables of training in probabilistic reasoning and participation in collaborative teams that shared information and discussed rationales (Mellers, Ungar, et al., 2014); and (c) behavioral variables of deliberation time and frequency of belief updating. We developed a profile of the best forecasters; they were better at inductive reasoning, pattern detection, cognitive flexibility, and open-mindedness. They had greater understanding of geopolitics, training in probabilistic reasoning, and opportunities to succeed in cognitively enriched team environments. Last but not least, they viewed forecasting as a skill that required deliberate practice, sustained effort, and constant monitoring of current affairs.

Keywords: forecasting, predictions, skill, probability judgment, accuracy

Supplemental materials: http://dx.doi.org/10.1037/xap0000040.supp

Predicting the future is an integral part of human cognition. We reach for an umbrella when we expect rain. We cross the street when the light turns green and expect cars to stop. We help others and expect reciprocity—they will help us in future situations. Without some ability to generate predictions, we could neither plan for the future nor interpret the past.

Psychologists have studied the accuracy of intuitive predictions in many settings, including eyewitness testimony (Loftus, 1996), affective forecasting (Wilson & Gilbert, 2005), and probability judgment (Kahneman, Slovic, & Tversky, 1982). This literature paints a disappointing picture. Eyewitness testimonies are often faulty (Wells, 2014; Wells & Olson, 2003), affective forecasts stray far from affective experiences (Schkade & Kahneman, 1998), and probability estimates are highly susceptible to overconfidence, base rate neglect, and hindsight bias (Fischhoff & Bruine de Bruin, 1999; Fischhoff, Slovic, & Lichtenstein, 1977; Kahneman et al., 1982).

To make matters worse, intuitive predictions are often inferior to simple statistical models in domains ranging from graduate school admissions to parole violations (Dawes, Faust, & Meehl, 1989; Swets, Dawes, & Monahan, 2000). In political forecasting, Tetlock (2005) asked professionals to estimate the probabilities of events

This article was published Online First January 12, 2015.

Barbara Mellers, Eric Stone, Pavel Atanasov, Nick Rohrbaugh, S. Emlen Metz, Department of Psychology, University of Pennsylvania; Lyle Ungar, Department of Computer Science, University of Pennsylvania; Michael M. Bishop, Department of Psychology, University of Pennsylvania; Michael Horowitz, Department of Political Science, University of Pennsylvania; Ed Merkle, Department of Psychology, University of Missouri; Philip Tetlock, Department of Psychology, University of Pennsylvania.

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center (DoI/NBC) contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. The authors declare that they had no conflicts of interest with respect to their authorship or the publication of this article. The authors thank Jonathan Baron for helpful comments on previous drafts of the article.

Correspondence concerning this article should be addressed to Barbara Mellers, Department of Psychology, 3720 Walnut Street, Solomon Labs, University of Pennsylvania, Philadelphia, PA 19104. E-mail: mellers@wharton.upenn.edu

up to 5 years into the future—from the standpoint of 1988. Would there be a nonviolent end to apartheid in South Africa? Would Gorbachev be ousted in a coup? Would the United States go to war in the Persian Gulf? Experts were frequently hard-pressed to beat simple actuarial models or even chance baselines (see also Green and Armstrong, 2007).

A Forecasting Competition

It was against this backdrop that the National Academy of Sciences issued a report on the quality of intelligence analysis (Fischhoff & Chauvin, 2011). A key theme was the need to systematically track the accuracy of probabilistic forecasts that analysts routinely (albeit covertly) make. In response, the Intelligence Advanced Research Projects Activity (IARPA), the research and development branch of the Office of the Director of National Intelligence, launched a large-scale forecasting tournament designed to monitor the accuracy of probabilistic forecasts about events that occurred around the world. Five university-based research groups competed to develop methods to elicit and aggregate forecasts to arrive at the most accurate predictions. Our research group consistently outperformed the other groups 2 years in a row.

Within the tournament, accuracy of probability judgments was assessed by the Brier scoring rule (Brier, 1950), a widely used measure in fields ranging from meteorology (Murphy & Winkler, 1984) to medical imaging (Itoh et al., 2002; Steyerberg, 2009). The Brier scoring rule is "strictly proper" in the sense that it incentivizes forecasters to report their true beliefs—and avoid making false-positive versus false-negative judgments. These scores are sums of squared deviations between probability forecasts and reality (in which reality is coded as "1" for the event and "0" otherwise). They range from 0 (best) to 2 (worst). Suppose a forecaster reported that one outcome of a two-option question was 75% likely, and that outcome occurred. The forecaster's Brier score would be $(1 - 0.75)^2 + (0 - 0.25)^2 = 0.125$. This measure of accuracy is central to the question of whether forecasters can perform well over extended periods and what factors predict their success.

Consistency in Forecasting Skills

In this article, we study variation in the degree to which people possess, and are capable of developing, geopolitical forecasting skill. Skill acquisition and expertise has been examined in numerous domains. We were unsure whether it was even possible to develop skill in this domain. Geopolitical forecasting problems can be complex, requiring a balance of clashing causal forces. It is no wonder that some attribute forecasting success to skill, whereas others attribute it to luck. Skeptics argue that accurate forecasts are fortuitous match-ups between reality and observers' preconceptions in a radically unpredictable world. From this perspective, we would find little or no consistency in individual accuracy across questions (Almond & Genco, 1977; Taleb, 2007).

Our prediction task involves several factors usually associated with poor performance, including a dynamic prediction environment, a long delay before feedback on most questions, the lack of empirically tested decision aids, and a reliance on subjective judgment (Shanteau, 1992). Indeed, Reyna, Chick, Corbin, and Hsia (2014) showed that intelligence analysts were more susceptible to risky-choice framing effects than either college students or postcollegiate adults, perhaps because they had developed bad habits in an "unfriendly" environment. Although experts may, on average, be poor at exercising good judgment in complex domains like geopolitical forecasting, others suspect that there are systematic individual differences—and that some forecasters will consistently outperform others (Bueno de Mesquita, 2009; Tetlock, 2005).

As we shall soon show, striking individual differences in forecasting accuracy emerged, and these differences created the opportunity to test hypotheses about which assortment of dispositional variables (e.g., cognitive abilities and political understanding), situational variables (e.g., cognitive-debiasing exercises), and/or behavioral variables (e.g., willingness to revisit and update one's beliefs) could predict judgmental accuracy. Insofar as all three classes of variables matter, how are they interrelated? And what are the characteristics of the best forecasters?

Dispositional Variables

Accurate predictions of global events require an array of skills. One needs diverse pockets of content knowledge, a judicious capacity to choose among causal models for integrating and applying content knowledge, and a rapid-fire Bayesian capacity to change one's mind quickly in response to news about shifting base rates and case-specific cues. A natural starting hypothesis is intelligence, a well replicated predictor of success, including job performance (Ree & Earles, 1992; Schmidt & Hunter, 2004), socioeconomic status (Strenze, 2007), academic achievement (Furnham & Monsen, 2009), and decision competence (Del Missier, Mäntylä, & Bruine de Bruin, 2012; Parker & Fischhoff, 2005).

Intelligence

Theories of intelligence vary in complexity, starting with the single-factor model widely known as *g* (Spearman, 1927), the two-factor fluid/crystallized intelligence framework (Cattell, 1963; Cattell & Horn, 1978), the seven basic abilities (Thurstone & Thurstone, 1941), and, finally, the 120-factor cube derived from combinations of content, operation, and product (Guilford & Hoepfner, 1971). Carroll (1993) reanalyzed over 400 data sets that measured cognitive abilities and found overwhelming evidence for a general intelligence factor (interpreted as *g*, fluid intelligence, with domain-specific forms of crystallized intelligence defining additional factors).

Three aspects of intelligence seem most relevant to geopolitical forecasting. One is the ability to engage in *inductive reasoning*, or make associations between a current problem—say, the likelihood of an African leader falling from power—and potential historical analogies. Individuals must look for regularities, form hypotheses, and test them. The second is *cognitive control* (also known as cognitive reflection). Someone with greater cognitive control has the ability to override seemingly obvious but incorrect responses and engage in more prolonged and deeper thought. The third skill is *numerical reasoning*. Numeracy would be especially important for economic questions such as, "Will the price per barrel for November, 2011 Brent Crude oil futures exceed $115 by a given date?" A more numerate forecaster would be likelier to recognize that the answer hinged, in part, on how close the current price was

to the target price and how often price fluctuations of the necessary magnitude occurred within the specified time frame. Our first hypothesis is therefore as follows:

> *Hypothesis 1:* Individuals with greater skill at inductive reasoning, cognitive control, and numerical reasoning will be more accurate forecasters.

Researchers disagree on the relationship between intelligence and expertise. Some claim that experts, such as chess grandmasters, possess exceptional cognitive abilities that place them high up in the tail of the distribution (Plomin, Shakeshaft, McMillan, & Trzaskowski, 2014). Others claim that, beyond a certain moderately high threshold, intelligence is not necessary; what really matters is deep deliberative practice that promotes expertise by enabling the neural networking and consolidation of performance-enhancing knowledge structures (Ericsson, 2014).

The forecasting tournament let us explore the relationship between intelligence and skill development. If the correlation between intelligence and accuracy was positive and remained constant throughout the tournament, one could argue that superior intelligence is necessary for expertise. But if the correlation between intelligence and accuracy were stronger at the beginning and weaker toward the end of the tournament (after deliberative practice), one could argue that deliberative practice is a cognitive leveler, at least within the ability range of the above-average IARPA forecasters.

Thinking Style

Cognitive styles capture *how* people typically think—as opposed to what they think about (e.g., causal theories) and how well they can think (ability). There are as many frameworks for cognitive styles as taxonomies of cognitive abilities (Riding & Cheema, 1991; Vannoy, 1965; Witkin, Oltman, Raskin, & Karp, 1971).

A relevant cognitive style is the tendency to evaluate arguments and evidence without undue bias from one's own prior beliefs—and with recognition of the fallibility of one's judgment (Nickerson, 1987). High scorers on this dimension are *actively open-minded thinkers*. They avoid the "myside bias"—the tendency to bolster one's own views and dismiss contradictory evidence (Baron, 2000). Actively open-minded thinkers have also been found to be more accurate at estimating uncertain quantities (Haran, Ritov, & Mellers, 2013), a task that is arguably similar to estimating the likelihood of future events.

Actively open-minded thinkers also have greater tolerance for ambiguity and weaker *need for closure* (the tendency to want to reach conclusions quickly, often before all the evidence has been gathered, coupled with an aversion to ambiguity; Kruglanski & Webster, 1996; Webster & Kruglanski, 1994). Previous research has found that experts with a greater need for closure reject counterfactual scenarios that prove their theories wrong while embracing counterfactual scenarios that prove their theories right (Tetlock, 1998), a form of motivated reasoning that is likely to hinder attempts to accurately model uncertainty in the real world.

In a related vein, the concept of *hedgehogs versus foxes*, developed by Tetlock (2005), draws on need for closure and taps into a preference for parsimony in political explanations (the hedgehog knows one big thing) versus a preference for eclectic blends of causal precepts (the fox knows many, not-so-big things). Tetlock found that the foxes were less prone to overconfidence in their political predictions. Although we measured actively open-minded thinking, need for closure, and hedgehog versus fox separately, these constructs reflect distinct but related features of cognitive flexibility. Given the strong family resemblance among openness to self-correction, cognitive flexibility, foxiness, and tolerance for ambiguity, we bundle them into our next hypothesis. Forecasters with more open-minded and flexible cognitive styles should be more nuanced in applying pet theories to real-world events—or, more simply,

> *Hypothesis 2:* More open-minded forecasters will be more accurate forecasters.

Political Knowledge

Political knowledge refers to content information necessary for answering factual questions about states of the world. Even the most intelligent and open-minded forecasters need political knowledge to execute multidimensional comparisons of current events with pertinent precedents. Consider the question, "Will the United Nations General Assembly recognize a Palestinian state by September, 30, 2011?" Forecasters with no institutional knowledge would be at a disadvantage. They might read headlines that a majority of the General Assembly favored recognition and infer that recognition was imminent. But someone who knew more about the United Nations might know that permanent members of the Security Council have many ways to delay a vote, such as "tabling the resolution" for a later point in time. This brings us to our third hypothesis:

> *Hypothesis 3:* More politically knowledgeable forecasters will be more accurate forecasters.

Situational Variables

Forecasting accuracy also depends on the environment; forecasters need opportunities for deliberative practice to cultivate skills (Arkes, 2001; Ericsson, Krampe, & Tesch-Romer, 1993; Kahneman & Klein, 2009). Some environments lack these opportunities. Cue-impoverished environments stack the odds against forecasters who wish to cultivate their skills. Environments with delayed feedback, misleading feedback, or nonexistent feedback also restrict learning (Einhorn, 1982).

Mellers, Ungar, et al. (2014) reported two experimentally manipulated situational variables that boosted forecasting accuracy. One was training in probabilistic reasoning. Forecasters were taught to consider comparison classes and take the "outside" view. They were told to look for historical trends and update their beliefs by identifying and extrapolating persistent trends and accounting for the passage of time. They were told to average multiple estimates and use previously validated statistical models when available. When not available, forecasters were told to look for predictive variables from formal models that exploit past regularities. Finally, forecasters were warned against judgmental errors, such as wishful thinking, belief persistence, confirmation bias, and hindsight bias. This training module was informed by a large literature that investigates methods of debiasing (see Lichtenstein

and Fischhoff, 1980; Soll, Milkman, and Payne, in press; and Wilson and Brekke, 1994, for reviews).

The second situational factor was random assignment to teams. Drawing on research in group problem-solving (Laughlin, 2011; Laughlin, Hatch, Silver, & Boh, 2006; MacCoun, 2012; Steiner, 1972), Mellers, Ungar, et al. (2014) designed teams with the goal of ensuring the "process gains" of putting individuals into groups (e.g., benefits of diversity of knowledge, information sharing, motivating engagement, and accountability to rigorous norms) exceeded the "process losses" from teaming (e.g., conformity pressures, overweighting common information, poor coordination, factionalism). The manipulation was successful. Teamwork produced enlightened cognitive altruism: Forecasters in teams shared news articles, argued about the evidence, and exchanged rationales using self-critical epistemic norms. Forecasters who worked alone were less accurate. Here, we explore whether the dispositional variables discussed earlier add to the predictive accuracy of forecasting over and beyond the two situational variables already known to promote accuracy. Our fourth hypothesis is

> *Hypothesis 4:* Dispositional variables, such as intelligence, open-mindedness, and political knowledge will add to the prediction of forecasting accuracy, beyond situational variables of training and teamwork.

Behavioral Variables

Dweck (2006) argues that those with growth mind-sets who view learning and achievement as cultivatable skills are likelier to perform well than those who view learning as innately determined. More accurate forecasters are presumably those with growth mind-sets. In the tournament, behavioral indicators of motivation included the numbers of questions tried and the frequency of belief updating. Engaged forecasters should also spend more time researching, discussing, and deliberating before making a forecast. Our fifth hypothesis is

> *Hypothesis 5:* Behavioral variables that reflect engagement, including the number of questions tried, frequency of updating, and time spent viewing a question before forecasting will add to the prediction of forecasting accuracy, beyond dispositional and situational variables.

Overview

After testing these hypotheses, we build a structural equation model to summarize the interrelationships among variables. Then we develop a profile of the best forecasters. Finally, we take a practical stance and ask, when information is limited, which variables are best? Imagine a forecaster who "applies for the job" and takes a set of relevant tests (i.e., dispositional variables). We might also know the forecaster's working conditions (i.e., situational variables). We could then "hire" the forecaster and monitor work habits while "on the job" (i.e., behavioral variables). Which type of variables best identifies those who make the most accurate forecasts?

Method

The forecasting tournament was conducted over 2 years, with each year lasting about 9 months. The first period ran from September 2011 to April 2012, and the second one ran from June 2012 to April 2013. We recruited forecasters from professional societies, research centers, alumni associations, and science blogs, as well as word of mouth. Entry into the tournament required a bachelor's degree or higher and completion of a battery of psychological and political knowledge tests that took approximately 2 hr. Participants were largely U.S. citizens (76%) and males (83%), with an average age of 36. Almost two thirds (64%) had some postgraduate training.

Design

In Year 1, participants were randomly assigned to a 3 × 3 factorial design of Training (probabilistic-reasoning training, scenario training, and no training) × Group Influence (independent forecasters, crowd-belief forecasters, and team forecasters).[1] Training consisted of instructional modules. Probabilistic-reasoning training, consisted of tips about what information to look for and how to avoid judgmental biases. Scenario training taught forecasters to generate new futures, actively entertain more possibilities, use decision trees, and avoid overconfidence.

Group influence had three levels. We staked out a continuum with independent forecasters who worked alone at one end, and interdependent forecasters who worked in teams of approximately 15 people and interacted on a website at the other end. We also included a compromise level (crowd belief forecasters) in which forecasters worked alone, but had knowledge of others' beliefs. The benefit of this approach is that forecasters had access to a potentially potent signal—the numerical distribution of the crowd's opinions, but they could avoid the potential costs of social interaction, such as mindless "herding" or free-riding.

Those in team conditions also received training in how to create a well-functioning group. Members were encouraged to maintain high standards of proof and seek out high-quality information. They were given strategies for explaining their forecasts to others, offering constructive critiques, and building an effective team. Members could offer rationales for their thinking and critiques of others' thinking. They could share information, including their forecasts. But there was no systematic display of team members' predictions. Instructions, training, tests, and forecasting questions are available in the online supplemental materials.

At the end of Year 1, we learned that probabilistic-reasoning training was the most effective instructional module, and teamwork was the most effective form of group interaction. We decided to replicate only the most effective experimental methods from Year 1 using a reduced 2 × 2 factorial design of Training (probabilistic-reasoning training and no training) by Group Influence (team vs. independent forecasters).

We added another intervention—the tracking of top performers. We skimmed off the top 2% of forecasters and put them in five elite teams. This small group of forecasters had a distinctly different experience from others (see Mellers, Stone, et al., 2014) and therefore was not included in the analyses. However, results did not change when these forecasters were included.

[1] Results from the prediction market are discussed elsewhere because individual accuracy measures (in the form of Brier scores) cannot be computed (Atanasov et al., 2014).

The smaller experimental design of Year 2 required reassignment of forecasters from Year 1 conditions that were not continued in Year 2 (i.e., crowd belief forecasters and scenario training). Assignment of forecasters proceeded as follows: (a) if a Year 1 condition remained in Year 2, forecasters stayed in that condition; (b) crowd-belief forecasters were randomly assigned to independent or team conditions; (c) scenario trainees were randomly assigned to no training or probabilistic-reasoning training. Our analyses in this article focus only on the effectiveness of two situational variables: probabilistic-reasoning training and teamwork.

Questions

Questions were released throughout the tournament in small batches, and forecasters received 199 questions over 2 years. Questions covered political-economic issues around the world and were selected by the IARPA, not by the research team. Questions covered topics ranging from whether North Korea would test a nuclear device between January 9, 2012, and April 1, 2012, to whether Moody's would downgrade the sovereign debt rating of Greece between October 3, 2011, and November 30, 2011. Questions were open for an average of 109 days (range = 2 to 418 days).

Participants were free to answer any questions they wished within a season. There were no constraints on how many, except that payment for the season required that participants provide forecasts for at least 25 questions. One question asked, "Will there be a significant outbreak of H5N1 in China in 2012?" The word "significant" was defined as at least 20 infected individuals and five casualties. The question was launched on February 21, 2012, and was scheduled to close on December 30, 2012. If the outcome occurred prior to December 30, the question closed when the outcome occurred. Forecasters could enter their initial forecast or update their prior forecast until the resolution of the outcome.

One hundred fifty questions were binary. One binary question, released on November 8, 2011, asked, "Will Bashar al-Assad remain President of Syria through January, 31 2012?" Answers were "yes" or "no." Some questions had three to five outcomes. A three-option question, released on October 4, 2011, asked, "Who will win the January 2012 Taiwan Presidential election?" Answers were "Ma Ying-jeou," "Tsai Ing-wen," or "neither." Some questions had ordered outcomes. One with four ordered outcomes asked, "When will Nouri al-Maliki resign, lose confidence vote, or vacate the office of Prime Minister of Iraq?" Answers were "between July 16, 2012 and Sept 30, 2012," "between Oct 1, 2012 and Dec, 31 2012," between "Jan 1, 2013 and Mar 31, 2013," or "the event will not occur before April 1, 2013." Finally, another set was conditional questions, typically having two antecedents and two outcomes. One question asked,

> Before March 1, 2014, will North Korea conduct another successful nuclear detonation (a) if the United Nations committee established pursuant to Security Council resolution 1718 adds any further names to its list of designated persons or entities beforehand or (b) if the United Nations committee established pursuant to Security Council resolution 1718 *does not* add any further names to its list of designated persons or entities beforehand?

Answers to both possibilities were "yes" or "no."

All forecasters were given a brief Brier score tutorial and learned that their overarching goal was to minimize Brier scores.

Feedback given to forecasters during the tournament included Brier scores, averaged over days within a question and across questions. Forecasters were incentivized to answer questions if they believed they knew more than the average forecaster in their condition. If they did not answer a question, they received the average Brier score that others in their condition received on that question. Whenever a question closed, we recalculated individual Brier scores, thereby providing forecasters with constant feedback.

Brier scores used in our analyses *did not* include the average scores of others if a forecaster did not answer a question. Instead, we simply computed the Brier score for each forecast made by a participant and averaged over Brier scores if that participant made multiple forecasts on a given question. Inclusion of averages from others would simply have reduced differences among individuals.

Measures

Prior to each forecasting season, we administered a battery of psychological tests. Intelligence was measured by four scales. Inductive pattern recognition was assessed by a short form of the Ravens Advanced Progressive Matrices (Ravens APM; Bors & Stokes, 1998), which circumvents cultural or linguistic knowledge by testing spatial problem-solving skills. Cognitive control was measured by the three-item Cognitive Reflection Test (CRT; Frederick, 2005) and the four-item extension of the CRT (Baron, Scott, Fincher, & Metz, 2014), with questions such as, "All flowers have petals. Roses have petals. If these two statements are true can we conclude that roses are flowers?" Mathematical reasoning was measured by a three-item Numeracy scale. The first item came from Lipkus, Samsa, and Rimer (2001), and the second two were from Peters et al. (2006).

We had three measures of open-mindedness. The first was a seven-item actively open-minded thinking test (Haran et al., 2013) that used a 7-point rating scale (1 = *completely disagree* and 7 = *completely agree*). Actively open-minded thinking predicts both persistence in information searches and accuracy in estimating uncertain quantities (Haran et al., 2013). The second was an 11-item Need-For-Closure scale (Kruglanski & Webster, 1996; Webster & Kruglanski, 1994). Responses were made on the same 7-point rating scale. The third was a single question: "In a famous essay, Isaiah Berlin classified thinkers as hedgehogs and foxes: The hedgehog knows one big thing and tries to explain as much as possible using that theory or framework. The fox knows many small things and is content to improvise explanations on a case-by-case basis. When it comes to making predictions, would you describe yourself as more of a hedgehog or more of a fox?" Responses were made on a 5-point rating scale (1 = *very much more fox-like*; 5 = *very much more hedgehog-like*).

Political knowledge was assessed by two true–false tests of current affairs, one given each year. The first was a 35-item test with items such as "Azerbaijan and Armenia have formally settled their border dispute." The second was a 50-item test with items such as "India's liberalization reforms now allow for 100% Foreign Direct Investment (FDI) stake in ventures" or "The GINI coefficient measures the rate of economic expansion."

Participants

Year 1 began with 1,593 survey participants who were randomly assigned to nine conditions, with an average of 177 per condition. Attrition was 7%. Year 2 started with 943 respondents. Attrition in Year 2 fell to 3%, perhaps because most participants were returnees and familiar with the task. We wanted forecasters who made many predictions and for whom we could get stable estimates of forecasting ability. To that end, we used only 743 forecasters who participated in both years of the tournament and had made at least 30 predictions.

Payments

Forecasters who met the minimum participation requirements received $150 at the end of Year 1 and $250 at the end of Year 2, regardless of their accuracy. Those who returned from Year 1 received a $100 retention bonus. Forecasters also received status rewards for their performance via leader boards that displayed Brier scores for the top 20 forecasters (10%) in each condition and full Brier score rankings of teams. Team Brier scores were the median of scores for individuals within a team.

Results

Individual Differences and Consistency Over Time

Our primary goal was to investigate the consistency and predictability of individual forecasting accuracy, defined as a Brier score averaged over days and questions. Participants attempted an average of 121 forecasting questions. Figure 1 shows the distribution of overall Brier scores, revealing a wide range of forecasting abilities.

A common approach in studies of accuracy is to compare intuitive predictions with simple benchmarks, such as the score one would receive by assigning a uniform distribution over outcomes for all questions. The raw Brier score would be 0.53, on a scale ranging from 0 (best) to 2 (worst). The average raw Brier score of our participants was 0.30, much better than random guessing, $t(741) = -61.79$, $p < .001$. Overall, forecasters were significantly better than chance.

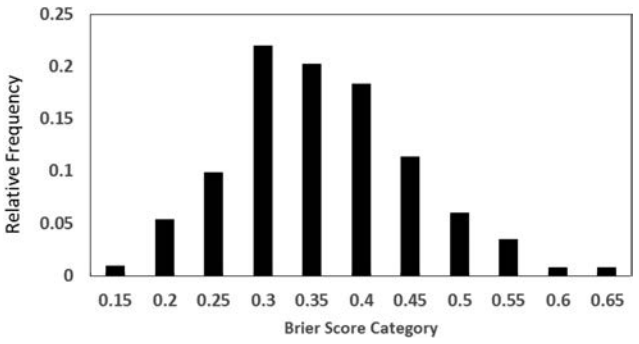

Figure 1. Distribution of Brier scores over forecasters plotted against category bins of size .049. The category labeled .15 refers to Brier scores between .10 and .149.

An alternative measure of forecast accuracy is the proportion of days on which forecasters' estimates were on the correct side of 50%. This measure is calculated by counting the days on which forecasters were active and correct (i.e., they placed estimates of 51% or above for events that occurred and 49% or below for events that did not occur). For multinomial questions, forecasts were considered correct if the realized option was associated with the highest probability. We counted all days after the first forecast was placed, and we carried forward estimates until a participant updated his or her forecast or the question closed. A perfect score would be 100%, and a chance score for binary questions would be 50%. For all questions in the sample, a chance score was 47%. The mean proportion of days with correct estimates was 75%, significantly better than random guessing for binary questions, $t(740) = 79.70$, $p < .001$.

Figure 2 shows the distribution. The correlation between mean Brier score and proportion of correct days was very high, $r = .89$, $t(741) > 54.25$, $p < .0001$. The proportion of correct days is just another way to illustrate accuracy. All subsequent analyses focus on Brier scores, unless otherwise specified.

Next we turn to the question of consistency, but first we make a simple adjustment to the accuracy metric. Forecasters selected their own questions, and this feature of the experimental design allowed people to get a better Brier score if they could select events that were easier to predict. To handle this problem, we standardized Brier scores within questions. Standardization minimizes differences in difficulty across questions and allowed us to focus on relative, rather than absolute, performance. If accuracy were largely attributable to luck, there would be little internal consistency in Brier scores over questions. However, Cronbach's alpha (a gauge of the internal consistency of Brier scores on questions) was 0.88, suggesting high internal consistency. Figure 3 illustrates how the best and worst forecasters differed in skill across time. We constructed two groups based on average standardized Brier scores after the first 25 questions had closed and forecasters had attempted an average of 15 questions. The black and gray lines represent the 100 best and worst forecasters, respectively. Figure 3 tracks their progress over time; average Brier scores are presented for each group on 26th to the 199th question, plotted against the order that questions closed.[2] Using relatively little initial knowledge about forecaster skill, we could identify differences in performance that continued for a period of 2 years. These groups differed by an average of 0.54—more than half a standard deviation—across the tournament. If we could identify good forecasters early, there was a reasonable chance they would be good later.

There are several ways to look for individual consistency across questions. We sorted questions on the basis of response format (binary, multinomial, conditional, ordered), region (Eurzone, Latin America, China, etc.), and duration of question (short, medium, and long). We computed accuracy scores for each individual on each variable within each set (e.g., binary, multinomial, conditional, and ordered) and then constructed correlation matrices. For all three question types, correlations were positive; an individual who scored high on binary questions tended to score higher on

[2] For each forecaster, we averaged predictions over days, regardless of the day on which the prediction was made.

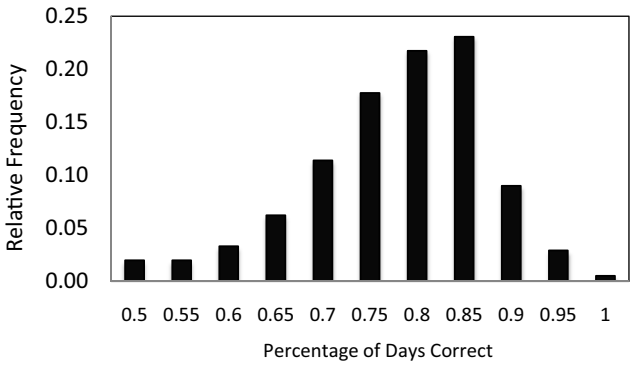

Figure 2. Distribution of days on which estimates were on the correct side of 50% plotted against bins of size .049. The category labeled 0.55 refers to forecasters who were correct for 55% to 59.9% of the days on which they had active forecasts.

multinomial questions. Then we conducted factor analyses. For each question type, a large proportion of the variance was captured by a single factor, consistent with the hypothesis that one underlying dimension was necessary to capture correlations among response formats, regions, and question duration.

Dispositional Variables

Are individual dispositional variables of intelligence, open-mindedness, and political knowledge associated with forecasting accuracy? Table 1 shows means and variances of predictor variables. The mean score on the short version of the Ravens APM was 8.56 out of 12, which was considerably higher than 7.07, the mean score of a random sample of first-year students at the University of Toronto (Bors & Stokes, 1998). Our forecasters scored 2.10 on the CRT, virtually equivalent to 2.18, the average score of MIT students (Frederick, 2005). The extended CRT and the numeracy items had no comparable norms.

Table 1 also presents reliability estimates, when applicable. Values were .71 for the Ravens APM, .55 for the three-item CRT, .70 for the extended CRT, .11 for Numeracy, .65 for actively open-minded thinking, .55 for Need for Closure, .53 for the Year 1 political knowledge test, and .64 for the Year 2 test. The most troubling reliability estimate was that of Numeracy. Most people found it very easy; the percentages correct on the three items were 93%, 92%, and 86%.

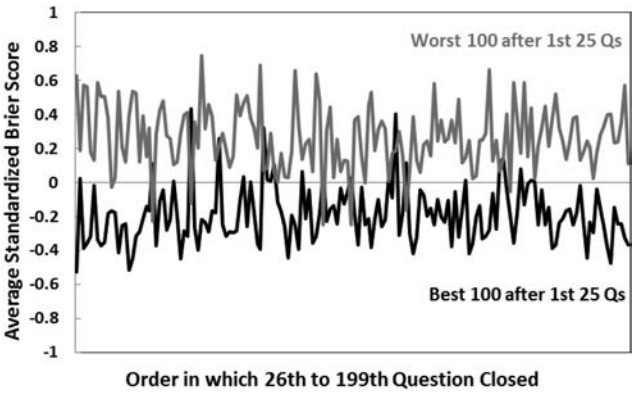

Figure 3. Average scores for 100 best forecasters (black) and 100 worst forecasters (gray) defined after the close of the first 25 questions. The *x*-axis represents the order in which questions closed throughout the rest of the tournament. Differences defined early in the tournament remained for 2 years, as seen by the space between the two lines.

Table 1
Descriptive Statistics

	Mean	*SD*	Min	Max	Alpha
Standardized Brier score	0	0.29	−0.57	1.32	0.88
Ravens	8.56	2.43	0	12	0.71
Cog Reflection Test	2.1	0.97	0	3	0.55
Extended Cog Reflection Test	3.37	1.03	0	4	0.70
Numeracy	2.71	0.53	0	3	0.11
Actively open-minded think	5.91	0.6	4	7	0.65
Need for closure	3.34	0.58	1.45	5.09	0.55
Fox vs. Hedgehog	3.82	0.54	1.9	6	
Political knowledge Year 1	28.79	3.07	18	35	0.53
Political knowledge Year 2	36.5	4.64	19	48	0.64
Number predictions per Q	1.58	0.77	1	6.33	
Number of questions	21	51	13	199	
Deliberation time (s)	3.6	0.71	2	4	

Note. Min = minimum; Max = maximum; SD = signaled donation; Cog = cognition.
[a] Cronbach's alpha for Brier scores is calculated at the question level. All other alphas are calculated at the participant level. Alphas are not reported for scales with three or fewer items and for behavioral variables.

Table 2 shows all possible pairwise correlations. Three of the four measures of intelligence—the Ravens APM, the CRT, and the extended CRT—were significantly correlated with standardized Brier score accuracy: Correlations were −.23, −.15, and −.14, respectively, $t(741) = -6.38$, $p < .001$, $t(741) = -4.17$, $p < .001$, and $t(599) = -3.56$, $p = .001$.[3] Lower Brier scores indicate higher accuracy, so negative correlations mean that greater accuracy is associated with higher intelligence scores. We combined these variables into a single factor using the first dimension of a principal axis factor analysis. The correlation between standardized Brier scores and factor scores was −.22, $t(741) = -5.51$, $p < .001$. Greater intelligence predicted better performance, consistent with our first hypothesis.

Next we turn to open-mindedness and examined whether three measures—actively open-minded thinking, need for closure, and hedgehog–fox orientation—predicted forecasting accuracy. The average score on actively open-mindedness was 5.91, relatively high on a 1 to 7 response scale. The average need for closure score was 3.34, close to the middle of the scale, and the average fox–hedgehog response was 3.82, which indicated that, on average, forecasters viewed themselves as slightly more hedgehog-like. More actively open-minded participants had less need for closure, $r = -.20$, $t(742) = -5.56$, $p < .001$, and more hedgehog-like participants had more need for closure, $r = .24$, $t(742) = -6.73$, $p < .001$. Only one of the measures, actively open-minded thinking, was significantly related to standardized

[3] The correlation between reaction time on the Ravens APM test and forecasting accuracy was also significant; those who spent more time on the Ravens APM test also tended to be better forecasters, $r = -0.12$, $t(741) = 3.29$, $p < .001$.

Table 2
Correlations Among Dispositional, Situational, and Behavioral Variables

	Std BS	Ravens	CRT	ExCRT	Numeracy	AOMT	Nfclo	Foxhed	PKY1	PKY2	Train	Teams	Npredq	Nquest
Std BS	1.00													
Ravens	**−0.23**	1.00												
CRT	**−0.15**	**0.38**	1.00											
ExCRT	**−0.14**	**0.34**	**0.39**	1.00										
Numeracy	−0.09	**0.16**	**0.12**	**0.14**	1.00									
AOMT	**−0.10**	**0.10**	0.08	**0.22**	0.09	1.00								
Nfclo	0.03	0.03	−0.02	−0.05	0.10	**−0.20**	1.00							
Foxhed	0.09	0.05	0.01	0.02	0.02	−0.09	**0.24**	1.00						
PKY1	**−0.18**	0.05	0.06	0.08	0.03	**0.13**	−0.03	−0.03	1.00					
PKY2	**−0.20**	0.08	0.08	0.12	0.01	**0.12**	−0.07	−0.09	**0.59**	1.00				
Train	**−0.17**	0.02	−0.01	0.06	0.06	0.05	−0.03	0.02	0.04	0.02	1.00			
Teams	**−0.30**	−0.04	0.01	0.01	0.04	0.04	−0.05	−0.06	0.02	0.02	0.00	1.00		
Npredq	**−0.49**	**0.17**	**0.12**	**0.12**	0.09	0.05	0.01	−0.02	**0.14**	**0.19**	0.08	**0.11**	1.00	
Nquest	0.07	−0.02	0.04	0.04	−0.05	−0.02	0.06	0.07	0.07	0.07	−0.02	**−0.17**	**0.23**	1.00
Del time	**−0.30**	0.08	−0.09	−0.05	0.03	0.05	−0.09	−0.08	−0.01	0.05	0.06	**0.28**	**0.15**	**−0.25**

Note. Bold values are significant at the .001 level. Std BS = Standardized Brier score; CRT = cognitive reflection test; ExCRT = extended cognitive reflection test; AOMT = actively open-minded thinking; Nfclo = need for closure; Foxhed = fox versus hedgehog; PKY1 = political knowledge year 1; PKY2 = political knowledge year 2; train = training; Npredq = number of predictions per question; Nonquest = number of questions answered.

Brier score accuracy, $r = -.10$, $t(742) = -2.51$, $p < .01$. Thus, we had partial support for the second hypothesis that flexible and open-minded cognitive styles predicted forecasting accuracy.

The third hypothesis stated that political knowledge would predict Brier score accuracy. Percent correct scores on these true–false questions were 82% and 76%, respectively. Test scores were highly correlated with forecasting accuracy, $r = .59$, $t(742) = -19.91$, $p < .001$. We have no comparable norms, but the obvious difficulty of the tests makes these scores seem high. The correlation between political knowledge scores in Years 1 and 2 and relative forecasting accuracy was −.18 and −.20, respectively, $t(741) = -4.85$, $p < .001$, and $t(648) = -5.06$, $p < .001$. Again, we constructed a single measure of content knowledge using the first factor of a principal axis factor analysis. The correlation between relative accuracy and these factor scores was −.22, $t(599) = -5.52$, $p < .001$. Political knowledge was predictive of forecasting accuracy, consistent with our third hypothesis.

Earlier, we mentioned the debate about the role of intelligence versus deliberative practice in the development of expertise. One hypothesis was that the correlation between intelligence and performance would be strongest early on and gradually disappear as forecasters engage in more deliberate practice. In past studies, the Ravens APM is a common measure of cognitive ability (e.g., Ruthsatz, Detterman, Griscom, & Cirullo, 2008). We correlated Ravens APM scores with accuracy early and late in the tournament. "Early" and "late" are vague terms, so we used multiple definitions, including the first 50 and last 50 questions, the first 40 and last 40 questions, and the first 30 and last 30 questions (out of 199).

Correlations between the Ravens APM scores and accuracy based on the first and last 50 questions representing early and late stages were −.22 and −.10, respectively. The relationship between intelligence and performance was stronger earlier than it was later, $t(624) = 2.57$, $p < .01$. Similar results occurred with cutoffs of 30 and 40 questions. This analysis is based on only 2 years of deliberative practice, not 10,000 hr (i.e., the length of deliberative practice necessary to achieve expertise, according to Ericsson et al., 1993). Nonetheless, the difficulty of the questions and the breadth of topics suggest that one would do poorly in our tournament without some degree of sustained effort and engagement.

Situational Variables

Mellers, Ungar, et al. (2014) showed that forecasters who were trained in probabilistic reasoning and worked together in teams were more accurate than others. However, effect sizes in the form of correlations were not presented. Table 2 shows that relative accuracy and training in probabilistic reasoning had a correlation of −.17, $t(741) = -4.56$, $p < .001$. In addition, relative accuracy of team participation had a correlation of −.30, $t(741) = -8.55$, $p < .001$. These findings illustrate how the prediction environment influences forecaster accuracy, independent of all else.

To test the fourth hypothesis—dispositional variables predict forecasting skill beyond situational variables—we conducted a multiple regression predicting standardized Brier scores from intelligence factor scores, actively open-minded thinking, political knowledge factor scores, probability training, and teamwork. The latter two variables were dummy coded. The multiple correlation was .43, $F(5, 587) = 26.13$, $p < .001$. Standardized regression coefficients for two of the three dispositional variables—intelligence and political knowledge—were statistically significant, −0.18 and −0.15, $t(587) = -4.65$, $p < .001$, and $t(587) = -3.89$, $p < .001$. Intelligence and political knowledge added to the prediction of accuracy beyond the situational variables.

Behavioral Variables

Effort and engagement manifest themselves in several ways, including the number of predictions made per question (belief updating), the time spent before making a forecast (deliberation time), and the number of forecasting questions attempted. The

average number of predictions made per forecasting question was 1.58, or slightly more than 1.5 forecasts per person per question. Deliberation time, which was only measured in Year 2, was transformed by a logarithmic function (to reduce tail effects) and averaged over questions. The average length of deliberation time was 3.60 min, and the average number of questions tried throughout the 2-year period was 121 out of 199 (61% of all questions). Correlations between standardized Brier score accuracy and effort were statistically significant for belief updating, −.49, $t(740) = -15.29$, $p < .001$, and deliberation time, −.30, $t(694) = -8{,}28$, $p < .001$, but not for number of forecasting questions attempted. Thus, two of three behavioral variables predicted accuracy, in partial support of the fourth hypothesis.

The fifth hypothesis stated that behavioral variables would contribute to the predictability of skill over and beyond dispositional and situational variables. To test this hypothesis, we conducted a multiple regression predicting standardized Brier scores from belief updating, deliberation time, and number of questions attempted, as well as intelligence factor scores, actively open-minded thinking, political knowledge factor scores, training, and teamwork. The multiple correlation was .64, $F(8, 581) = 52.24$, $p < .001$. Behavioral variables with significant standardized regression coefficients were belief updating, −0.45, $t(581) = -12.89$, $p < .001$, and deliberation time, −0.13, $t(581) = -3.69$, $p < .001$. Results were thus consistent with the fifth hypothesis that behavioral variables provide valuable independent information, in addition to dispositional and situational variables.

Structural Equation Model

To further explore interconnections among these variables, we used a structural equation model that enabled us to synthesize our results in a single analysis, incorporate latent variables, and perform simultaneous regressions to test our hypotheses. The initial model only included variables that were significant predictors of standardized Brier score accuracy on their own. These variables included two latent constructs (political knowledge and intelligence), open-mindedness, probabilistic training, teamwork, belief updating, and deliberation time.

We then conducted mediation analyses to see whether behavioral variables mediated the relationship between dispositional variables and accuracy, and situational variables and accuracy.[4] Results are shown in Table 3. For simplicity, we removed pathways whose inclusion neither improved nor changed the model fit significantly, and ultimately we arrived at the model in Figure 4. Yellow ovals are latent dispositional variables, yellow rectangles are observed dispositional variables, pink rectangles are experimentally manipulated situational variables, and green rectangles are observed behavioral variables.

Dispositional variables of political knowledge and intelligence had direct and indirect effects on Brier score accuracy. Better forecasters had greater knowledge and ability, and part of that relationship was accounted for by greater belief updating. Actively open-minded thinking, our best cognitive-style predictor, had only direct effects on accuracy. Situational variables of teamwork and training had direct effects on accuracy, but teamwork also had indirect effects. Those in teams did better than those who worked alone, especially when they updated their beliefs often and deliberated longer.

The structural equation modeling required the fit of five simultaneous regressions shown in Table 4. In one regression, the latent variable of fluid intelligence was predicted from the Ravens APM, the CRT, the extended CRT, and Numeracy. The coefficient for the Ravens APM was set to 1.0, and others were estimated relative to it. In the next regression, the latent variable of political knowledge was predicted from tests in Years 1 and 2. In the third regression, belief updating was predicted by the two latent variables and teamwork, and in the fourth, deliberation time was predicted from teamwork. The last regression was the prediction of forecaster accuracy from fluid intelligence, political knowledge, intelligence, actively open-minded thinking, teams, probability training, belief updating, and deliberation time. Coefficients for these regressions with observed variables, along with standard errors, Z statistics, p values, and bootstrapped confidence intervals (when appropriate) are provided in Table 4.

This model provided a reasonable account of relative forecaster accuracy. The Tucker-Lewis Reliability Index was 0.92, and the comparative fit index was 0.95. The root mean squared error of approximation was 0.04. In addition to fitting a model to individuals' average relative accuracy scores, we fit models to their first and last forecasts. Using only first forecasts, the effect of belief updating disappeared, as expected, but the remaining associations remained strong. Using only last forecasts, the effect of belief updating increased, as we would expect, and all other associations did not change.

One way to test the model's stability is to calculate confidence intervals on coefficients using bootstrapping methods. These intervals are presented in Table 4, and all of these exclude zero, supporting the validity of the relationships. Another way to test the model's stability is to conduct cross validations which examine the extent to which the model would have fit with different subsets of questions or of participants. For the cross-validation of questions, we randomly assigned each question to one of 10 folds (or subsets of questions) with equal numbers of questions in each fold. In each validation, we used 90% of the questions, computed average standardized Brier scores for each participant, and refit the structural model. We repeated this process for each fold and examined the distributions of resulting coefficients over all 10 validations. Coefficients for all of the parameters were consistent with the full model.[5] We conducted a similar cross-validation analysis using subsets of participants, and again, coefficients for all of the parameters were consistent with the full model in each of the 10 validation sets.

Which Types of Variables Best Predict Relative Accuracy?

Table 5 shows multiple correlations and tests of nested comparisons. Using only dispositional information (intelligence, political

[4] We conducted mediation analyses to determine which of our predictors of accuracy might be mediated by effort. We used those results to determine which pathways to include in the structural equation model. Results of the mediation analyses are summarized in Table 3 and Figure 3.

[5] Consistency refers to all coefficients in the validation models maintaining significance ($p \leq .05$), and similar magnitude to the full model, across all 10 cross-validation folds.

Table 3
Indirect and Total Contributions in Mediation Analyses

Independent	Mediator	Dependent	Indirect	*p* value	Total	*p* value
IQ	Bel updating	Std Br score	−0.18	<0.001	−0.54	<0.001
Pol know	Bel updating	Std Br score	−0.16	<0.001	−0.36	<0.001
AOMT	Bel updating	Std Br score	−0.03	0.10	−0.12	<0.001
Train	Bel updating	Std Br score	−0.07	0.03	−0.32	<0.001
Team	Bel updating	Std Br score	−0.08	0.01	−0.53	<0.001
Team	Deliberation time	Std Br score	−0.07	<0.001	−0.29	<0.001

Note. Independent refers to the independent variable, and Dependent is the dependent variable. Indirect is the product of the correlation between the independent variable and the mediator and that between the mediator and the dependent variable. Total is the indirect plus the direct effects, where the direct effect is the correlation between the independent and dependent variable. IQ = intelligence; Pol know = political knowledge; AOMT = actively open-minded thinking; Train = probablistic reasoning training; Teams = on a collaborative team.

knowledge, and actively open-minded thinking), the multiple correlation for accuracy was .31, close to the famous .3 barrier in personality research, which is sometimes supposed to be the upper bound of the validities on many personality scales. Using only situational variables that describe the conditions under which forecasters worked, the multiple correlation was .34, similar in magnitude to that obtained from the dispositional variables, replicating the more general conclusion of Funder and Ozer (1983) that many individual difference effects and situational manipulations appear to have similar effect-size upper bounds.

Adding behavioral information on forecaster belief updating and deliberation time, predictability improved and the multiple correlation rose to .54. Not surprisingly, it is harder to identify the best forecasters from abstract dispositional constructs than it is from specific features of their behavior in the prediction environment. Nonetheless, as we saw in the structural model, and confirm here, the best model uses dispositional, situational, and behavioral variables. The combination produced a multiple correlation of .64. Each model provided a better fit as more variables were included. *F* tests for the nested model deviance showed that larger models provided a significantly better fit than their simpler counterparts. The person, the situation, and related behavior all contribute to identifying superior geopolitical forecasting performance.[6]

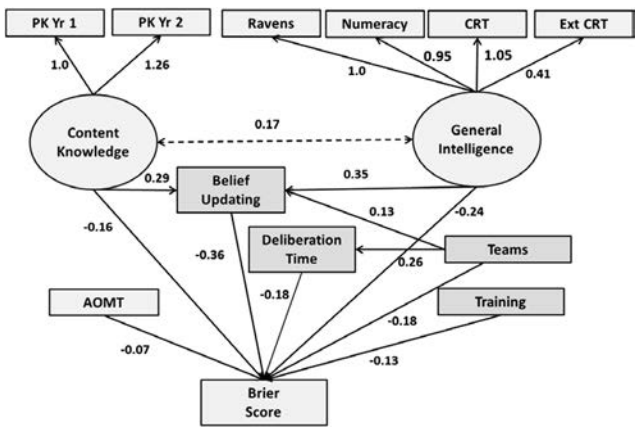

Figure 4. Structural equation model with standardized coefficients. See the online article for the color version of this figure.

Discussion

We examine the geopolitical forecasting skills of participants in a 2-year tournament. Drawing on diverse literatures, we tested three categories of hypotheses about the psychological drivers of judgmental accuracy: (a) dispositional variables measured prior to the tournament, such as cognitive abilities and styles; (b) situational variables that were experimentally manipulated prior to the competition, such as training in probabilistic reasoning and collaborative engagement in teams organized around self-critical, epistemic norms; and (c) behavioral variables measured during the competition, such as the willingness to revisit and update older beliefs.

The best forecasters scored higher on both intelligence and political knowledge than the already well-above-average group of forecasters. The best forecasters had more open-minded cognitive styles. They benefited from better working environments with probability training and collaborative teams. And while making predictions, they spent more time deliberating and updating their forecasts.

These predictors of accuracy proved robust over several subsets of questions. With few exceptions, variables that captured the best forecasters overall predicted accuracy across different temporal periods within a question (early, middle, and late), across questions that differed in length (short, medium, and long durations), and across questions that differed in mutability (close calls vs. clear-cut outcomes).

We offer a structural equation model to capture the interrelationships among variables. Measures that reflected behavior within tournaments served as mediators. Belief updating partly mediated the relationship between intelligence and accuracy, between political knowledge and accuracy, and teamwork and accuracy. Deliberation time mediated the relationship between teamwork and accuracy. This association has different causal interpretations. Those with more political knowledge and greater intelligence might have enjoyed the task more—and that enjoyment may have motivated engagement. Alternatively, once forecasters became more engaged, they may have become more politically knowledgeable. Furthermore, those who worked in teams may also have been

[6] These correlations were fit directly to the data. Cross-validated correlations would obviously be smaller.

Table 4
Regressions in Structural Equation Model

Regressions	Estimate	Std error	*z*	*p*	Bootstrap CI	
1. Intelligence (FS)						
CRT	1.01	0.12	8.73	0.00		
Ex-CRT	0.89	0.10	8.63	0.00		
Numeracy	0.35	0.08	4.17	0.00		
2. Pol knowledge (FS)						
Pol know Year 2	1.22	0.22	5.49	0.00		
3. Belief updating						
Teams	0.12	0.04	2.91	0.00	0.04	0.21
Pol know	0.30	0.08	3.97	0.00	0.15	0.44
Intelligence	0.34	0.09	3.86	0.00	0.17	0.51
4. Deliberation time						
Teams	0.27	0.04	7.18	0.00	0.20	0.34
5. Std Brier score						
Belief updating	−0.35	0.03	−10.72	0.00	−0.42	−0.29
Deliberation time	−0.17	0.04	−4.82	0.00	−0.24	−0.10
Pol know	−0.16	0.06	−2.80	0.01	−0.28	−0.05
Intelligence	−0.24	0.07	−3.54	0.00	−0.38	−0.11
Teams	−0.19	0.03	−5.65	0.00	−0.26	−0.12
P train	−0.13	0.03	−4.15	0.00	−0.20	−0.07
AOMT	−0.07	0.03	−2.13	0.03	−0.13	−0.01

Note. FS refers to factor score. The regression coefficients for Ravens and the political knowledge test Year 1 were set to 1.0 in the first and second regression, respectively. CRT = cognitive reflection test; ExCRT = extended cognitive reflection test; Pol know = political knowledge; P train = probablistic reasoning training; AOMT = actively open-minded thinking.

motivated to do well for the sake of the group, which could also produce greater updating and ultimately greater accuracy.

Actively open-minded thinking predicted accuracy but less consistently than other variables. This cognitive style is associated fewer cognitive errors, including the myside bias, biased argument evaluation, superstitious thinking, and outcome bias (Stanovich & West, 1997, 1998, 2007). Laboratory evidence shows that actively open-minded thinking predicts accuracy of estimates of uncertain

Table 5
Predicting Overall Forecasting Accuracy From Different Types of Variables

	Multiple *R*	*F*	Sig
Variable type			
Dispositional (2 latent, 1 observed variable)	0.31	21.12	$p < .001$
Situational (2 variables)	0.34	49.44	$p < .001$
Behavioral (2 variables)	0.54	142.52	$p < .001$
Dis + Sit (2 latent, 3 observed variables)	0.45	30.17	$p < .001$
Dis + Beh (2 latent, 3 variables)	0.58	60.06	$p < .001$
Sit + Beh (4 variables)	0.58	89.43	$p < .001$
Dis + Sit + Beh (7 variables)	0.64	52.99	$p < .001$

	Delta SS	Delta DF	*F*	Sig
Nested comparisons				
Dis vs. Dis + Sit	64.76	2	40.38	$p < .001$
Sit vs. Dis + Sit	56.16	7	10.01	$p < .001$
Dis vs. Dis + Beh	154.51	3	79.41	$p < .001$
Beh vs. Dis + Beh	22.75	7	5.01	$p < .001$
Sit vs. Beh + Sit	148.88	3	77.79	$p < .001$
Beh vs. Sit + Beh	25.72	2	25.72	$p < .001$
Dis + Sit vs. Dis + Sit + Beh	115.51	3	63.51	$p < .001$
Dis + Beh vs. Dis + Sit + Beh	25.77	2	21.25	$p < .001$
Sit + Beh vs. Dis + Sit + Beh	22.80	7	5.37	$p < .001$

Note. Dispositional variables refer to principle factor scores for general intelligence (Ravens, CRT, exCRT, numeracy) and political knowledge (Year 1 and Year 2 tests), as well as actively open-minded thinking. Situational variables refer to teams and training. Behavioral variables are the average number of forecasts made per question and average time spent deliberating about a question. The *F* test for nested comparisons tests the probability that the difference in sum of squares between the smaller and larger models is >0 by comparing the mean square error of the smaller model to the residual sum of squares for the larger one.

quantities (Haran et al., 2013), but no prior studies have demonstrated an association between actively open-minded thinking and forecasting performance on real-world problems. We believe this cognitive style translates into more accurate political forecasts through its association with better norms of thinking.

Kahneman and Klein (2009) argued that for any type of skill to develop, two conditions must be present: (a) an environment with sufficient deterministic stability to permit learning, and (b) opportunities for practice. Skill development occurs to the extent that people care enough to engage in deliberative rehearsal (Ericsson, 2006). Our forecasters received constant feedback in the form of Brier scores and leaderboard rankings. They had many chances to learn; there were 199 questions over a period of 2 years, and, on average, forecasters made predictions for 121 of them. These conditions enabled a process of learning-by-doing and help to explain why some forecasters achieved far-better-than-chance accuracy.

In the real world, intelligence analysts use data from diverse sources. They frequently make nonnumerical forecasts that are vague and hard to score for accuracy, so feedback is often absent. Intelligence analysts shift their response thresholds depending on the cost of the errors. That is, they are likelier to say "signal" when the costs of a miss are high, and they are likelier to say "noise" when the costs of a false alarm are high. Although our forecasters knew that the Brier-score costs of errors were symmetric, the real world is much more complicated.

Analysts also operate under bureaucratic-political pressure—and are tempted to respond to previous mistakes by shifting their response thresholds. They are likelier to say "signal" when recently accused of underconnecting the dots (i.e., 9/11) and to say "noise" when recently accused of overconnecting the dots (i.e., weapons of mass destruction in Iraq). Tetlock and Mellers (2011) describe this process as accountability ping-pong. One escape from this cycle is to make a public organizational commitment to using tournaments to monitor long-term accuracy, not just avoidance of the most recent mistake (McGraw, Todorov, & Kunreuther, 2011).

Our study is the first to keep score and track categories of variables that predict performance in the politically sensitive domain of intelligence analysis. The study demonstrates the value of tournaments in identifying top forecasters. If the National Academy of Science Report on improving intelligence analysis is correct about the power of measuring accuracy and providing feedback to boost performance, tournaments should become a regular feature of research on improving accuracy in organizational systems and evaluating the track records of intelligence analysts.

References

Almond, G. A., & Genco, S. J. (1977). Clouds, clocks, and the study of politics. *World Politics, 29,* 489–522. http://dx.doi.org/10.2307/2010037

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In S. Armstrong (Ed.), *Principles of forecasting* (pp. 495–515). New York, NY: Springer.

Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2014). *Crowd forecasting with prediction polls and prediction markets*. Manuscript submitted for publication.

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 & Cognition*. Advance online publication. http://dx.doi.org/10.1016/j.jarmar.2014.09.003

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58,* 382–398. http://dx.doi.org/10.1177/0013164498058003002

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Bueno de Mesquita, B. (2009). *The predictioneer's game: Using the logic of brazen self-interest to see and shape the future*. New York, NY: Random House.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22. http://dx.doi.org/10.1037/h0046743

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. http://dx.doi.org/10.1111/j.1745-3984.1978.tb00065.x

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25,* 331–351. http://dx.doi.org/10.1002/bdm.731

Dweck, C. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.

Einhorn, H. J. (1982). Learning from experience and suboptimal rules in decision making. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 268–286). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.020

Ericsson, K. A. (2006). The influence of experience and deliberative practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, R. Hoffman, & P. Feltovich (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 683–704). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816796.038

Ericsson, K. A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence, 45,* 81–103. http://dx.doi.org/10.1016/j.intell.2013.12.001

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty–Fifty=50%? *Journal of Behavioral Decision Making, 12,* 149–163. http://dx.doi.org/10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J

Fischhoff, B., & Chauvin, C. (Eds.). (2011). *Intelligence analysis: Behavioral and social scientific foundations*. Washington, DC: National Academies 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,* 552–564. http://dx.doi.org/10.1037/0096-1523.3.4.552

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19,* 25–42. http://dx.doi.org/10.1257/089533005775196732

Funder, D., & Ozer, D. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44,* 107–112. http://dx.doi.org/10.1037/0022-3514.44.1.107

Furnham, A., & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19,* 28–33. http://dx.doi.org/10.1016/j.lindif.2008.02.001

Green, K. C., & Armstrong, J. S. (2007). Global warming: Forecasts by scientists versus scientific forecasts. *Energy & Environment, 18,* 997–1021. http://dx.doi.org/10.1260/095830507782616887

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York, NY: McGraw-Hill.

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,* 188–201.

Itoh, S., Ikeda, M., Mori, Y., Suzuki, K., Sawaki, A., Iwano, S., . . . Ishigaki, T. (2002). Lung: Feasibility of a method for changing tube current during low-dose helical CT. *Radiology, 224,* 905–912. http://dx.doi.org/10.1148/radiol.2243010874

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526. http://dx.doi.org/10.1037/a0016755

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing." *Psychological Review, 103,* 263–283. http://dx.doi.org/10.1037/0033-295X.103.2.263

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90,* 644–651. http://dx.doi.org/10.1037/0022-3514.90.4.644

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior & Human Performance, 26,* 149–171. http://dx.doi.org/10.1016/0030-5073(80)90052-5

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. http://dx.doi.org/10.1177/0272989X0102100105

Loftus, E. (1996). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

MacCoun, R. J. (2012). The burden of social proof: Shared thresholds and social influence. *Psychological Review, 119,* 345–372. http://dx.doi.org/10.1037/a0027121

McGraw, P., Todorov, A., & Kunreuther, H. (2011). A policy maker's dilemma: Preventing terrorism or preventing blame? *Organizational Behavior and Human Decision Processes, 115,* 25–34. http://dx.doi.org/10.1016/j.obhdp.2011.01.004

Mellers, B. A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . Tetlock, P. (2014). *Improving probabilistic predictions by identifying and cultivating "superforecasters"*. Manuscript submitted for publication.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25,* 1106–1115. http://dx.doi.org/10.1177/0956797614524255

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79,* 489–500.

Nickerson, R. S. (1987). *Understanding understanding*. New York, NY: Bolt Beranek and Newman.

Parker, A., & Fischhoff, B. (2005). Decision making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making, 18,* 1–27. http://dx.doi.org/10.1002/bdm.481

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17,* 407–413. http://dx.doi.org/10.1111/j.1467-9280.2006.01720.x

Plomin, R., Shakeshaft, N. G., McMillan, A., & Trzaskowski, M. (2014). Nature, nurture, and expertise. *Intelligence, 45,* 46–59. http://dx.doi.org/10.1016/j.intell.2013.06.008

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89. http://dx.doi.org/10.1111/1467-8721.ep10768746

Reyna, V. F., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2014). Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students. *Psychological Science, 25,* 76–84. http://dx.doi.org/10.1177/0956797613497022

Riding, R., & Cheema, I. (1991). Cognitive styles—An overview and integration. *Educational Psychology, 11,* 193–215. http://dx.doi.org/10.1080/0144341910110301

Ruthsatz, J., Detterman, D. K., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence, 36,* 330–338. http://dx.doi.org/10.1016/j.intell.2007.08.003

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9,* 340–346. http://dx.doi.org/10.1111/1467-9280.00066

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162–173. http://dx.doi.org/10.1037/0022-3514.86.1.162

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252–266. http://dx.doi.org/10.1016/0749-5978(92)90064-E

Soll, J. B., Milkman, K. L., & Payne, J. W. (in press). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of judgment and decision making*. New York: Wiley.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89,* 342–357. http://dx.doi.org/10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188. http://dx.doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13,* 225–247. http://dx.doi.org/10.1080/13546780600780796

Steiner, I. D. (1972). *Group processes and group productivity*. New York, NY: Academic Press.

Steyerberg, E. W. (2009). *Clinical prediction models*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-77244-8

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35,* 401–426. http://dx.doi.org/10.1016/j.intell.2006.09.004

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26. http://dx.doi.org/10.1111/1529-1006.001

Taleb, N. N. (2007). Black swans and the domains of statistics. *The American Statistician, 61,* 198–200. http://dx.doi.org/10.1198/000313007X219996

Tetlock, P. E. (1998). Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right. *Journal of*

Personality and Social Psychology, 75, 639–652. http://dx.doi.org/10.1037/0022-3514.75.3.639

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66,* 542–554. http://dx.doi.org/10.1037/a0024285

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago, IL: Psychometric Monographs.

Vannoy, J. S. (1965). Generality of cognitive complexity-simplicity as a personality construct. *Journal of Personality and Social Psychology, 2,* 385–396. http://dx.doi.org/10.1037/h0022270

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049–1062. http://dx.doi.org/10.1037/0022-3514.67.6.1049

Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science, 23,* 11–16. http://dx.doi.org/10.1177/0963721413504781

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277–295. http://dx.doi.org/10.1146/annurev.psych.54.101601.145028

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117–142. http://dx.doi.org/10.1037/0033-2909.116.1.117

Wilson, T. D., & Gilbert, D. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science, 14,* 131–134. http://dx.doi.org/10.1111/j.0963-7214.2005.00355.x

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *A manual for the group embedded figures test*. Palo Alto, CA: Consulting Psychologists Press.

Received April 21, 2014
Revision received October 13, 2014
Accepted October 21, 2014 ■