Confidence Calibration in a Multi-year Geopolitical Forecasting Competition

Don A. Moore
University of California, Berkeley

Samuel A. Swift
Betterment, Inc.

Angela Minster, Barbara Mellers, Lyle Ungar, Philip Tetlock
University of Pennsylvania

Heather H. J. Yang
Massachusetts Institute of Technology

Elizabeth R. Tenney
University of Utah

Draft of April 16, 2015

Corresponding author
Don A. Moore
University of California, Berkeley
545 Student Services Building #1900
Berkeley, CA 94720-1900
don.moore@alumni.carleton.edu

Author Note
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 (IARPA) via the Department of Interior National Business Center 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.

Abstract

This research examines the development of confidence and accuracy over time in the context of geopolitical intelligence 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 in ways that would help forecasters be more accurate. Forecasts began well-calibrated: forecasters' confidence roughly matched their accuracy. Over time, as information came in, forecasts became more accurate. Confidence increased at roughly the same rate as accuracy, and good calibration persisted. Nevertheless, there was evidence of a small amount of overconfidence (3%). Training helped reduce overconfidence and having forecasters work together on teams 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

Confidence Calibration in a Multi-year Geopolitical Forecasting Competition

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 facilitating other biases and errors (Bazerman & Moore, 2013; Fischhoff, 1982; Kahneman, 2011). Research has identified overconfidence in tests of declarative knowledge, bets, and predictions of the future (Ben-David, Graham, & Harvey, 2013; Fischhoff, Slovic, & Lichtenstein, 1977; Massey, Simmons, & Armor, 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 ameliorated.

On the one hand, Tetlock's (2005) long-term survey of political experts suggests pessimism. The experts in his sample were persistently overconfident. Although they clearly believed they had expertise relevant to making accurate forecasts, 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 some of the conditions that evidence suggests might be important for debiasing overconfidence, including incentives rewarding accuracy, training in the use and interpretation of probability scales, practice, and feedback that is frequent, timely, and unambiguous (Benson & Onkal, 1992; Hoelzl & Rustichini, 2005; Larrick, 2004). These conditions have each proven at least somewhat successful in reducing overconfidence, but effects have generally been studied over short time horizons, usually constrained by the duration of laboratory experimental sessions (Soll, Milkman, & Payne, 2015). There are legitimate

questions about the degree to which these results generalize over longer time horizons and in more consequential domains (Dawes & Mulford, 1996; Gigerenzer, 1991; Juslin, Winman, & Olsson, 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 of the United States federal government (IARPA). Our team was one of five research teams that provided 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 is accompanied by a confidence judgment that reflects how sure the forecaster is that he or she knows what is going to happen. We examine these judgments for the presence of overconfidence and see how that changes with time and experience.

The tournament was IARPA's attempt to improve geopolitical forecasting and intelligence analysis. Current systems rely primarily on qualitative assessments of probabilities and risk (Mandel & Barnes, 2014). Qualitative probability estimates are difficult to score, aggregate, and analyze. They limit accountability because it is unclear what constitutes a good forecast. And they limit usefulness because qualitative forecasts cannot be fed into expected value calculations that could inform policy decisions by helping estimate likely consequences. IARPA's forecasting tournament was designed to serve as an important proof of the viability of quantitative scored forecasting. IARPA scored each research team's forecasts using 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. 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

There have been a number of studies of training. One of the more ambitious ones provided participants with 45 minutes of training on the calibration of confidence judgments, followed by 22 testing sessions, each an hour long (Lichtenstein & Fischhoff, 1980). Another study employed six testing sessions, each 2 hours long (Stone & Opel, 2000). These studies 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, we wanted to test the potential for training to endure over a year on a diverse set of forecasting questions across many different domains.

The role of group interaction

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, Messervey, & Griffin, 2005). This seems to be due, at least in part, to the potential for discussion to polarize attitudes (Moscovici & Zavalloni, 1969). However, when discussion shares useful information it can increase accuracy (Stasser & Davis, 1981). Furthermore, an increase in perceived accountability to the group can increase self-critical reflection and help reduce overconfidence (Lerner & Tetlock, 1999; Sniezek & Henry, 1989). 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.

Mellers and colleagues (2014) present data on forecast accuracy across our experimental conditions, as measured by Brier scores, from the first two years of the tournament. At the end of the second year, our team's accuracy was so far superior to that of the other four teams that the project sponsor (IARPA) elected to cut funding to all four of the other teams. Our team was the only one that continued into the third year of forecasting. The present paper examines data from all 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?

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 & Weber, 2007; Park &

Santos-Pinto, 2010). The few longitudinal studies suffer from sporadic sampling and relatively few judgments (Ben-David et al., 2013; Dunlosky & Rawson, 2011; Simon & 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, Larrick, & Klayman, 2006; Kruger & Dunning, 1999). However, some kinds of information increase confidence without increasing accuracy and vice versa, even for experts (Griffin & 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, Lüders, & Schröder, 2010). And then, 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, Liersch, & Yaniv, 2008).

To preview our results, we find that forecasters making predictions about real world events can be remarkably well calibrated when they are evaluated on prediction accuracy. 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 are analyzed by Mellers et al. (2015) in more detail than we can do justice to here.

Participants who stuck with it for the entire year 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 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 PhDs, 37% had masters degrees, 36% had only an undergraduate education, and less than 1% had not graduated from college; 78% were US citizens.

Materials

Questions. A total of 344 specific questions, created by IARPA, had resolved by the end of Year 3 and were included in the present analyses. A list of the questions appears at http://learnmoore.org/CAC/. New questions were released roughly every week in batches of about four or five. Each question was open for between 1 to 549 days (M = 114), during which forecasters could update their forecasts as frequently as they wished. The average forecaster submitted a forecast 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*, or *neither*. Like this one, there were 27 multinomials that asked about 3 outcomes, 31 that asked about 4, and 9 that asked about 5.
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, Moore, & Morewedge, 2010; Tversky & 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 multi-dimensional and reflects a number of different components, which it is useful to decompose (Yates, 1982). In this paper, we focus on calibration. For each question, we identify the outcome that the forecaster reported to be most likely and took the associated probability as the forecaster's confidence. In order 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.

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 assigned participants to one of four conditions in a 2 (individual vs. team) X 2 (no training vs. 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.

In the individual condition, forecasters worked alone and did not interact with one another.

In the team condition, forecasters were assigned to groups of up to 25. 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

[1] We omit discussion of other conditions because their data are not as well suited to the analyses we run: a prediction-market condition in which people bet against each other in a market and a crowd-prediction condition in which people knew the consensus forecast when they made their own, which only existed in Year 1. Moreover, we omit discussion of a scenario-training condition that was only used in Year 1. For more information about these other conditions, see Mellers et al. (2014). For more detail about the prediction-market conditions, see Atanasov et al. (2015).

disagreed. Examples of the suggestions we gave to forecasters in the team condition can be found in the online supplement (http://learnmoore.org/CAC/).

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 full details of this training are available in the supplementary materials.[2]

Leaderboard and Incentives. Brier scores, averaged across questions, determined the order in which individual forecasters' chosen user names appeared on leaderboards that ranked forecasters within each condition and were visible to all forecasters in that condition. Scores were posted after the first 10 questions closed and were updated every time a question closed after that, providing forecasters with regular feedback.

For teams, the forecasters' chosen team names appeared on leaderboards that were visible to all teams. Team members received their own individual Brier scores privately, but only team scores were posted to the leaderboard.

In Years 1 and 2, in addition to an outcome leaderboard, teams had access to a process leaderboard, updated twice annually, which showed who the top contributors were in terms of engagement with their teams (e.g., most number of comments). Two experimenters also selected

[2] http://learnmoore.org/CAC/ Note that in re-assigning 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 the paper.

exemplary comments (based on the experimenters' perceptions of helpfulness to the team) and posted them to this board.

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 others had. Note that these imputed scores are not part of any of the results we report in this paper.

Individuals in the team condition who declined to forecast on a particular question received the median score from all of 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 forecasters were quite well-calibrated and exhibited only a small amount 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, Fischhoff, & Phillips, 1977), and the 9% overconfidence estimated in Juslin, Winman, and Olsson's (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 were .13. The average Brier score of forecasters in the last week of forecasting on each question was .14. For the last day of forecasting, it was .10.[3]

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, Wallsten, & Budescu, 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% 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. This result raises questions about the origins of these extreme forecasts. Figure 1 makes it clear that the most extreme 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. 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 this paper's supplementary materials (http://learnmoore.org/CAC/).

[3] Lower Brier scores indicate better accuracy.

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, Rowe, Bolger, & Gammack, 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 4, which shows that self-reported expertise was not a reliable moderator of the relationship 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 5-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.

This is surprising because two different varieties of overconfidence appear to be at odds with one another. Forecasters exhibited underconfidence, underplacing themselves relative to other forecasters with regard to their relative expertise, even while they exhibited overprecision by overestimating the probability that their forecasts were correct. This replicates other results showing that “better than average” beliefs are far from ubiquitous and are routinely unrelated to other forms of overconfidence (Moore & Healy, 2008; Moore, 2007). 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 each year (Year 1: M = 2.18, SD = 0.92, n = 141,186; Year 3: M = 2.69, SD = 1.14, n = 203,553).

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 lock-step. 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 5 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. But there was also a persistent gap between them: confidence systematically exceeded accuracy by a small but persistent amount.

While we do find that calibration increased from the beginning of a tournament year to the end, we do not find that having more years of forecasting experience (forecasting tenure) leads to an improvement in calibration. Figure 6 shows that the calibration curves of forecasters with one, two or three tournament years of forecasting experience showed about the same level of calibration. Statistically, through analysis of variance, we find no significant differences in calibration between forecasters with more or less experience.

Variation by Experimental Treatment

In support of some evidence suggesting that groups can reduce overconfidence (Sniezek & Henry, 1989), we find that forecasters in the team condition were even better calibrated than those in the solo forecasting condition. See Table 1. Working on teams significantly improves accuracy and slightly reduces overconfidence. Training, for its part, slightly improves accuracy but mostly improves calibration by reducing confidence.

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. 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. When forecasting on a particular question, all forecasters had to specify the probabilities of the full set of mutually exclusive possible outcomes. All forecasters got frequent feedback using an incentive-compatible scoring rule (i.e., the Brier scores).

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 get feedback on how accurate their forecasts are; 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. The results also reveal an interesting pattern in the development of confidence over time. As our participants gained information, their confidence increased and accuracy improved. Indeed, the parallel increases in both confidence and accuracy may be the single most remarkable feature of the results we present.

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. 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. So long as confidence goes up at the same rate as accuracy,

then good calibration will persist. Although our results do find evidence of overconfidence, the overall effect is smaller than in prior studies.

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 & Winkler, 1977). The unique conditions of our forecasting tournament are, no doubt, key to our forecasters' performance. The fact that their forecasts would be scored against a clear standard for accuracy was undoubtedly crucial (Armor & Sackett, 2006; Clark & Friesen, 2009). It is also likely that our forecasters felt accountable to us and to each other, especially in the team condition (see Lerner & Tetlock, 1999). We strongly suspect that the quality and regularity of feedback is likely to have been important (Butler, Fazio, & Marsh, 2011; González-Vallejo & Bonham, 2007; Lichtenstein & 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 real life provides the kind of practice, clarity, and prompt feedback we provided our forecasters, we have reason to believe that calibration in everyday life should look more like what we observe in our study and less like what we might expect based on short-term lab studies. 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 & Arkes, 2005; Yates, Lee, Shinotsuka, Patalano, & Sieck, 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 & Mulford, 1996; Gigerenzer, 1991; Juslin et al., 2000). Still others maintain that overconfidence cannot be explained away so easily (Budescu, Wallsten, & Au, 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 and colleagues (2000). Moreover, we can say little about what it was about training or about team collaboration that helped. Both our manipulations are complicated, including a variety of different components. Our experimental design does not allow us to distinguish the relative influence of these different manipulations or how they might interact with one another. 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., 2015). If feedback and experience were essential to our forecasters' performance, then their calibration should have improved over time as they gained experience, but 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.

Proof of Concept

The good calibration of our forecasters offers a hopeful sign for the quantifiability of intelligence forecasts. One striking feature of most formal intelligence reports is how rarely they contain quantified estimations of probabilities (Chauvin & 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 the scorability of quantified forecasts 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 & 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 27th, 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 or slander, 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 scorability.

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

Lest our results be taken as some sort of redemption for expert judgment, which has taken quite a beating over the years (Camerer & Johnson, 1997; 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, Brier scores, and their 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 if they later turned out to be wrong. 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.

References

Armor, D. A., & Sackett, A. M. (2006). Accuracy, Error, and Bias in Predictions for Real Versus Hypothetical Events. *Journal of Personality and Social Psychology*, *91*(4), 583–600.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*. Boston, MA: Kluwer Academic.

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P. E., … Mellers, B. A. (2015). Distilling the wisdom of crowds: Prediction markets versus prediction polls. *Unpublished Manuscript*.

Bazerman, M. H., & Moore, D. A. (2013). *Judgment in managerial decision making* (8th ed.). New York: Wiley.

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *Quarterly Journal of Economics*.

Benson, P. G., & Onkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*(4), 559–573. doi:10.1016/0169-2070(92)90066-I

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

Budescu, D. V, Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment: Part II. Applying the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*(3), 173–188.

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

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*(1), 60–77.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin and Review*, *18*(6), 1238–1244.

Camerer, C. F., & Johnson, E. J. (1997). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In W. M. Goldstein & R. M. Hogarth (Eds.), *Research on judgment and decision making: Currents, connections, and controversies* (p. 342). Cambridge: Cambridge University Press.

Chauvin, C., & Fischhoff, B. (Eds.). (2011). *Intelligence analysis: Behavioral and social scientific foundations*. Washington, DC: National Academies Press.

Clark, J., & Friesen, L. (2009). Overconfidence in Forecasts of Own Performance: An Experimental Study. *Economic Journal*, *119*(534), 229–251.

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

Deaves, R., Lüders, E., & Schröder, M. (2010). The dynamics of overconfidence: Evidence from stock market forecasters. *Journal of Economic Behavior & Organization*, *75*(3), 402–412. doi:http://dx.doi.org/10.1016/j.jebo.2010.05.001

Dunlosky, J., & Rawson, K. A. (2011). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271–280.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519–527.

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

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

Gardner, D. (2010). *Future babble: Why expert predictions fail-and why we believe them anyway*. New York: Random House.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases." *European Review of Social Psychology*, *2*(1), 83–115.

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

González-Vallejo, C., & Bonham, A. (2007). Aligning confidence with accuracy: revisiting the role of feedback. *Acta Psychologica*, *125*(2), 221–39. doi:10.1016/j.actpsy.2006.07.010

Griffin, D. W., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, *24*(3), 411–435.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making*, *5*(7), 467–476.

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

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

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

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus and Giroux.

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

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

Larrick, R. P. (2004). Debiasing. In D. J. Koehler & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*. Oxford, England: Blackwell Publishers.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255–275.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Decision Processes, 26*(2), 149–171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. DeZeeuw (Eds.), *Decision making and change in human affairs*. Amsterdam: D. Reidel.

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences, 111*(30), 10984–10989. doi:10.1073/pnas.1406138111

Massey, C., Simmons, J. P., & Armor, D. A. (2011). Hope over experience: Desirability and the persistence of optimism. *Psychological Science, 22*.

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior & Human Decision Processes, 107*, 179–191.

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

Mellers, B. A., 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*(4).

Moore, D. A. (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. *Organizational Behavior and Human Decision Processes*, *102*(1), 42–58.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*, 125–135.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest*, *2*, 2–9.

Oskamp, S. (1965). Attitudes toward U S and Russian actions: A double standard. *Psychological Reports*, *16*(1), 43–46.

Park, Y. J., & Santos-Pinto, L. (2010). Overconfidence in tournaments: evidence from the field. *Theory and Decision*, *69*(1), 143–166.

Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral 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.

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

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and cofidences in group judgment. *Organizational Behavior and Human Decision Processes*, *4*(3), 1–28.

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of Judgment and Decision Making*. New York: Wiley.

Stasser, G., & Davis, J. H. (1981). Group decision making and social influence: A social interaction sequence model. *Psychological Review*, *88*(6), 523–551.

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, *83*(2), 282–309.

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. *The American Psychologist, 66*(6), 542–54. doi:10.1037/a0024285

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547–567.

Wright, G., Rowe, G., Bolger, F., & Gammack, J. (1994). Coherence, calibration, and expertise in judgmental probability forecasting. *Organizational Behavior and Human Decision Processes, 57*(1), 1–25. doi:10.1006/obhd.1994.1001

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior & Human Decision Processes, 30*(1), 132–156.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*(2), 89–117.

Table 1

Working on teams primarily improves accuracy, while training primarily reduces overconfidence. Overconfidence measures with different subscripts are significantly different from one another; these significance groupings are calculated using Bonferroni-adjusted 95% confidence intervals.

Elicitation	Training	Confidence	Hit rate	Overconfidence
Autonomous	None	64.9%	60.1%	4.3% $_{a}$
Autonomous	Training	63.5%	61.5%	2.0% $_{b}$
Team	None	65.8%	62.9%	2.8% $_{a,b}$
Team	Training	64.3%	63.6%	0.6% $_{b}$

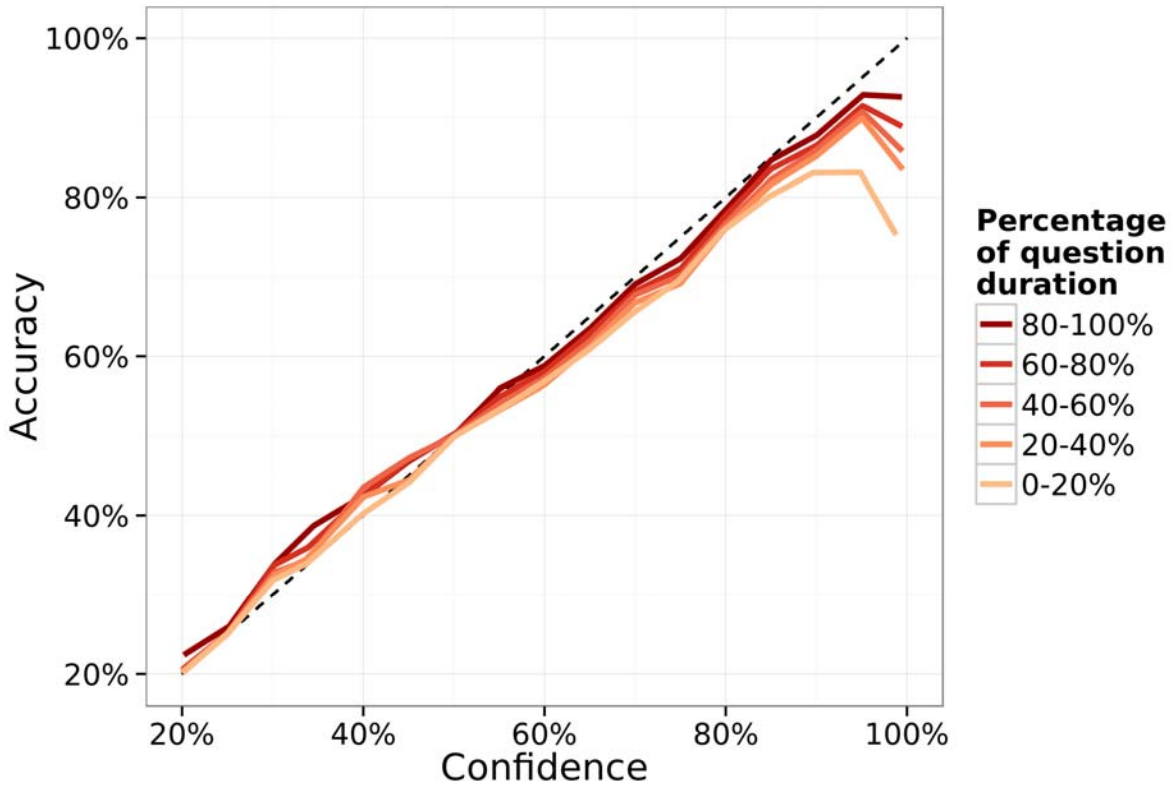

Figure 1. Calibration curves, conditional on when in the question's life the forecast was made.

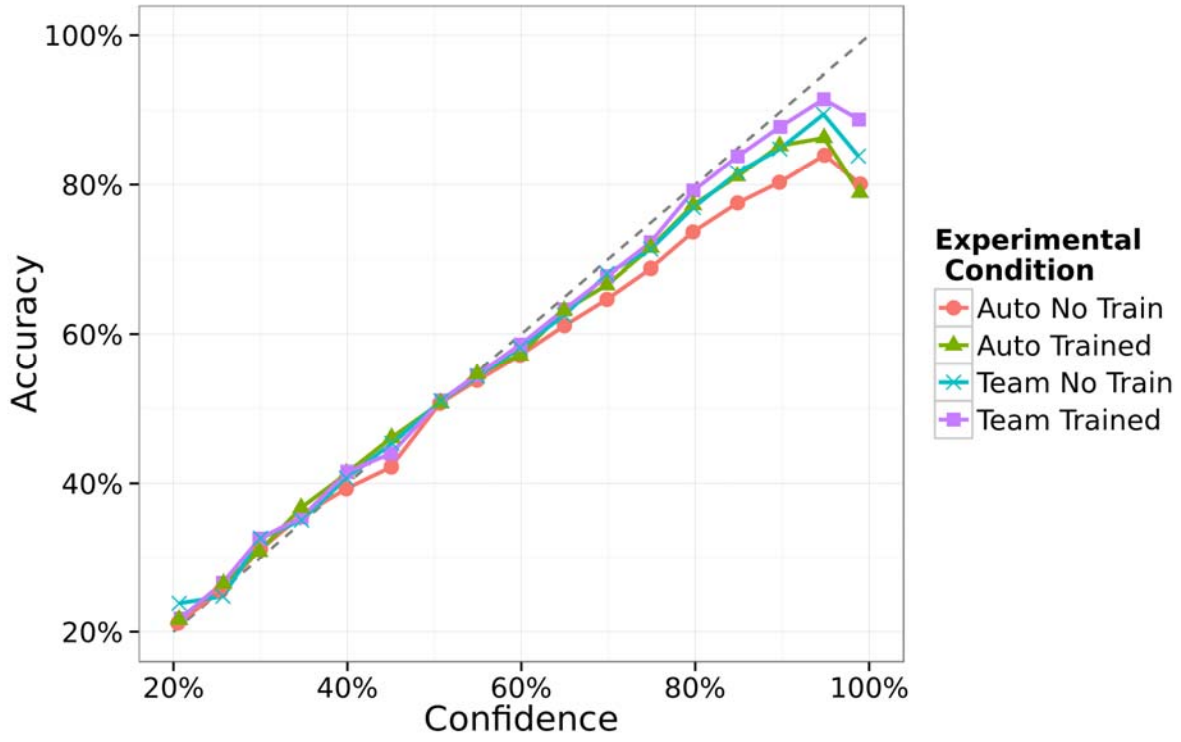

Figure 2. Confidence and accuracy curves as a function of experimental condition (Individual (i.e., “Auto” for “autonomous”) vs. Team X Training vs. No Training).

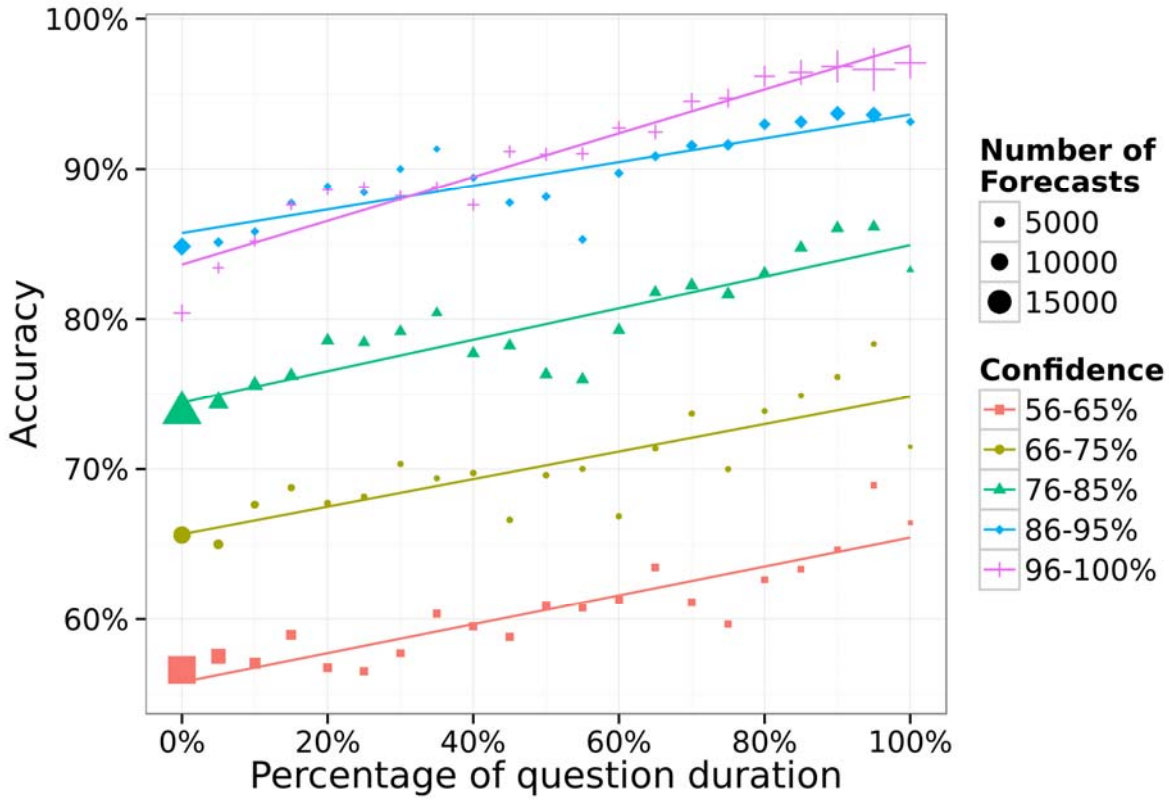

Figure 3. Accuracy, expressed in hit rates, as a function of the forecast confidence and when the forecast was made during the duration of a question.

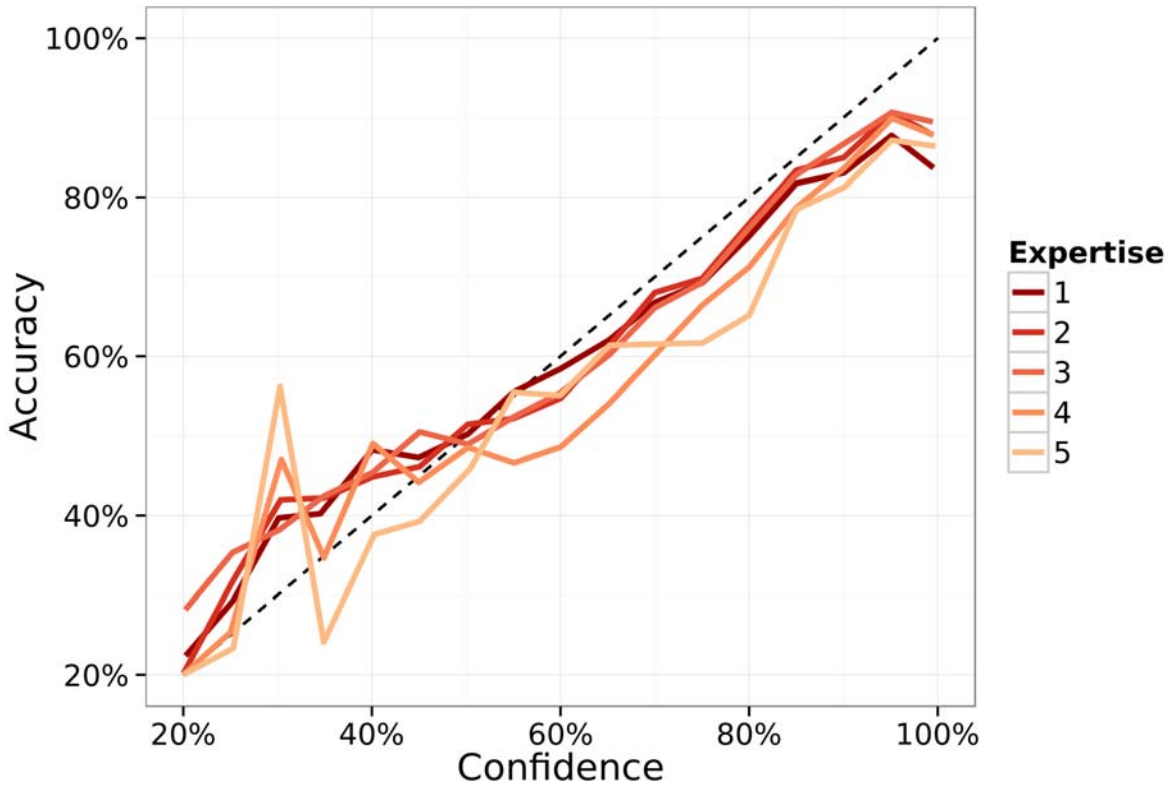

Figure 4. Confidence and accuracy as a function of forecasters' self-rated expertise on the question.

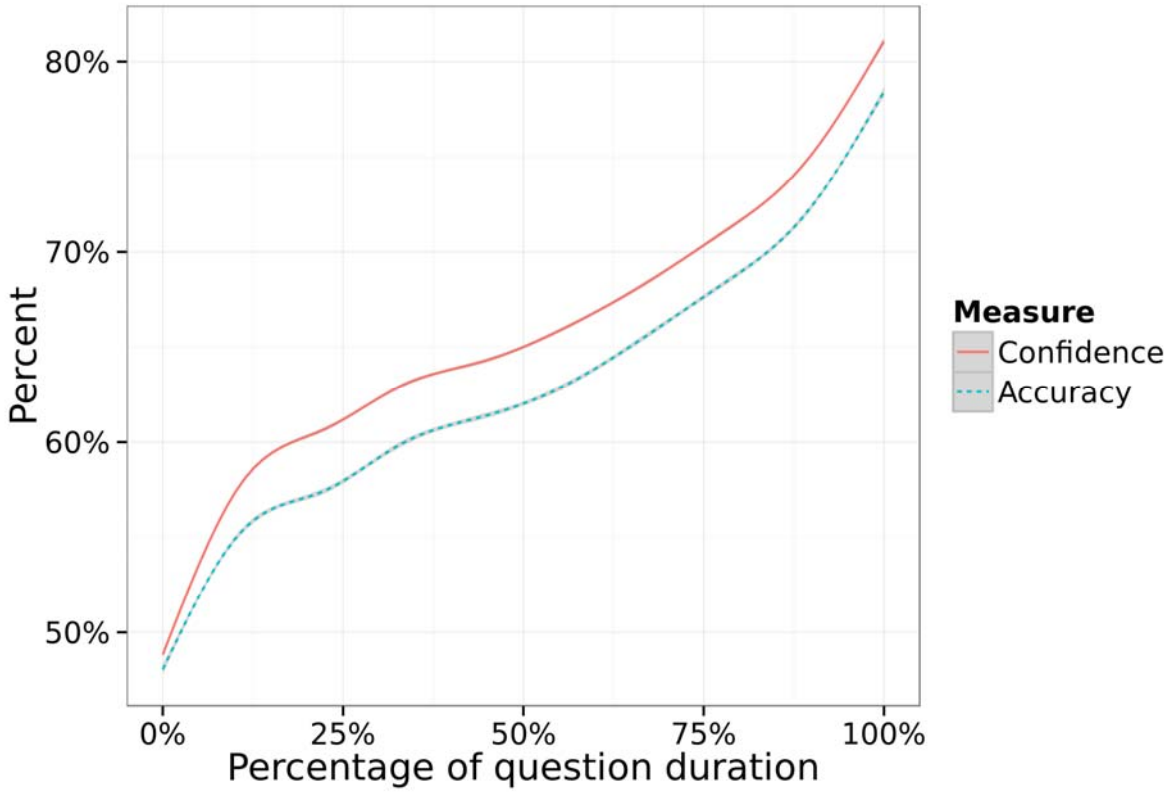

Figure 5. The persistent gap between confidence and accuracy over the duration of forecasting questions.

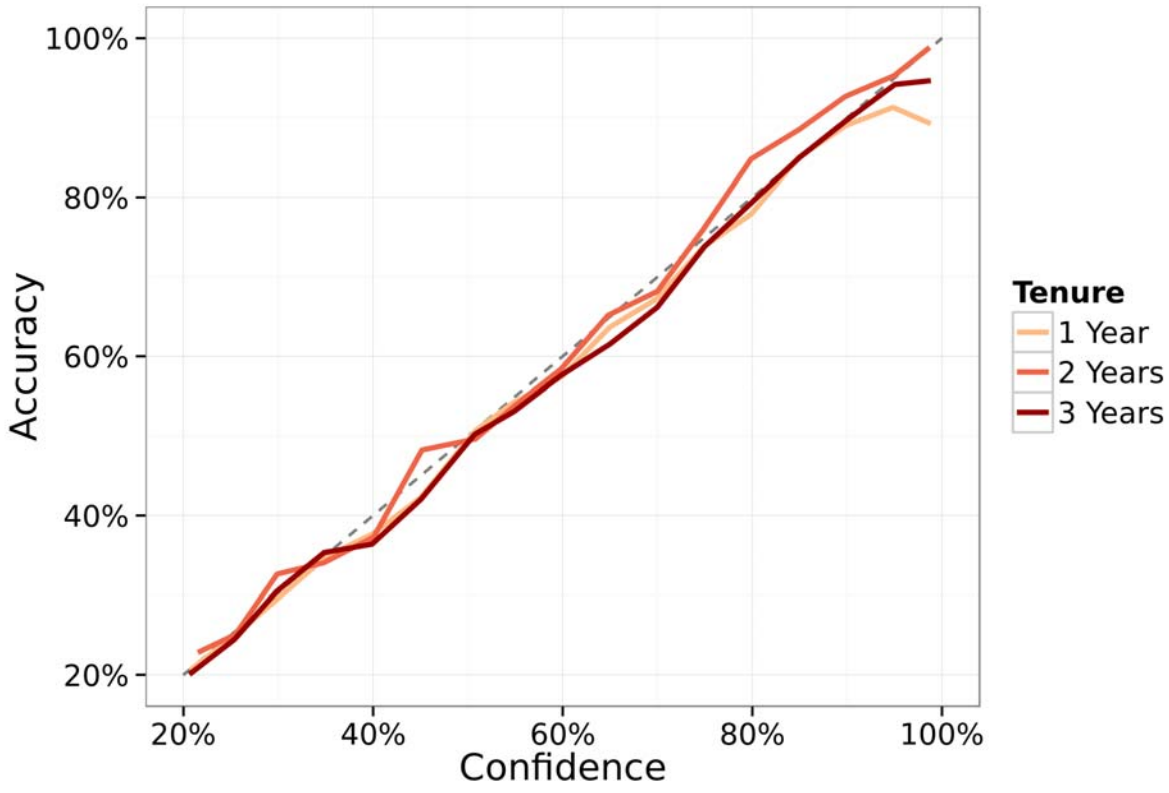

Figure 6. Confidence and accuracy as a function of forecasters' years of experience in the tournament (tenure).