MANAGEMENT SCIENCE
Articles in Advance, pp. 1–15
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2016.2525

Confidence Calibration in a Multiyear Geopolitical Forecasting Competition

Don A. Moore
University of California, Berkeley, Berkeley, California 94720, don.moore@alumni.carleton.edu

Samuel A. Swift
Betterment, LLC, New York, New York 10010, samswift@gmail.com

Angela Minster, Barbara Mellers, Lyle Ungar, Philip Tetlock
University of Pennsylvania, Philadelphia, Pennsylvania 19104
{angelaminster@gmail.com, mellers@wharton.upenn.edu, ungar@cis.upenn.edu, tetlock@wharton.upenn.edu}

Heather H. J. Yang
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, heatherhjyang@gmail.com

Elizabeth R. Tenney
University of Utah, Salt Lake City, Utah 84112, tenney@business.utah.edu

This research examines the development of confidence and accuracy over time in the context of forecasting. Although overconfidence has been studied in many contexts, little research examines its progression over long periods of time or in consequential policy domains. This study employs a unique data set from a geopolitical forecasting tournament spanning three years in which thousands of forecasters predicted the outcomes of hundreds of events. We sought to apply insights from research to structure the questions, interactions, and elicitations to improve forecasts. Indeed, forecasters' confidence roughly matched their accuracy. As information came in, accuracy increased. Confidence increased at approximately the same rate as accuracy, and good calibration persisted. Nevertheless, there was evidence of a small amount of overconfidence (3%), especially on the most confident forecasts. Training helped reduce overconfidence, and team collaboration improved forecast accuracy. Together, teams and training reduced overconfidence to 1%. Our results provide reason for tempered optimism regarding confidence calibration and its development over time in consequential field contexts.

Keywords: confidence; overconfidence; forecasting; prediction
History: Received April 17, 2015; accepted April 13, 2016, by Yuval Rottenstreich, judgment and decision making. Published online in *Articles in Advance* August 22, 2016.

Introduction

Overconfidence may be the most consequential of the many biases to which human judgment is vulnerable, both because of its ubiquity and because of its role in facilitating other biases and errors (Bazerman and Moore 2013, Fischhoff 1982, Kahneman 2011). Overconfidence affects the judgments of physicians (Oskamp 1965), entrepreneurs (Cooper et al. 1988), bridge players (Keren 1987), government planners (Flyvbjerg et al. 2002), investors (Statman et al. 2006), and basketball players (Jagacinski et al. 1977), to name but a few examples. Research has identified overconfidence in tests of declarative knowledge, bets, and predictions of the future (Ben-David et al. 2013, Fischhoff et al. 1977, Massey et al. 2011). Perhaps it should come as no surprise that forecasts of geopolitical events, so central to intelligence analysis and policy formulation, are also biased by overconfidence (Gardner 2010, Silver 2012). The question we ask in this paper is whether there are conditions under which this bias can be reduced or even eliminated.

On the one hand, Tetlock's (2005) long-term survey of political experts suggests pessimism, as the experts in his sample were persistently overconfident. Although they clearly believed they had expertise, the evidence suggests their expertise was not as useful as they seemed to think it was. Dilettantes forecasting outside their domain of expertise were no less accurate than those who claimed to be experts (Tetlock 2005). Yet these experts lacked incentives rewarding accuracy, training in the use and interpretation of probability scales, and practice, and perhaps most importantly, they lacked timely and unambiguous feedback (Benson and Onkal 1992, Hoelzl and Rustichini 2005, Larrick 2004). Research has found that each of these can help reduce overconfidence, but effects have generally been studied over short time horizons, usually constrained by the duration

of laboratory experimental sessions (Soll et al. 2016). There are legitimate questions about the degree to which these debiasing effects generalize over longer time horizons and in more consequential domains (Dawes and Mulford 1996, Gigerenzer 1991, Juslin et al. 2000).

We have a unique opportunity to address these questions. Our data come from a geopolitical forecasting tournament sponsored by the Intelligence Advanced Research Projects Activity (IARPA) of the United States federal government. Our research group was one of five that provided the IARPA with daily probabilistic forecasts on a set of hundreds of world events. These daily forecasts represented the aggregation of the forecasts from hundreds of people making their own predictions of what would happen. Each forecast was accompanied by a confidence judgment that reflected how sure the forecaster was that he or she knew what was going to happen. We examine these judgments for the presence of overconfidence and see how that changes with time and experience.

The tournament was the IARPA's attempt to improve geopolitical forecasting and intelligence analysis. Current systems rely primarily on qualitative assessments of probabilities and risk (Mandel and Barnes 2014). Qualitative probability estimates are difficult to score, aggregate, and analyze. They limit accountability because it is unclear what constitutes a good forecast. They also limit usefulness because qualitative forecasts cannot be incorporated into expected value calculations that could inform policy decisions by estimating expected consequences. The IARPA's forecasting tournament was designed to serve as an important proof of the viability of quantitatively scored forecasting. The IARPA scored each research team's forecasts using the Brier score, an incentive-compatible scoring rule that rewarded researchers for helping forecasters make the best predictions they could. Each research team in the tournament independently recruited its own participants. As such, we sought to identify through our study the conditions that would provide the best opportunity for accurate and well-calibrated forecasts.

In the design of our study, we faced innumerable decisions, large and small, about how to recruit participants, how to train and orient them, how to compensate them, how to elicit their beliefs, and how to provide them with feedback, among many other things. We were guided in these decisions by the research evidence and, when none existed, our own intuitions. Whenever possible, we sought to employ recruiting tools, situations, incentives, question formats, and response formats that had the best chance of producing accurate, reliable, and well-calibrated forecasts. It was not possible for us to vary all these things in our research design, of course. Instead, we focused on two dimensions on which we had reason to believe that experimental variation would provide the most interesting and informative results: probability training and group interaction.

The Role of Training

One of the most ambitious studies of training provided participants with 45 minutes of training on the calibration of confidence judgments, followed by 22 testing sessions, each an hour long (Lichtenstein and Fischhoff 1980). Another study employed 6 testing sessions, each two hours long (Stone and Opel 2000). These interventions showed some benefits of training for reducing overconfidence and improving the accuracy of probability judgments, but with degradation over time and limited generalization beyond the training context. Although both studies suggest that training should be helpful, they do not attempt to examine its effectiveness over the span of years. We set out to test the potential for training to endure over a year on a diverse set of forecasting questions across many different domains.

Our approach to training included four different components that we hoped might help forecasters in their work. First, we encouraged them to take the outside view by considering how often, under similar circumstances, something like the event in question took place (Kahneman and Lovallo 1993). This outside view demands consideration of relevant comparisons and historical events. Second, the training encouraged forecasters to average across opinions, either others' or their own. This advice was an attempt to help them exploit the wisdom of the crowd, thereby averaging some of the idiosyncrasies and noise out of individual predictions and strengthening the signal value they contain (Larrick and Soll 2006). Third, we suggested that they employ mathematical and statistical models where appropriate. For those forecasters who understood them, tools like Bayes' theorem could prove useful. Fourth, we provided some education regarding biases relevant to forecasting. In particular, the materials discussed the twin risks of overconfidence and excess caution in estimating probabilities, noting their consequences on calibration and Brier scores. For more details on exactly what the training entailed, see the online supplementary material (available at http://dx.doi.org/10.1287/mnsc.2016.2525 and https://osf.io/ecmk6/).

The Role of Group Interaction

Examining the effect of group deliberation is of some practical interest, given that most important decisions made by organizations, institutions, and governments are made by groups. Intelligence analysis in particular is often conducted within the social context of an agency, where analysts discuss forecasts with one

another. Reports and recommendations are collaborative products.

Prior evidence presents a mixed picture on the potential benefits of group discussion. On the one hand, it can lead to increased confidence and thus contribute to overconfidence (Buehler et al. 2005), especially when the most confident people in the group are the most influential (Anderson et al. 2012). Group overconfidence is magnified, at least in part, by the potential for discussion to polarize attitudes (Moscovici and Zavalloni 1969). When dissent and disagreement are suppressed, the group will reinforce each other's biases rather than correct them (Stasser and Titus 1987). On the other hand, when discussions lead to the sharing of useful information, they can increase accuracy (Stasser and Davis 1981). Furthermore, an increase in perceived accountability to the group can increase self-critical reflection and help reduce overconfidence (Lerner and Tetlock 1999, Sniezek and Henry 1989).

Some of the pressures pushing groups toward a focus on common information have to do with seeking harmony (Janis and Mann 1977). Because it is more pleasant to be part of a cohesive team, people will dampen some dissent in the interest of collective consensus (Dobbins and Zaccaro 1986). It is important that our teams did not meet face to face, but instead interacted exclusively online via a website that made it possible to share forecasts and comments. At no point did the individuals have to chat, socialize, or interact face to face with others on their team. The geographically distributed nature of these teams, interacting only via technologically mediated communication, makes them distinct. However, geographically distributed teams are becoming more common in our increasingly wired world, where people in different places work together using technologically mediated communication.

Effects Over Time

Almost all of the small handful of studies that examine calibration outside of the lab examine confidence judgments taken at one point in time (Glaser and Weber 2007, Park and Santos-Pinto 2010). The few longitudinal studies suffer from sporadic sampling and relatively few judgments (Ben-David et al. 2013, Dunlosky and Rawson 2012, Simon and Houghton 2003). In the current study, we examine probabilistic forecasts of important events over a period of three years. Our data allow us to examine the development of confidence judgments over time with regular updating and hundreds of forecasts from each participant. We can track forecast accuracy and observe the degree to which forecasters learn from experience and feedback.

Many people share the intuition that calibration should improve as people become better informed. The more information people have about a forecast question topic, the better they might be at detecting when they are right and when they should be less certain (Burson et al. 2006, Kruger and Dunning 1999). However, some kinds of information increase confidence without increasing accuracy, and vice versa, even for experts (Griffin and Tversky 1992). As Oskamp (1965) memorably demonstrated, psychologists who learned more details of patients' life histories grew more confident in their diagnoses, without commensurate increases in accuracy. If additional information enhances confidence more than accuracy, it could drive up overconfidence (Deaves et al. 2010). On the other hand, of course, there is the possibility that confidence and accuracy change according to different inputs but ultimately balance each other, and that across time confidence increases at roughly the same rate as accuracy (McKenzie et al. 2008).

Self-Rated Expertise

One striking feature of the forecasting questions we examine is the value of specialized domain knowledge. Accurately forecasting the probability that Greece would exit the euro was facilitated by understanding Greek national identity and the political viability of a return to the drachma as its national currency. It seems reasonable to think that those forecasters with the most knowledge of Greek politics would know enough to make well-calibrated forecasts. After all, as Kruger and Dunning (1999) noted, the most ignorant may also lack an appreciation for how much they do not know (Burson et al. 2006). It is also simply the case that when accuracy is lower, there is more room to be overconfident. So if those with less expertise are less accurate, there is good reason to expect them to be more overconfident.

However, this prediction depends on a strong correlation between self-rated expertise and actual accuracy. This correlation will be driven down when specialized local knowledge increases confidence without increasing accuracy (Oskamp 1965, Wells and Olson 2003), such as when the most distinguishing feature of experts is their willingness to take strong, opinionated stances (Tetlock 2005). The correlation will be driven down further by knowledge that increases accuracy without a commensurate effect on confidence (Griffin and Tversky 1992). These sorts of influences will drive down the correlation between accuracy and expertise, leading us to expect a weak correlation between self-rated expertise and calibration in confidence judgments.

The Present Research

The IARPA forecasting tournament pitted five research groups against each other. At the end of the second year, our research group's accuracy was sufficiently superior to that of the other four groups that

the project sponsor elected to cut funding to all four of the other groups. Our group (modestly dubbed the "Good Judgment Project") was the only one that continued into the third year of forecasting. The present paper examines data from three years of the forecasting competition, focusing on the calibration of our forecasters' confidence judgments. In particular, we analyze the development of confidence and accuracy over time. Where do we observe effects of time, experience, and learning?

It is worth distinguishing this paper from others that have emerged from the Good Judgment Project. Three other papers examine the conditions that contribute to individual forecasting accuracy. Mellers et al. (2014) provide an overview of the forecasting tournament and discuss the positive impacts of three behavioral interventions—training, teaming, and tracking—on individual performance in prediction polls using data from the first two years of the tournament. Mellers et al. (2014) do not, however, examine confidence and accuracy over time to understand how they develop. Mellers et al. (2015a) explore the profiles of individual forecasters using dispositional, situational, and behavioral variables. In another paper, Mellers et al. (2015b) document the performance of the most accurate performers, known as superforecasters, and identify reasons for their success.

Aggregation techniques for prediction poll data are discussed in three other papers from the Good Judgment Project. Satopää et al. (2014a) offer a simple method for combining probability estimates in log-odds space. This method discounts older forecasts and recalibrates or "extremizes" forecasts to reflect the amount of overlapping information of individual opinions. Satopää et al. (2014b) describe a time-series model for combining expert estimates that are updated infrequently. Baron et al. (2014) provide a theoretical justification and empirical evidence in favor of transforming aggregated probability predictions toward the extremes. Atanasov et al. (2016) develop a method for aggregating probability estimates in prediction markets when probabilities are inferred from individual market orders and combined using statistical aggregation approaches. Tetlock et al. (2014) discuss the role that tournaments can play in society by both increasing transparency and improving the quality of scientific and political debates by opening closed minds and holding partisans accountable to evidence and proof.

The research questions we ask in this paper are distinct from those in these other papers. To preview our results, we find that forecasters making predictions about consequential world events can be remarkably well calibrated. Confidence and accuracy move upward together in parallel over time as forecasters gain information. In addition, training is astoundingly effective: an hour of training halves overconfidence over the following year. Our distributed teams are also slightly better calibrated than individuals.

Method

Our data comprise 494,552 forecasts on 344 individual forecasting questions over a period of three years from 2,860 forecasters. Each of the three forecasting "years" lasted about nine months, roughly coinciding with the academic year.

Participants

We recruited forecasters from professional societies, research centers, alumni associations, science blogs, and word of mouth. Once forecasters had provided their consent to participate in the research, they had to complete roughly two hours' worth of psychological and political tests and training exercises. This included several individual difference scales whose results were analyzed by Mellers et al. (2015a) in more detail than we can do justice to here.

Participants who stuck with it for the entire year and made at least 25 forecasts received a payment at the end of the year ($150 after year 1 and $250 after years 2 and 3). Those who persisted from one year to the next received a $100 bonus. Despite this modest compensation, forecasters' dedication was impressive. Most spent several hours each week collecting information, reading the news, and researching issues related to their forecasts. Some spent more than 10 hours per week. The most dedicated forecasters built their own analytical tools for comparing particular questions to relevant reference classes or updating their probability estimates based on relevant evidence.

Our data come from all participants who submitted at least one valid forecast. They had a median age of 35 years (SD = 13.7); 83% of them were male; 26% had Ph.D.s, 37% had master's degrees, 36% had only an undergraduate education, and less than 1% had not graduated from college; and 78% were U.S. citizens.

Materials

Questions. A total of 344 specific questions, created by the IARPA, had resolved by the end of year 3 and were included in the present analyses. The IARPA selected these questions so that they were relevant to decisions in U.S. government policy, had unambiguous resolution criteria, had to be resolvable within a reasonable time frame (generally less than a year), and were sufficiently difficult to forecast. In particular, they deemed forecasts with below 10% or above 90% chance of occurring as too easy, and instead aimed for events with more middling probabilities of occurrence. They were, in short, the tough calls.

A list of all the questions appears in this paper's online supplement. New questions were released roughly every week in batches of about four or five. Questions were open from 1 to 549 days (mean = 114), during which forecasters could update their forecasts as frequently as they wished. The average forecaster submitted forecasts on 65 different questions. There were three types of questions:

1. The majority of questions (227 of 344) asked about binary outcomes. Examples include, "Will the United Nations General Assembly recognize a Palestinian state by 30 September 2011?" and "Will cardinal Peter Turkson be the next pope?"

2. Multinomial questions (45 of 344) asked about more than two outcome categories. An example is, "Who will win the January 2012 Taiwan presidential election?" Answers were "Ma Ying-jeou,'' "Tsai Ing-wen,'' and "neither.'' There were 27 multinomials that asked about three outcomes, 31 that asked about four, and 9 that asked about five.

3. Conditional questions (72 of 344) had two antecedents and two outcomes each. For example, one of these conditional questions 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?" Forecasters provided probabilities for both arms of the conditional, but only forecasts for the realized condition were scorable.

Confidence and Calibration. Each forecast specified the probability of each of the possible outcomes for a given question. The elicitation interface forced the forecaster to consider all possible outcomes and specify the probability of each, such that they summed to 100%. This approach to elicitation has proven useful for producing better-calibrated confidence judgments and reducing the inflation of probabilities observed following narrow focus on a specific outcome (Haran et al. 2010, Tversky and Koehler 1994). Forecasters knew that after a question closed and its outcome was known, we would score each day's forecast using the Brier (1950) scoring rule to compute the score for that one question. Since the Brier score rewards accurate reporting, it provided useful incentive properties.

However, the Brier score is multidimensional and reflects a number of different components that can be decomposed (Yates 1982). In this paper, we focus on calibration and resolution. For each question, we identified the outcome that the forecaster reported to be most likely and took the associated probability as the forecaster's confidence. To assess their calibration, we grouped forecasts with similar degrees of confidence and then compared them to the actual frequency with which these forecasts proved correct. Identifying the one outcome the forecaster deemed most likely also allows us to measure hit rates. The hit rate is the proportion of the time the outcome identified as most likely was actually the outcome that occurred. Occasionally, a forecaster selected more than one outcome as most likely. This happened, for instance, when two of four possible outcomes each received a 50% chance of occurring. In this case, if either of these two outcomes occurred it counted as 50% of a hit.

It is possible to have good calibration but bad resolution. This would be the case for a weather forecaster who simply predicts a 50% chance of rain every day in a city where it rains on half of all days. Perfect resolution, on the other hand, would constitute accurate forecasts predicting rain with either 100% or 0% probability. Our results also examine resolution—discrimination between events that occur and those that do not. Good resolution is evident in a range of forecast probabilities that correspond well with the actual probability of events' occurrence.

Expertise. Forecasters rated their expertise (using a 1 to 5 scale) on each question they answered. In year 1 the response scale ran from "uninformed" to "complete expert." In year 2, the question asked forecasters to place themselves in one of the five expertise quintiles relative to others answering the same question. In year 3, participants indicated their confidence in their forecast from "not at all" to "extremely."

Design and Procedure

We randomly assigned participants to one of four conditions in a 2 (individual versus team) × 2 (no training versus training) factorial design.[1] All forecasters in all conditions could update their forecasts as often as they wished. A forecast stood until the question was resolved or the forecaster updated it.

Individual vs. Team Conditions. The first experimental factor varied the amount of interaction between forecasters.

[1] Because the other conditions are not well suited to testing our research questions, we omit discussion of those conditions: a crowd-prediction condition in which people knew the consensus forecast when they made their own, which only existed in year 1, and a prediction market condition. Moreover, we omit discussion of a scenario-training condition that was only used in year 1. We also omit data from the select group of "superforecasters" in years 2 and 3. For more information about these other conditions, see Mellers et al. (2014). For more detail about the prediction-market conditions, see Atanasov et al. (2016).

In the individual conditions, forecasters worked alone and did not interact with one another. In the team conditions, forecasters were assigned to groups of approximately 15. Interaction between team members occurred exclusively via an online forecasting platform, which we provided. We sought to structure team interaction to maximize its potential benefit. We encouraged team forecasters to justify their forecasts by providing reasons and to discuss those reasons with their teams. Those in team conditions also received guidance on how to create a well-functioning group. Members were encouraged to maintain high standards of proof and seek out high-quality information. They were encouraged to explain their forecasts to others and offer constructive critiques when they saw opportunities to do so. Members could offer rationales for their thinking and critiques of others' thinking. They could share information, including their forecasts. Forecasters were encouraged to challenge each other with logical arguments and evidence, especially when they observed group members make forecasts with which they disagreed. Examples of the suggestions we gave to forecasters in the team condition can be found in the online supplement.

Probability Training. The second experimental manipulation varied the provision of probability training. The training coached participants on how to think about uncertainties in terms of probabilities and frequencies. It warned them specifically against the dangers of overconfidence. The training included a test of knowledge in which participants provided confidence estimates on the accuracy of their answers and then received feedback on their accuracy. Participants in this condition completed the one-hour training online before they submitted any forecasts. The details of this training are available in the online supplementary materials.[2]

Leaderboard and Incentives. Brier scores, averaged across questions, were calculated after the first ten questions closed and were updated every time a question closed after that, providing forecasters with regular feedback. These scores determined the order in which individual forecasters' chosen user names appeared on leaderboards that ranked forecasters within condition and were visible to other forecasters in the individual condition. Members of teams were ranked on a leaderboard relative to other members of their team in years 1 and 2. Because we were concerned that this intrateam ranking might have stimulated competition within the team, in year 3 we eliminated the intrateam leaderboard and replaced it with one that ranked all teams relative to each other.

Forecasters in the individual condition who declined to provide a forecast for a particular question received the median score from others in the same condition. This provided an incentive to forecast only if the forecaster thought he or she could provide a forecast more accurate than those of the other forecasters. Note that these imputed scores are not part of any of the results we report in this paper.

Forecasters in the team conditions who declined to forecast on a particular question received the median score from their team members who did make forecasts. This scheme rewarded individuals for helping their teammates make the most accurate forecasts they could, and forecasting themselves when they thought they could be more accurate than the median. They were, however, not forced to come to consensus; different group members could make different forecasts.

Results

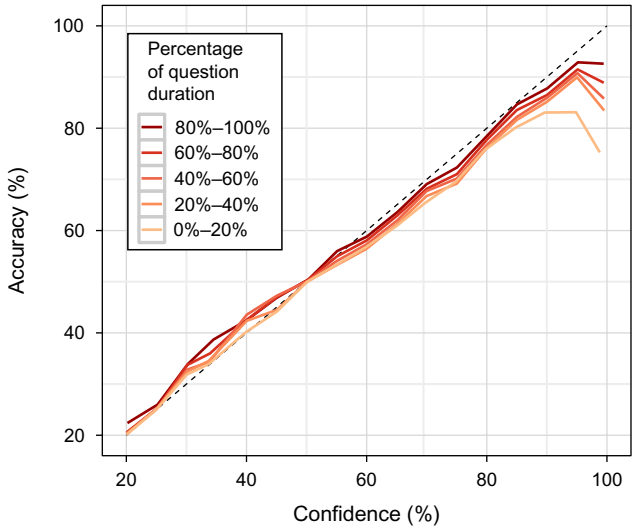
Our evidence suggests that, overall, forecasters were quite well calibrated and exhibited only a small degree of overconfidence. On average, our forecasters reported being 65.4% sure that they had correctly predicted what would happen. In fact, they were correct 63.3% of the time, for an overall level of 2.1% overconfidence. The difference between accuracy and confidence exhibits a small effect size with a Cohen's d of 0.21, with a 95% confidence interval of (0.166, 0.259).

Our forecasters were not equally overconfident across the range of confidence. Figure 1 divides confidence into bins, as is common practice (Keren 1991). The most striking result is how well calibrated forecasters were: the dots lie close to the identity line. This stands in contrast to the standard findings from laboratory studies of overconfidence (Lichtenstein et al. 1977), and the 9% overconfidence estimated in the Juslin et al. (2000) review of the literature. Instead, our forecasters show a degree of calibration akin to the famously well-calibrated meteorologists studied by Murphy and Winkler (1977). The average Brier (1950) score of the meteorologists' predictions regarding the probability of precipitation the next day was 0.13. The average Brier score of forecasters in the last week of forecasting on each question was 0.14. For the last day of forecasting, it was 0.10.[3] This represents impressively good calibration, especially because the forecasting questions were selected to be difficult and overconfidence tends to be greatest on difficult questions (Erev et al. 1994, Juslin et al. 2000, Klayman et al. 1999).

[2] Note that in reassigning participants to experimental conditions for the second forecasting year, some of those who had received scenario training in Year 1 went on to receive either training or no training in Year 2. The scenario training condition did not affect calibration or overconfidence and thus is not discussed further in this paper.

[3] Lower Brier scores indicate better accuracy.

Figure 1 (Color online) Calibration Curves, Conditional on When in the Question's Life the Forecast Was Made



The axes on Figure 1 go down to 20% because some forecasting questions had five possible outcomes. Naturally, if questions had only two alternatives, then the axes would just go down to 50% since that would be the lower expected limit of accuracy, among those who were just guessing. Obviously, overconfidence is greatest when confidence is high. This is no surprise—there is simply more room for hit rates to fall below forecast confidence as confidence approaches 100% (Erev et al. 1994). What is also striking about the calibration curve is the downturn in hit rates at confidence levels near 100%, a result that holds across experimental conditions, as shown in Figure 2. This downturn arises largely from the 7.8%

Figure 2 (Color online) Confidence and Accuracy Curves as a Function of Experimental Condition (Individual vs. Team × Training vs. No Training)

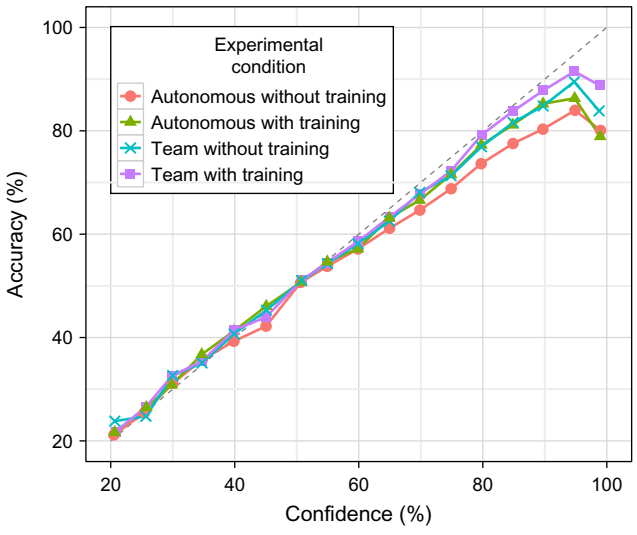
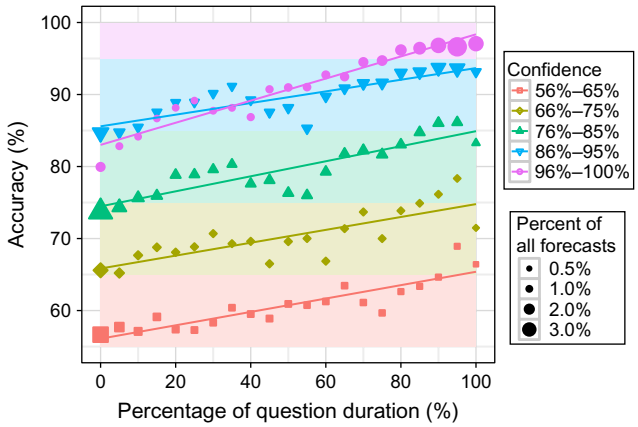


Figure 3 (Color online) Accuracy, Expressed in Hit Rates, as a Function of the Forecast Confidence and When the Forecast Was Made During the Duration of a Question



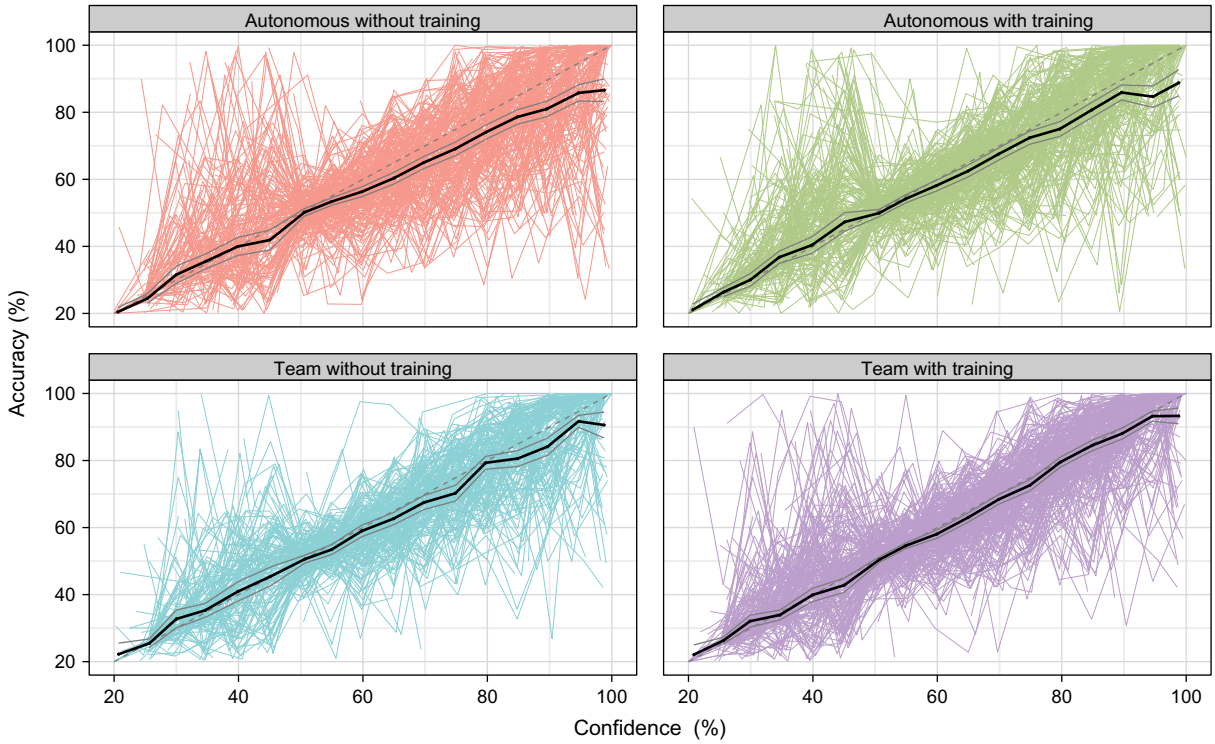
Note. Lines indicate the linear best fit of the duration–accuracy relationship among forecasts in each confidence bin.

of forecasts that indicated the forecaster was absolutely certain of the outcome. These forecasts only came true 84% of the time, whereas forecasts made with 95% confidence occurred 90% of the time.

Where do these extreme forecasts arise? Figure 1 makes it clear that the largest drop in accuracy occurred for those making extreme forecasts early in the life of a question. Figure 3 shows that extremely confident forecasts (forecast confidence of greater than 95%) were generally no more accurate than forecasts with 86%–95% certainty, but their accuracy was especially poor when made early in the life of a question. Making an extreme forecast early in a question's life represents a bold and high-risk bet that might best be characterized by forecasters "swinging for the fences" of accuracy—hoping for a home run, and simultaneously increasing the risk of striking out. Note that the length of time a question was open varied substantially, so the timing of forecasts in Figures 1 and 3 is measured as a percentage of the duration of each question. For additional analyses employing different operationalizations of time, see the online supplementary materials.

The finding that our forecasters' probability estimates are well calibrated might lead to questions of resolution. Resolution measures the ability to discriminate between events that do and do not occur. It is, of course, possible for confidence to match accuracy if everyone simply predicts the base rate. However, the analyses we present clearly go well beyond this simple calibration score. We show that forecasts span the entire range from the ignorance prior to 100% confident, that this variation exists not just between forecasters but also within individual forecasters across different questions, and that hit rates match forecast probabilities at those different levels of confidence. Figure 4 presents calibration curves for each of

Figure 4 **(Color online) Calibration Curves for the Four Experimental Conditions**



Notes. Each condition's mean is indicated by a dark like with 95% confidence intervals. Each curve is also surrounded by a cloud of individual lines, one for each forecaster in that condition.

our four conditions, accompanied by the variability between individual forecasters around the average. Additional comparisons of resolution scores across our experimental conditions are presented by Mellers et al. (2014).

Variation by Experimental Treatment

In support of some evidence suggesting that groups can reduce overconfidence (Sniezek and Henry 1989), we find that forecasters in the team conditions were even better calibrated than those in the solo forecasting conditions. As Table 1 shows, working in teams significantly improved accuracy and slightly reduced overconfidence. Training, for its part, slightly improved accuracy, but mostly improved calibration by reducing confidence. Perhaps somewhat surprisingly, we do not find that these treatment effects interacted with time. In other words, the beneficial effects of training and teaming hold systematically across time.

How Does Self-Rated Expertise Moderate the Confidence–Accuracy Relationship?

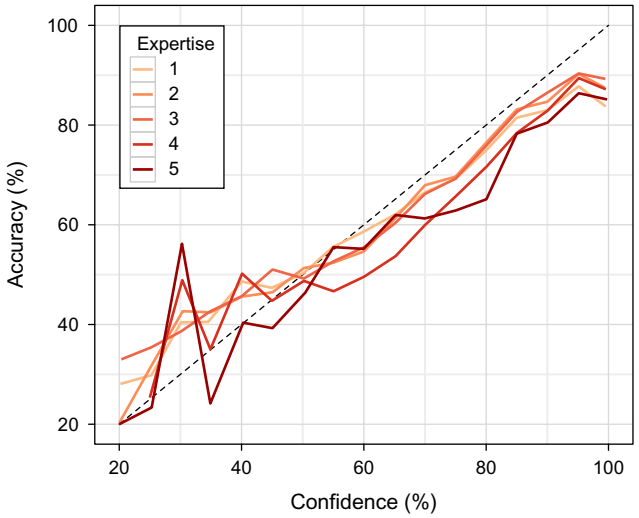
Some prior results have found that those who rated themselves as experts attained higher proportions of correct predictions, better calibration, and less overconfidence (Wright et al. 1994). Yet experts do not always perform better (Armstrong 2001, Tetlock 2005). In our data, self-rated expertise was not strongly related to calibration, accuracy, or Brier score. See Figure 5, which shows that self-reported expertise was not a reliable moderator of the relationship

Table 1 **Working in Teams Primarily Improves Accuracy, While Training Primarily Reduces Overconfidence**

Elicitation	Training	Confidence (%)	Hit rate (%)	Overconfidence (%)
Autonomous	None	64.9 (8.5)	60.7 (10.3_c)	4.3_a (9.4_e)
Autonomous	Training	63.5 (8.2)	61.5 (10.6_c)	2.0_b (9.7_e)
Team	None	65.8 (7.9)	62.9 (12.7_d)	$2.8_{a,b}$ (12.5_f)
Team	Training	64.3 (8.4)	63.6 (11.0_c)	0.7_b (7.3_g)

Notes. Overconfidence measures with different subscripts (a, b) are significantly different from one another; these significance groupings are calculated using Bonferroni-adjusted 95% confidence intervals. The variance in overconfidence across individuals is heterogeneous across conditions. The standard deviation of each measure is reported in parentheses. Standard deviations with different subscripts (c–g) are significantly different from one another.

Figure 5 **(Color online) Confidence and Accuracy as a Function of Forecasters' Self-Rated Expertise on the Question**



between confidence and accuracy. These results do not vary substantially across Years 1, 2, and 3 and the different ways we posed the expertise question.

The lack of a strong correspondence between self-reported expertise and actual accuracy raises the question of whether our forecasters were systematically biased in their assessments of expertise. The answer to this question is yes, but in a surprising way. Forecasters reported themselves (on average) to be less expert than other forecasters. In year 2, when forecasters placed themselves into expertise quintiles, if they were well calibrated, they should have divided themselves evenly between the five categories of expertise, and the mean should have been in the middle category—a 3 on the five-point scale. In fact, mean self-reported expertise in year 2 was 2.44 (SD = 1.07, $n = 152{,}660$), well below this midpoint, implying that forecasters, on average, believed that they were less expert than others. The absolute phrasing of the expertise question used in years 1 and 3 does not allow this check on collective rationality, but mean expertise was below the scale midpoint in both those years (year 1, mean = 2.18, SD = 0.92, $n = 141{,}186$; year 3, mean = 2.69, SD = 1.14, $n = 203{,}553$).

Finding underplacement is surprising because two different varieties of overconfidence appear to be at odds with one another. Forecasters exhibited underconfidence by underplacing themselves relative to other forecasters with regard to their relative expertise, even while they overestimated the probability that their forecasts were correct. Our finding of underplacement replicates other results showing that "better than average" beliefs are far from ubiquitous (Moore and Healy 2008, Moore 2007). On difficult tasks, people routinely believe that they are worse than others (Kruger 1999). Indeed, it is on these difficult tasks that people are most likely to overestimate their performance (Larrick et al. 2007).

Does Good Calibration Change Over Time?

Our results find a remarkable balance between people's confidence and accuracy. Confidence and accuracy increased over time in lockstep. In the first month of forecasting in year 1, confidence was 59.0% and accuracy was 57.0%. In the final month of the third year, confidence was 76.4%, and accuracy was 76.1%. However, this result glosses over important differences across questions. The population of questions changed over time, and confidence and accuracy varied widely across questions. To control for those differences, we examined confidence and accuracy within question as the closing date approached.

Figure 6 shows confidence and hit rate averaged across all forecasting questions as the day on which the question closed drew nearer. Both confidence and hit rate reliably went up as a question's close drew near, demonstrating impressive calibration. This same pattern is reflected in Figure 3: all the lines slope up as time passes because accuracy moves up as forecasters gain information, even if that information is simply the passage of time. The accompanying increase in confidence manifests itself in the larger numbers of more confident forecasts over time as a question's closing approaches. But Figure 6 also shows there was also a persistent gap between confidence and accuracy: confidence systematically exceeded accuracy by a small but consistent amount.

While we do find that calibration increased from the beginning of a tournament year to the end, we do

Figure 6 **(Color online) The Persistent Gap Between Confidence and Accuracy Over the Duration of Forecasting Questions**

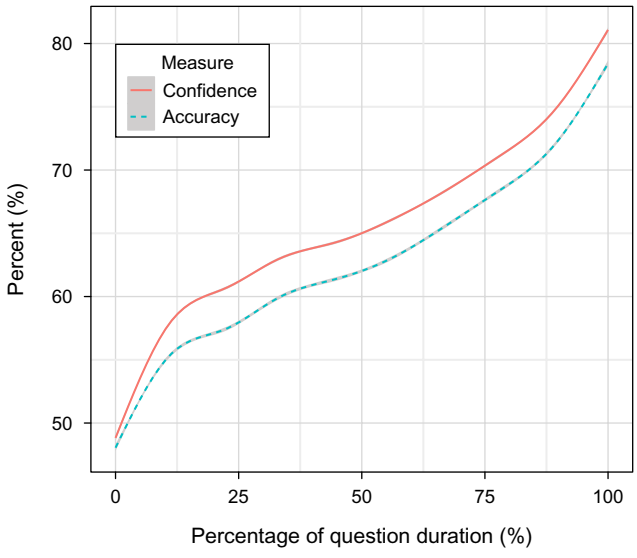
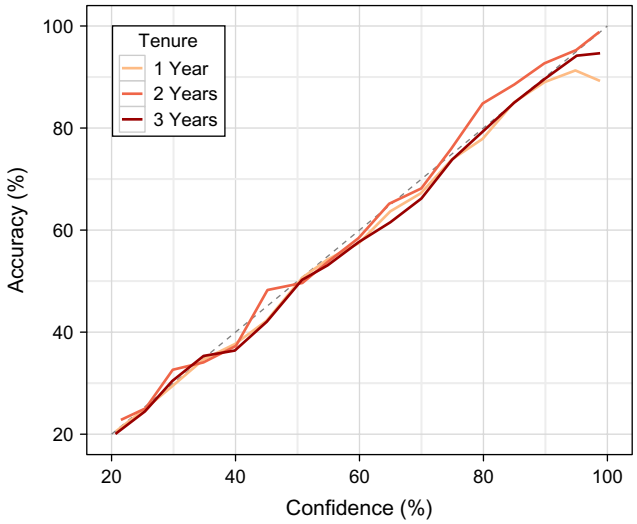


Figure 7 **(Color online) Confidence and Accuracy as a Function of Forecasters' Years of Participation in the Tournament (Tenure)**



not find that having more years of forecasting experience (forecasting tenure) led to an improvement in calibration. Figure 7 shows that the calibration curves of forecasters with one, two, and three tournament years of forecasting experience showed about the same level of calibration, even though questions varied in difficulty across the years. Statistically, through analysis of variance, we find no significant differences in calibration between forecasters with more or less experience.

Discussion

We began this paper by asking whether we could identify conditions under which we would observe good calibration in forecasts of consequential geopolitical events. Our results provide an affirmative answer. By applying some of the best insights from decades of research on judgment and decision making, we were able to structure the situation, incentives, and composition of a crowd of forecasters so that they provided accurate and well-calibrated forecasts of important geopolitical events.

There were some features of our approach that did not vary experimentally. When forecasting on a particular question, all forecasters had to specify the probabilities of the full set of mutually exclusive and exhaustive possible outcomes. All forecasters got frequent feedback using an incentive-compatible scoring rule (i.e., the Brier scores). All the forecasters were treated not so much as research subjects, but as partners in an important and path-breaking project testing the viability and accuracy of probabilistic forecasting of important world events.

There were some other features that we systematically varied. Our training did prove useful for improving calibration and reducing overconfidence. What surprised us was the durability of this intervention. Training appeared to reduce overconfidence similarly over the entire forecasting year, even many months after the actual intervention. Of course, a key aspect of our study is that forecasters got feedback on how accurate their forecasts were; this may have been important in cementing the benefits of training and helping them maintain good calibration. Perhaps more surprisingly, interaction in teams improved calibration. When forecasters collaborated with others, their forecasts became more accurate and better calibrated.

Our results replicate key findings of prior research, including the presence of overconfidence. But what is impressive is that the magnitude of overconfidence is smaller than in prior studies. Forecasters were extremely well calibrated. Our results also reveal an interesting pattern in the development of confidence over time. As our participants gained information, their confidence increased and accuracy improved (McKenzie et al. 2008). Indeed, the parallel increases in both confidence and accuracy may be the single most remarkable feature of our results. It is remarkable because the two alternatives are so salient. On the one hand, there are circumstances in which new information increases confidence more than accuracy, exacerbating overconfidence (Hall et al. 2007, Heath and Gonzalez 1995, Oskamp 1965). On the other hand, there are other circumstances in which increasing expertise can increase accuracy faster than confidence, reducing overconfidence (Koriat et al. 2002). The parallel development of confidence and accuracy over time implies that our forecasters' judgments navigated adeptly between the Scylla and Charybdis of these twin risks.

Moreover, the close development of confidence and accuracy over time implies that people are aware of how much better their forecasts are getting as they gather information. But this sort of self-awareness is contradicted by the weak effect of self-rated expertise. What is the crucial difference? We speculate that expertise that develops over time contributes to calibration because it allows a within-person comparison in which people can compare their level of knowledge on a question to their prior level of knowledge. The question of how expert you are is a more difficult question if it requires you to guess about how your expertise compares with that of others. We also note that our result is consistent with prior evidence suggesting that experts' confidence calibration is not necessarily better (McKenzie et al. 2008, Tetlock 2005).

However, some features of our results are at odds with prior research. Prior research would lead one to expect that interacting groups could exacerbate rather than ameliorate bias (Buehler et al. 2005, Kerr et al.

1996, Moscovici and Zavalloni 1969). Why then do we observe the opposite? We speculate that the beneficial effect of teams in our study is dependent on the unique nature of these groups. The forecasters were not friends or colleagues. Their only reason for existing and interacting as a group was to make more accurate forecasts. They did not need to get along, impress one another, or work together on other tasks. Their prime directive was accuracy, and their interaction was not complicated by interpersonal or social motives that can lead to the suppression of dissenting views in the interests of group harmony (Stasser and Titus 1987).

On the Importance of Forecasting

Every decision depends on forecasts of the future. Whether to bring an umbrella depends on the chances of rain. Whether to cash out one's investments depends on the future changes in capital gains taxes. Whether to launch a product depends on how it would sell. Over time we gain expertise that should increase our accuracy (Keren 1987). What happens to our confidence? The data we present offer a partial answer to this important question: because confidence increases along with increased accuracy, people continue to display overconfidence, even in the presence of good calibration and even as expertise and accuracy increase.

Our results show that increases in the presence of useful information increase accuracy over time. But greater information also increases forecasters' confidence in the accuracy of their forecasts, perhaps for good reason. As long as confidence goes up at the same rate as accuracy, good calibration will persist. Although our results do find evidence of overconfidence, the overall effect is smaller than in prior studies.

Reviews of prior studies report a 9% average difference between confidence and accuracy (Juslin et al. 2000). However, few of these prior findings come from forecasting. Some that do include Tetlock's (2005) study of expert political judgment. He documents an average of 12% overconfidence among the political experts in his study. This includes over 20% overconfidence among experts forecasting long-term outcomes within their domains of expertise. On the other hand, he finds about 3% overconfidence among those forecasting short-term outcomes in domains outside of their expertise.

This 3% overconfidence figure, as it turns out, is higher in studies that use difficult questions (Gigerenzer et al. 1991). Studies that include easier items produce less overconfidence (Klayman et al. 1999). There is simply more room to be overconfident on difficult items (on which most people are guessing) than on easy items (that most people get right). This is the so-called hard–easy effect in confidence judgments (Erev et al. 1994). Were our forecasting items easy or hard? We believe it would be fair to categorize them as hard. Indeed, that is explicitly how they were chosen: both important and highly uncertain. This fact ought to make it easier for our forecasters to show overconfidence, making their good calibration that much more impressive.

In fact, the performance of our forecasters rivals that of the legendary weather forecasters that scholars routinely hold up as the paragons of disciplined calibration (Murphy and Winkler 1977). The unique conditions of our forecasting tournament are probably key to our forecasters' performance. The fact that their forecasts would be scored against a clear standard for accuracy was, as with weather forecasters, undoubtedly crucial (Armor and Sackett 2006, Clark and Friesen 2009). It is also likely that our forecasters felt accountable to us and to each other, especially in the team condition (see Lerner and Tetlock 1999). We strongly suspect that the quality and regularity of feedback is likely to have been important (Butler et al. 2011, González-Vallejo and Bonham 2007, Lichtenstein and Fischhoff 1980), as it is for weather forecasters. We would be rash to assert that the consistent relationship between confidence and accuracy in our data is somehow necessary or universal. However, to the extent that daily life provides the kind of practice, clarity, and prompt feedback we provided our forecasters, we have reason to believe that calibration should look more like what we observe in our study and less like what lab studies might predict. At the same time, we must admit that life rarely calls upon us to make scorable, quantitative forecasts, and it is even rarer for forecasts to be followed by prompt, unambiguous feedback on actual outcomes and the performance of our forecasts.

Research evidence suggests that overconfidence persists across cultures and domains and can be robust to feedback (Harvey 1997, Sieck and Arkes 2005, Yates et al. 1998). Yet, some have argued that empirical evidence of overconfidence may be a consequence of artificial and unfamiliar tasks. Lab experiments in particular have been accused of making overconfidence seem more pronounced than it is. Indeed, there is some evidence that overconfidence shrinks as the domains of judgment become more similar to the information we encounter every day (Dawes and Mulford 1996, Gigerenzer 1991, Juslin et al. 2000). Still others maintain that overconfidence cannot be explained away so easily (Budescu et al. 1997). Questions about the robust persistence of overconfidence over the longer term shed light on this debate. If overconfidence reduces with experience and feedback, the laboratory findings of overconfidence

on novel tasks might be of little real consequence outside the lab. On the other hand, if overconfidence persists over extended periods of time, its importance and the potential need for debiasing interventions become stronger.

Limitations

Although our data have the benefit of a large sample size of diverse participants working on a task of obvious importance, they come with a number of limitations. First, the results offer frustratingly few clues regarding why exactly our forecasters are so well calibrated. We can point to beneficial effects of training and collaboration, but even forecasters in the solitary untrained condition display better calibration than prior research has documented. They were only 4% overconfident as opposed to 9% found by Juslin et al. (2000). Moreover, we can say little about what aspects of training or team collaboration helped. Determining why they were effective will require future research that investigates their elements more systematically, with more fine-grained experimental treatments and more complex experimental designs.

What natural variation does occur in our data provides little insight into the explanations for our forecasters' good accuracy and calibration. We are reluctant to conclude that our forecasters were better calibrated than the students in prior lab studies because they were older and better educated. Age and education are weak predictors of performance among our forecasters (Mellers et al. 2015a). If feedback and experience were essential to our forecasters' performance, then their calibration should have improved as they gained experience over the course of the forecasting year, or over the three forecasting years. However, we find no evidence for such improvement in our data. Their good calibration is evident from the outset. If the lessons of training were most effective immediately thereafter and waned over time, we should have seen performance degrade, yet we do not find evidence of such degradation. It is possible, of course, that degradation and improvement from experience were balancing each other enough that it interfered with our ability to detect either one, but that is just speculation about the lack of an effect in the results.

It is also worth noting again the unique nature of our participant population. They were exceptionally well educated, motivated, and informed. They differed in many ways, large and small, from the populations of undergraduates and workers on Amazon Mechanical Turk who populate so many studies of judgment and decision making. It could be that their age and experience contributed to their good calibration, but other studies have not reliably found that age is associated with better calibration in judgment. Indeed, age is sometimes correlated with greater overconfidence (Hansson et al. 2008, Prims and Moore 2015).

Proof of Concept

The good calibration of our forecasters offers a hopeful sign for the quantification of intelligence forecasts. One striking feature of most formal intelligence reports is how rarely they contain quantified estimations of probabilities (Chauvin and Fischhoff 2011). This omission is problematic for systematic approaches to decision making that might include a decision tree or an attempt to calculate the expected values of different policy choices (Armstrong 2001). However, we also acknowledge that quantification may, in fact, be a political liability. Intelligence analysts aware of their accountability to a political establishment prone to seeking blame when things go wrong may be skittish about making their forecasts clear enough to be tested and scored (Tetlock and Gardner 2015, Tetlock and Mellers 2011).

Politically, there will always be risks on either side of any probability estimate. On the one hand, there is the risk of a false positive: forecasting an event that does not occur, such as New York mayor Bill de Blasio's prediction that the blizzard of January 27, 2015, would be "the worst in the city's history." New York shut down its entire public transit system on that day, but in fact only received a mild dusting of snow. On the other hand, there is the risk of the false negative: the failure to forecast the storm's severity, as with hurricane Katrina's strike on New Orleans in August of 2005. But just as the truth is a strong defense against charges of libel, a well-calibrated analyst can point to the performance of a set of forecasts over time as evidence of his or her performance. Accuracy of the type our results demonstrate ought to be so valuable for planning and decision making that we hope it would outweigh the political risks of greater clarity and quantification.

We hope that the approaches to forecasting that we developed will prove useful. However, we acknowledge that our project is but one small experimental endeavor in relation to an enormous intelligence establishment with entrenched practices that is slow to change. Nevertheless, we see potential value not only in forecasting world events for intelligence agencies and governmental policy makers, but innumerable private organizations that must make important strategic decisions based on forecasts of future states of the world. Hedge funds want to forecast political trends that could affect commodity prices. Investors need to forecast government policies that could affect investment returns. And nonprofits need to forecast the economic conditions and tax policies that will affect donors' contributions.

Final Word

There has been a conspicuous shortage of rigorous field tests of calibration in confidence judgments (Griffin and Brenner 2004, Koehler et al. 2002). Our study employed forecasting questions that were of enormous practical importance and came from dedicated forecasters working outside the experimental laboratory. Lest our results be taken as some sort of redemption for expert judgment, which has taken quite a beating over the years (Camerer and Johnson 1991, Tetlock 2005), we must point out that our forecasters were not selected to be experts on the topics they were forecasting. They were educated citizens who worked to stay abreast of the relevant news, and what limited incentives we gave them for accuracy came in the form of feedback, a small monetary reward, and the social prestige of names on a leaderboard. In contrast to experts from academia, quoted in the media, and sold in book stores, the forecasters in our study had less to gain from grandiose claims and bold assertions. By contrast, what made our forecasters good was not so much that they always knew what would happen, but that they had an accurate sense of how much they knew. In the right context, it appears that confidence judgments can be well calibrated after all.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2016.2525 and https://osf.io/ecmk6/.

Acknowledgments

This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity via the U.S. Department of the Interior (DoI) National Business Center (NBC) [Contract D11PC20061]. 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 IARPA, DoI NBC, or the U.S. government.

References

Anderson C, Brion S, Moore DA, Kennedy JA (2012) A social-functional account of overconfidence. *J. Personality Soc. Psych.* 103(4):718–735.

Armor DA, Sackett AM (2006) Accuracy, error, and bias in predictions for real versus hypothetical events. *J. Personality Soc. Psych.* 91(4):583–600.

Armstrong JS (2001) *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Kluwer Academic, Boston).

Atanasov P, Rescober P, Stone E, Swift SA, Servan-Schreiber E, Tetlock P, Unger L, Mellers B (2016) Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Sci.*, ePub ahead of print April 22, http://dx.doi.org/10.1287/mnsc.2015.2374.

Baron J, Ungar L, Mellers BA, Tetlock PE (2014) Two reasons to make aggregated probability forecasts more extreme. *Decision Anal.* 11(2):133–145.

Bazerman MH, Moore DA (2013) *Judgment in Managerial Decision Making*, 8th ed. (John Wiley & Sons, New York).

Ben-David I, Graham JR, Harvey CR (2013) Managerial miscalibration. *Quart. J. Econom.* 128(4):1547–1584.

Benson PG, Onkal D (1992) The effects of feedback and training on the performance of probability forecasters. *Internat. J. Forecasting* 8(4):559–573.

Brier GW (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Rev.* 78(1):1–3.

Budescu DV, Wallsten TS, Au WT (1997) On the importance of random error in the study of probability judgment: Part II. Applying the stochastic judgment model to detect systematic trends. *J. Behavioral Decision Making* 10(3):173–188.

Buehler R, Messervey D, Griffin DW (2005) Collaborative planning and prediction: Does group discussion affect optimistic biases in time estimation. *Organ. Behav. Human Decision Processes* 97(1):47–63.

Burson KA, Larrick RP, Klayman J (2006) Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *J. Personality Soc. Psych.* 90(1):60–77.

Butler AC, Fazio LK, Marsh EJ (2011) The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bull. Rev.* 18(6):1238–1244.

Camerer CF, Johnson EJ (1991) The process-performance paradox in expert judgment: How can experts know so much and predict so badly? Ericsson KA, Smith J, eds. *Toward a General Theory of Expertise: Prospects and Limits* (Cambridge University Press, Cambridge, UK), 195–217.

Chauvin C, Fischhoff B, eds. (2011) *Intelligence Analysis: Behavioral and Social Scientific Foundations* (National Academies Press, Washington, DC).

Clark J, Friesen L (2009) Overconfidence in forecasts of own performance: An experimental study. *Econom. J.* 119(534):229–251.

Cooper AC, Woo CY, Dunkelberg WC (1988) Entrepreneurs' perceived chances for success. *J. Bus. Venturing* 3(2):97–109.

Dawes RM, Mulford M (1996) The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? *Organ. Behav. Human Decision Processes* 65(3):201–211.

Deaves R, Lüders E, Schröder M (2010) The dynamics of overconfidence: Evidence from stock market forecasters. *J. Econom. Behav. Organ.* 75(3):402–412.

Dobbins GH, Zaccaro SJ (1986) The effects of group cohesion and leader behavior on subordinate satisfaction. *Group Organ. Management* 11(3):203–219.

Dunlosky J, Rawson KA (2012) Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learn. Instruction* 22(4):271–280.

Erev I, Wallsten TS, Budescu DV (1994) Simultaneous over- and underconfidence: The role of error in judgment processes. *Psych. Rev.* 101(3):519–527.

Fischhoff B (1982) Debiasing. Kahneman D, Slovic P, Tversky A, eds. *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, MA), 422–444.

Fischhoff B, Slovic P, Lichtenstein S (1977) Knowing with certainty: The appropriateness of extreme confidence. *J. Experiment. Psych.: Human Perception Performance* 3(4):552–564.

Flyvbjerg B, Holm MKS, Buhl SL (2002) Underestimating costs in public works projects: Error or lie? *J. Amer. Planning Assoc.* 68(3):279–295.

Gardner D (2010) *Future Babble: Why Expert Predictions Fail – and Why We Believe Them Anyway* (Random House, New York).

Gigerenzer G (1991) How to make cognitive illusions disappear: Beyond "heuristics and biases." *Eur. Rev. Soc. Psych.* 2(1): 83–115.

Gigerenzer G, Hoffrage U, Kleinbölting H (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psych. Rev.* 98(4):506–528.

Glaser M, Weber M (2007) Overconfidence and trading volume. *Geneva Risk Insurance Rev.* 32(1):1–36.

González-Vallejo C, Bonham A (2007) Aligning confidence with accuracy: Revisiting the role of feedback. *Acta Psychologica* 125(2):221–39.

Griffin DW, Brenner L (2004) Perspectives on probability judgment calibration. Koehler DJ, Harvey N, eds. *Blackwell Handbook of Judgment and Decision Making* (Blackwell, Malden, MA), 177–199.

Griffin DW, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Hall CC, Ariss L, Todorov A (2007) The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organ. Behav, Human Decision Processes* 103(2): 277–290.

Hansson P, Rönnlund M, Juslin P, Nilsson L-G (2008) Adult age differences in the realism of confidence judgments: Overconfidence, format dependence, and cognitive predictors. *Psych. Aging* 23(3):531–544.

Haran U, Moore DA, Morewedge CK (2010) A simple remedy for overprecision in judgment. *Judgment and Decision Making* 5(7):467–476.

Harvey N (1997) Confidence in judgment. *Trends Cognitive Sci.* 1(2):78–82.

Heath C, Gonzalez R (1995) Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making. *Organ. Behav. Human Decision Processes* 61(3):305–326.

Hoelzl E, Rustichini A (2005) Overconfident: Do you put your money on it? *Econom. J.* 115(503):305–318.

Jagacinski RJ, Isaac PD, Burke MW (1977) Application of signal detection theory to perceptual-motor skills: Decision processes in basketball shooting. *J. Motor Behav.* 9(3):225–234.

Janis IL, Mann L (1977) *Decision Making: A Psychological Analysis of Conflict, Choice, and Commitment* (Free Press, New York).

Juslin P, Winman A, Olsson H (2000) Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psych. Rev.* 107(2):384–396.

Kahneman D (2011) *Thinking Fast and Slow* (Farrar, Straus and Giroux, New York).

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: A cognitive perspective on risk and risk taking. *Management Sci.* 39(1):17–31.

Keren G (1987) Facing uncertainty in the game of bridge: A calibration study. *Organ. Behav. Human Decision Processes* 39(1): 98–114.

Keren G (1991) Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica* 77(3):217–273.

Kerr NL, MacCoun RJ, Kramer GP (1996) Bias in judgment: Comparing individuals and groups. *Psych. Rev.* 103(4):687–719.

Klayman J, Soll JB, Gonzalez-Vallejo C, Barlas S (1999) Overconfidence: It depends on how, what, and whom you ask. *Organ. Behav. Human Decision Processes* 79(3):216–247.

Koehler DJ, Brenner L, Griffin DW (2002) The calibration of expert judgment: Heuristics and biases beyond the laboratory. Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK), 686–715.

Koriat A, Sheffer L, Ma'ayan H (2002) Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *J. Experiment. Psych.: General* 131(2):147–162.

Kruger J (1999) Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *J. Personality Soc. Psych.* 77(2):221–232.

Kruger J, Dunning D (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Personality Soc. Psych.* 77(6):1121–1134.

Larrick RP (2004) Debiasing. Koehler DJ, Harvey N, eds. *Blackwell Handbook of Judgment and Decision Making* (Blackwell Publishers, Malden, MA), 316–337.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1):111–127.

Larrick RP, Burson KA, Soll JB (2007) Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organ. Behav. Human Decision Processes* 102(1):76–94.

Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psych. Bull.* 125(2):255–275.

Lichtenstein S, Fischhoff B (1980) Training for calibration. *Organ. Behav. Human Decision Processes* 26(2):149–171.

Lichtenstein S, Fischhoff B, Phillips LD (1977) Calibration of probabilities: The state of the art. Jungermann H, DeZeeuw G, eds. *Decision Making and Change in Human Affairs* (D. Reidel, Amsterdam), 275–324.

Mandel DR, Barnes A (2014) Accuracy of forecasts in strategic intelligence. *Proc. Natl. Acad. Sci. USA* 111(30):10984–10989.

Massey C, Simmons JP, Armor DA (2011) Hope over experience: Desirability and the persistence of optimism. *Psych. Sci.* 22(2):274–281.

McKenzie CRM, Liersch MJ, Yaniv I (2008) Overconfidence in interval estimates: What does expertise buy you? *Organ. Behav. Human Decision Processes* 107(2):179–191.

Mellers BA, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, Tetlock PE (2015a) The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *J. Experiment. Psych.: Appl.* 21(1):1–14.

Mellers BA, Stone ER, Murray T, Minster A, Rohrbaugh N, Bishop MM, Chen MM, et al. (2015b) Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspect. Psych. Sci.* 10(3):267–281.

Mellers BA, Ungar L, Baron J, Ramos J, Gurcay B, Fincher K, Scott SE, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psych. Sci.* 25(5):1106–1115.

Moore DA (2007) Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organ. Behav. Human Decision Processes* 102(1):42–58.

Moore DA, Healy PJ (2008) The trouble with overconfidence. *Psych. Rev.* 115(2):502–517.

Moscovici S, Zavalloni M (1969) The group as a polarizer of attitudes. *J. Personality Soc. Psych.* 12(2):125–135.

Murphy AH, Winkler RL (1977) Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *Natl. Weather Digest* 2:2–9.

Oskamp S (1965) Overconfidence in case-study judgments. *J. Consulting Psych.* 29(3):261–265.

Park YJ, Santos-Pinto L (2010) Overconfidence in tournaments: Evidence from the field. *Theory Decision* 69(1):143–166.

Prims J, Moore DA (2015) Overconfidence over the lifespan. Working paper, University of Illinois at Chicago, Chicago.

Satopää V, Baron J, Foster B, Mellers BA, Tetlock PE, Ungar L (2014a) Combining multiple probability predictions using a simple logit model. *Internat. J. Forecasting* 30(2):344–356.

Satopää V, Jensen B, Mellers BA, Tetlock PE, Ungar L (2014b) Probability aggregation in time series: Dynamic hierarchical modeling of sparse expert beliefs. *Ann. Appl. Statist.* 8: 1256–1280.

Sieck WR, Arkes HR (2005) The recalcitrance of overconfidence and its contribution to decision aid neglect. *J. Behav. Decision Making* 18(1):29–53.

Silver N (2012) *The Signal and the Noise: Why so Many Predictions Fail—But Some Don't* (Penguin Press, New York).

Simon M, Houghton SM (2003) The relationship between overconfidence and the introduction of risky products: Evidence from a field study. *Acad. Management J.* 46(2):139–150.

Sniezek JA, Henry RA (1989) Accuracy and confidences in group judgment. *Organ. Behav. Human Decision Processes* 4(3):1–28.

Soll JB, Milkman KL, Payne JW (2016) A user's guide to debiasing. Wu G, Keren G, eds. *Handbook of Judgment and Decision Making* (John Wiley & Sons, New York), 924–951.

Stasser G, Davis JH (1981) Group decision making and social influence: A social interaction sequence model. *Psych. Rev.* 88(6):523–551.

Stasser G, Titus W (1987) Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *J. Personality Soc. Psych.* 53:81–93.

Statman M, Thorley S, Vorkink K (2006) Investor overconfidence and trading volume. *Rev. Financial Stud.* 19(4):1531–1565.

Stone ER, Opel RB (2000) Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organ. Behav. Human Decision Processes* 83(2):282–309.

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

Tetlock PE, Gardner D (2015) *Superforecasting: The Art and Science of Prediction* (Signal, New York).

Tetlock PE, Mellers BA (2011) Intelligent management of intelligence agencies: Beyond accountability ping-pong. *Amer. Psychologist* 66(6):542–554.

Tetlock PE, Mellers BA, Rohrbaugh N, Chen E (2014) Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions Psychological Sci.* 23:290–295.

Tversky A, Koehler DJ (1994) Support theory: A nonextensional representation of subjective probability. *Psych. Rev.* 101(4): 547–567.

Wells GL, Olson EA (2003) Eyewitness testimony. *Annual Rev. Psych.* 54:277–295.

Wright G, Rowe G, Bolger F, Gammack J (1994) Coherence, calibration, and expertise in judgmental probability forecasting. *Organ. Behav. Human Decision Processes* 57(1):1–25.

Yates JF (1982) External correspondence: Decompositions of the mean probability score. *Organ. Behav. Human Decision Processes* 30(1):132–156.

Yates JF, Lee JW, Shinotsuka H, Patalano AL, Sieck WR (1998) Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organ. Behav. Human Decision Processes* 74(2):89–117.