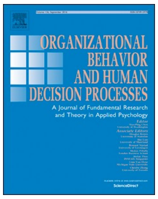
Organizational Behavior and Human Decision Processes 160 (2020) 19–35

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp



# Small steps to accuracy: Incremental belief updaters are better forecasters

Pavel Atanasov[a,*], Jens Witkowski[b], Lyle Ungar[c], Barbara Mellers[c], Philip Tetlock[c]

[a] *Pytho LLC, United States*
[b] *Frankfurt School of Finance & Management, Germany*
[c] *University of Pennsylvania, United States*

ARTICLE INFO

*Keywords:*
Belief revision
Belief updating
Prediction
Forecasting

ABSTRACT

Laboratory research has shown that both underreaction and overreaction to new information pose threats to forecasting accuracy. This article explores how real-world forecasters who vary in skill attempt to balance these threats. We distinguish among three aspects of updating: frequency, magnitude, and confirmation propensity. Drawing on data from a four-year forecasting tournament that elicited over 400,000 probabilistic predictions on almost 500 geopolitical questions, we found that the most accurate forecasters made frequent, small updates, while low-skill forecasters were prone to confirm initial judgments or make infrequent, large revisions. High-frequency updaters scored higher on crystallized intelligence and open-mindedness, accessed more information, and improved over time. Small-increment updaters had higher fluid intelligence scores, and derived their advantage from initial forecasts. Update magnitude mediated the causal effect of training on accuracy. Frequent, small revisions provided reliable and valid signals of skill. These updating patterns can help organizations identify talent for managing uncertain prospects.

## 1. Introduction

Can patterns of belief updating help organizations identify individuals with superior predictive judgment? Should organizations be more concerned that forecasters, analysts, and executive decision makers will underreact or overreact to the daily flow of new information across their desks? The literature on judgment and choice provides conflicting answers. The theoretical and empirical grounds for expecting underreaction to be more detrimental to judgment quality include Bayesian conservatism (Edwards, 1968) and the anchoring heuristic (Tversky & Kahneman, 1974). They describe a common tendency to update estimates in the right direction, but not far enough. Reasons for thinking overreaction poses the more serious problem include the availability (Tversky & Kahneman, 1973) and the representativeness heuristics (Kahneman & Tversky, 1972). These heuristics focus our attention on recent, memorable, and distinctive features (Bar-Hillel, 1980; Kahneman & Tversky, 1973), and have been cited as reasons behind suboptimal real-world behavior, such as excessive volatility in financial markets (Arrow, 1982; De Bondt & Thaler, 1985; Shiller, 1981).

In fact, we know relatively little about how real-world forecasters balance the threats of updating too little versus too much in naturalistic settings. A deeper understanding of belief updating could help organizations identify skilled forecasters, planners, advisors, and decision makers, all of whom need to grapple with evolving uncertainties about possible futures. Bayes' Theorem provides the normative standard for belief revision in simple settings and serves as a basis for normative models in more complex environments (e.g., Birnbaum, 1983).

Belief updating is an integral part of decision making in organizations, in contexts such as auditing (Ashton & Ashton, 1988; Hogarth, 1991), organizational expectation setting (Ehrig, 2015), and strategic decision making. Entrepreneurs face choices that rely on implicit predictions, such as choosing if, when, and how to pivot from one strategy to another (Ries, 2011). The propensity of leaders to update their beliefs and change their minds is an oft-debated aspect of business acumen (Pittampalli, 2016). Amazon founder Jeff Bezos has spoken publicly about the benefits of frequent belief revisions, noting that people who are *right a lot of the time are those who often change their minds* (Fried, 2012).

Does Bezos' heuristic help us identify those who are right about the future *a lot of the time*? To answer this question, we must first consider the extent to which forecasting performance is a matter of skill or luck. Mellers, Stone, Atanasov, et al. (2015) demonstrated that prediction accuracy was due in part to skill; past performance reliably predicted future performance. Furthermore, skillful forecasterstended to have

[*] Corresponding author at: 866 President Street, Brooklyn, NY 11215, United States.
*E-mail address:* pavel@pytho.io (P. Atanasov).

https://doi.org/10.1016/j.obhdp.2020.02.001
Received 19 February 2019; Received in revised form 15 September 2019; Accepted 13 February 2020
Available online 30 March 2020

specific cognitive and personality profiles. Atanasov et al. (2017) found that weighting predictions based on forecasters' past performance improved the accuracy of prediction polls, which feature direct probability elicitation and algorithmic aggregation, and enabled polls to outperform prediction markets.[1,2] Small crowds selected based on past performance can also outperform larger, less selective crowds (Goldstein, McAfee, & Suri, 2014; Mannes, Soll, & Larrick, 2014). Thus, past accuracy is useful in choosing how to weight opinions.

Historical accuracy data, however, is not always available (Witkowski, Atanasov, Ungar, & Krause, 2017). For example, newly hired political or financial analysts need time to build a track record. In the meantime, their managers may have little information on which to judge the quality of the analysts' forecasts. Furthermore, some forecasting questions, such as those regarding the impact of climate change or the emergence of disruptive technologies, only resolve far in the future. Alternative markers of forecasting skill, available earlier than historical accuracy, would be especially helpful in these settings.

Identifying those with better foresight is also useful to advice seekers. The extensive literature on advice taking and advice giving (Bonaccio & Dalal, 2006; Yaniv, 2004), has investigated the number of advisors one should ask and how to weight their inputs versus one's own prior judgments (Larrick & Soll, 2006). Advice seekers use the quality of past recommendations (Redd, 2002) and information about advisors' prediction strategies (Yates, Price, Lee, & Ramirez, 1996) when choosing how much to weight their advice. The updating patterns of potential advisors may serve as useful cues for judging the advisors' predictive skill.

We propose that belief updating patterns can be useful for identifying skilled forecasters and advisors, especially when there is little information about historical accuracy. But which patterns indicate high predictive skill?

We had the opportunity to examine the relationship between belief updating and prediction skill in a naturalistic environment with thousands of probability forecasts about real-world political events over four years. We characterize belief updating as the process of adjusting probabilistic forecasts by incorporating new information over time. Instead of taking a one-dimensional view of updating and assessing whether the best forecasters update too little or too much, we distinguish among three aspects of belief updating. First, forecasters can differ in the *frequency* with which they submit new probability estimates on a question. Second, they can vary in the absolute *magnitude* of their updates on the probability scale. Third, they can differ in their propensity to change their beliefs versus actively *confirm* their prior forecasts by simply restating their preceding judgments. We find that all three aspects of updating are predictive of forecaster accuracy.

Mellers, Stone, Atanasov, et al. (2015), and Chang, Chen, Mellers, and Tetlock (2016) previously examined the relationship between updating frequency and accuracy in forecasting tournaments. Frequent belief updaters, i.e. forecasters who revised their estimates more often, tended to be more accurate. Tetlock and Gardner (2015) discussed examples when incremental updating was associated with higher accuracy. The current work is the first to empirically examine the relationships among update magnitude, confirmation propensity, and predictive skill.

## 2. Theory and hypotheses

We present three hypotheses about how three aspects of belief updating—frequency, magnitude and confirmation propensity—relate to forecasting skill. These hypotheses are tested simultaneously, accounting for the effects of other predictors. We also pose three additional research questions focused on the reliability, predictability and malleability of updating behavior.

**Hypothesis 1.** Forecasters who make frequent updates tend to be more accurate.

This hypothesis was supported in prior work (Mellers, Stone, Atanasov, et al., 2015). We build on prior work by asking whether frequent updaters generate relatively accurate initial judgments or outperform only by improving their forecasts over time.

**Hypothesis 2.** Forecasters who make smaller belief updates tend to be more accurate.

There are compelling reasons to expect that small-increment (i.e. incremental) updates predict accuracy, but there are also good reasons to expect the opposite. To test this hypothesis, we examine how observed update magnitude related to accuracy, and perform counterfactual simulations. That is, we examine whether forecasters who make smaller revisions are more accurate than those who make larger ones. Using simulation, we further investigate whether forecasters would have done better had they made smaller or larger updates.

Let us first examine the hypothesis that larger updates predict greater accuracy. Research suggests that people generally update too little in the face of new evidence. Edwards (1968) describes belief updaters as conservative Bayesians, who update in the correct direction, but do so insufficiently. In this account, forecasters making larger updates would be seen as better Bayesians, which may help them achieve accuracy advantages over time.

The alternative is that smaller updates are associated with better accuracy. This pattern may hold because people are often bombarded by information. Overreaction to new data could result in excessive volatility and degrade accuracy. One such example is the dilution effect or the tendency to discount valid cues as more and more non-diagnostic information appears (Tetlock & Boettger, 1989). Market overreaction is a common occurrence (De Bondt & Thaler, 1985)[3] and could be due to the discounting of stable cues (e.g. base rates) in favor of noisy inside-view cues (e.g. case-specific information), especially when the inside cues are extreme (Griffin & Tversky, 1992).[4] Institutional practices may contribute as well. For example, U.S. intelligence training emphasizes the need to avoid underreaction to new evidence, which could increase the risks of overreaction (Chang & Tetlock, 2016), and bring about advantages to incremental updaters, i.e. forecasters who tend to make smaller revisions, and may be less prone to overreact to new information. .

Another reason to expect an association between smaller updates and higher accuracy is that incremental updaters might be more accurate from the start (Massey & Wu, 2005). Forecast updates provide signals about forecasters' metabeliefs: a small update represents a vote of confidence in one's previous forecast. If forecasters believed their prior forecasts were already accurate, they would see less need for large

[1] In contrast, Chen, Chu, Mullen, and Pennock (2005) found no benefit to weighting individuals based on prior accuracy. Goel, Reeves, Watts, and Pennock (2010) employed forecaster selection based on self-reported confidence. This "filtered poll" method performed on par with comparison conditions.

[2] Prediction markets rely on a built-in weighting mechanism to identify skills: in the long run, forecasters who make correct bets grow richer and gain more influence over market prices (Wolfers & Zitzewitz, 2004). This is especially true in play-money markets, where traders are not allowed to add outside funds. By contrast, in prediction polls, where forecasters provide direct probability estimates, platform designers must decide how to weight these estimates (Clemen & Winkler, 1999).

[3] Information cascades may produce aggregate-level overreaction even without individual-level overreaction.

[4] Koehler (1996) notes that base-rate neglect depends on the structure and representation of the task, and argues in favor of ignoring base rates that are ambiguous, unreliable or unstable. The validity of base rate cues is an open question in naturalistic tasks such as real-world forecasting.

revisions. The question is whether such confidence—or lack thereof—is justified. Research on metacognition suggests that people have better-than-chance estimates about the accuracy of their forecasts (Harvey, 1988), despite substantial individual differences (Tetlock, 2005). Thus, smaller belief revisions may indicate higher initial accuracy. We address this possibility by separately assessing the accuracy of early versus late forecasts on each question.

Of course, it is possible that both small and large updaters may benefit from larger updates. Small updaters could be more accurate from the start, but all forecasters could improve from their starting points. To explore whether large-increment updaters are overreacting, we examine counterfactual reruns of history—specifically reruns of the time series of forecasts for each individual—that let us gauge whether accuracy would have increased or decreased if we systematically dialed the magnitude of their updates up or down. We provide a supplementary measure of underreaction versus overreaction, by assessing volatility relative to the Bayesian benchmark proposed by Augenblick and Rabin (2017). This is described in Appendix Section A5.

**Hypothesis 3.** Forecasters who confirm their forecasts more often tend to be less accurate.

Psychologists have suggested that the most consequential belief updating error may be failing to update at all (Lord, Lepper, & Preston, 1984; Nisbett & Ross, 1980). They describe a common tendency to form initial impressions of the causal propensities at play with these causal schemata then biasing the interpretation of new data. The strong version of this bias is belief perseverance, a failure to adjust one's initial estimate even when faced with evidence that it was wrong. Confirmation bias, the tendency to seek information consistent with one's initial beliefs (Nickerson, 1998), may also produce a pattern of confirming one's previous judgments.[5] A failure to update may also result from satisficing (Simon, 1956)—a forecaster may consider her probability estimates to be "good enough."[6] In the current analysis, we operationalize forecast confirmations as commissions, i.e., cases in which forecasters actively re-enter their most recent forecast on a given question.

In addition to these hypotheses, we explore three additional questions. First, are updating frequency, magnitude, and confirmation propensity unique, stable individual characteristics of forecasters? Second, what are the psychometric and behavioral predictors of update frequency, magnitude and confirmation propensity? Third, can training influence updating patterns, and if so, do update measures mediate the effect of training on accuracy? Training increases update frequency and accuracy (Chang et al., 2016; Mellers et al., 2014), but the effect of training on update magnitude is unknown. We thus examine whether probability training reduces update magnitude, and if this reduction is associated with better performance.

## 3. Method

### 3.1. Subjects and data

Data were collected by the Good Judgment Project, a research team that competed in—and won—the Aggregative Contingent Estimation (ACE) project sponsored by the Intelligence Advanced Research Projects Activity (IARPA). The tournament took place between 2011 and 2015, and consisted of four forecasting seasons, each lasting approximately 9–10 months and featuring over 100 questions. Forecasters were recruited through mailing lists, personal connections, blogs, and media coverage. Actual or imminent completion of Bachelor degree was a pre-requisite for inclusion into the study, as was the completion of an inventory of psychometric tests administered before the start of each forecasting season.

Forecasters were experimentally assigned to conditions, including training and teaming. We report data from independent prediction polls, in which forecasters provided probability estimates without access to their peers' predictions. Forecasters had the option to update their estimates whenever they wished before questions resolved. Performance was assessed using the Brier score (Brier, 1950), also known as the quadratic score, a strictly proper scoring rule. The Brier score is defined in Equation (1), where $f_c$ denotes the probability forecast placed on the correct answer of a binary question.[7]

$$Brier\ Score = 2 \times [1 - f_c]^2 \tag{1}$$

Brier scores range from 0 (best) to 2 (worst). When a question closed, daily Brier scores were calculated after a participant's first forecast. Forecasts were carried forward across days until the forecaster made an update. For days before a participant's initial forecast on a question, we imputed Brier scores from the average Brier score of forecasters on that question and condition. If a forecaster skipped a question entirely, she received an imputed score for all days of the question that was equal to the mean Brier score of those in her condition who did report forecasts on that question. These imputation rules were intended to incentivize forecasters to attempt questions for which they believed they were better than average and update whenever their forecasts could be improved. Forecasters learned about these scoring procedures at the start of each season, and they had access to scoring rule descriptions, as well as their scores and accuracy rank, throughout a season. Imputed scores on questions that forecasters did not attempt were used only for incentive purposes; the current analysis excludes such scores.

Scores were averaged across days within a question. Because forecasters selected their own questions, scores were standardized within question, i.e., converted to z-scores, to emphasize relative forecaster skill while accounting for question difficulty. Finally, standardized Brier scores were averaged across questions for each forecaster. Brier score decomposition analyses followed the formulations described in Murphy and Winkler (1987).

The fifty most accurate forecasters in an experimental condition were featured on a leaderboard. Forecasters received compensation in the form of electronic retailer gift certificates if they had made at least one forecast on 25 or more questions. The value of the gift certificates was $150 in the first two seasons, and $250 in the last two seasons. There were no financial incentives based on forecast updating or accuracy. The top 2% of forecasters at the end of each season were awarded superforecaster status and invited to participate in small teams with other superforecasters in the following tournament season.

### 3.2. Belief updating measures

The current analysis assumes that forecasts are a reasonable proxy for beliefs. While it is possible that forecasters may not always report their true beliefs, we note that proper scoring rules, such as the Brier

[5] Confirmation bias may also lead to belief polarization, where individuals only seek and find confirmatory evidence for their favored beliefs, making their probability estimates more extreme over time. Belief polarization is more likely when people hold strong and relevant ideological positions. The tournament's focus on non-U.S. geopolitical questions and the wide variety of topics may limit the influence of ideological predispositions.

[6] On the other hand, maximizers, those who strenuously seek to select optimal rather than good-enough options, are not necessarily better at forecasting (Jain, Bearden, & Filipowicz, 2013).

[7] In this specification, Brier scores may vary between 0 and 2, and 50% forecasts result in 0.5 Brier scores. A generalized version of this scoring rule was used for questions with three or more possible outcomes. For questions that had outcome categories with a pre-defined order (e.g., from low to high), we used the ordered Brier scoring rule, which assigns better scores for placing high probabilities on categories closer to the correct one (Jose, Nau, & Winkler, 2009). Brier score decomposition analyses do not apply to this modification.

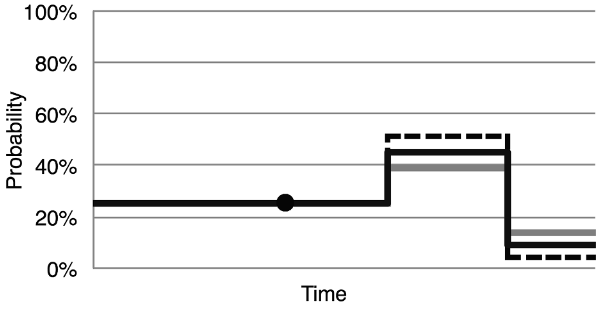


**Fig. 1.** An illustrative forecast stream for a forecaster who made three unique forecasts and one confirmation. Solid gray and dotted black lines are counterfactual forecast streams that would have resulted from proportionally smaller or large belief updates.

score, incentivize forecasters to state their beliefs truthfully and update them as necessary. In our analyses, we use more literal descriptions of changes in probabilistic estimates—forecast updates or simply updates—but posit that forecast changes correspond closely to revisions in underlying beliefs.

Imagine a forecaster faced with the question: "Will Bashar al-Assad cease to be president of Syria by May 1, 2020?" The solid black line in Fig. 1 illustrates a forecast stream. The forecaster provides an initial probability estimate of 25%. She later confirms the initial estimate, re-entering the same one, as denoted by the black dot. She then updates her forecast to 45%, and then lowers it to 9%. This forecaster has made 4 forecasts, including one forecast confirmation, and 25% of forecasts were confirmations. Update frequency in this example is 3, including the initial forecast but excluding confirmations. The forecaster's mean absolute update magnitude is (|25% − 45%| + |45% − 9%|) ÷ 2 = 28%.[8]

Average absolute update magnitude, hereafter referred to simply as magnitude, is first calculated across updates on a given question, then averaged across questions for each forecaster. Let $i$ be the index of forecasts within a question, and let $I$ be the total number of forecasts; let $q$ be the index of questions attempted by a forecaster, and let $Q$ be the total number of questions that the forecaster attempted; finally, let $p$ be the reported probability values.

$$Magnitude = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{I-1}\sum_{i=2}^{I}|p_i - p_{i-1}| \tag{2}$$

If a forecaster made only one forecast, magnitude was set to missing and thus did not affect magnitude calculations, frequency was set to one, and confirmation propensity was set to zero. Accuracy was measured for all forecasting questions, regardless of whether a forecaster updated his or her beliefs. Our sample included those forecasters who updated forecasts on at least ten questions.

### 3.3. Updating simulation

We also performed counterfactual simulations to determine whether modifying a forecaster's updating behavior would have improved accuracy. We simulated belief streams with larger or smaller updates than the actual ones. The solid black line in Fig. 1 represents the actual forecast stream, while the solid gray line and the dotted black lines illustrate two counterfactual forecast streams. The dotted black line depicts forecasts that would have resulted if update magnitudes were set to 130% of original values. The solid gray line depicts the forecast stream resulting from updates that were 70% as large as the observed values. Forecast streams were modified to fit the probability scale, i.e., simulated values below 0% and above 100% were truncated to 0% and 100%, respectively. The simulation had no stochastic components.

### 3.4. Predictors of updating and accuracy

We examined how behavioral and psychometric measures relate to updating patterns. Behavioral measures included: (a) the number of forecasting questions attempted by a forecaster; (b) the number of active forecasting sessions (i.e., the number of instances in which a forecaster logged into the web platform); and (c) the number of news-link clicks or instances in which forecasters clicked on links to news sources displayed on the platform in seasons 2 and 3. These links were based on the top results of Google News search queries featuring the keywords of forecasting questions.

Mellers, Stone, Atanasov, et al. (2015) showed that forecasters with high fluid and crystallized intelligence, and actively open-minded thinking styles tended to be more accurate. We include these measures as predictors of both updating measures and accuracy. Fluid intelligence was a combination of scores from Raven's progressive matrices test (Balboni, Naglieri, & Cubelli, 2010), Shipley's analytical reasoning test (Zachary, Paulson, & Gorsuch, 1985), a numeracy test (Lipkus, Samsa, & Rimer, 2001; Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007), and the cognitive reflection test (Frederick, 2005).

### 3.5. Forecasting training

Does forecasting training influence updating patterns? Approximately half of participants were randomly assigned to a training condition at the start of each season of the tournament. These forecasters were required to complete training in order to participate in the tournament. Forecasters who received training were later assigned to training conditions in subsequent seasons. Training content was designed to improve overall forecasting accuracy, not to produce specific updating patterns. Three topics were especially relevant: (a) constructing comparison classes and calculating base rates, (b) combining potentially conflicting information from multiple sources, and (c) updating forecasts in response to new information. Training took approximately one hour to complete, and resulted in approximately 8–10% improvement in accuracy over the course of each of four seasons. For more information on the content and effects of training, see Mellers et al. (2014) and Chang et al. (2016).

### 3.6. Cross-Sample validation

To provide robust tests of our hypotheses, we employed a cross-validation procedure, in which belief updating patterns were measured on one set of forecasting questions and then used to predict forecasting accuracy in another set of questions. This procedure consisted of three steps. First, for each forecaster in the sample, we independently and randomly assigned the questions attempted to two approximately equally-sized sets, that we refer to as A and B. Second, in each set of questions, we calculated the rate of belief confirmation (proportion of all forecasts that were confirmations), update frequency (number of forecasts per question), update magnitude (mean absolute difference between successive unique forecasts), and standardized Brier scores. Finally, we used these updating measures from set A to predict accuracy in set B (and, analogously, updating measures from set B to predict accuracy in set A), in several regression models. Individual forecasters were the units of analysis.

We repeated this procedure with 50 iterations of random half-sample splits, creating new sets, A and B, in each iteration. Each split yielded two sets of regression coefficients per model: one for predicting accuracy in set A based on behavioral measures in set B, and another for predicting accuracy in set B based on behavioral measures in set A.

[8] We used absolute magnitude, rather than squared magnitude. In this example, the mean squared belief update magnitude was 0.085. We applied squared update magnitude in a sensitivity analysis, listed in Appendix Table A2.3. Results were similar for absolute and squared magnitude measures.

Thus, we had 100 sets of regression coefficients. We report those regression estimates that were closest to the median set of *t*-test values. We used a similar procedure for estimating correlations, and present the median correlation coefficients across iterations.

### 3.7. Sample characteristics

Forecasting accuracy was assessed based on 481 forecasting questions. The median duration of questions was about three months (Median = 81 days, M = 110 days, SD = 93). These questions were posed and resolved over the course of four forecasting seasons, each lasting approximately 9–10 months. The core study sample consists of N = 515 participants who made at least one forecast update on at least 10 forecasting questions. This inclusion criterion is consistent with those used in prior research of individual forecasting accuracy (e.g., Mellers et al., 2014). Results held when we included the sample of all available forecasters, as long as they attempted two questions, the minimum needed for a split-sample analysis (see Appendix Section A2). Forecasters could choose to return from one forecasting season to another. As long as they fulfilled the inclusion criterion, we did not distinguish between forecasters who were active in one versus multiple forecasting seasons.

Forecasters in the study sample attempted an average of one hundred and thirteen questions (M = 113, SD = 73), and made at least one update on forty-three of those questions (M = 43, SD = 35). Frequency of updating was the number of forecasts per question, including the initial one. The average forecaster submitted two forecasts per question (M = 2.0, SD = 1.6), i.e., one initial forecast and one update. The pattern of update frequency was best approximated by a log-normal distribution. Frequency was log-transformed before being used in the regression models. The average absolute update magnitude was 0.20 (SD = 0.10) on the probability scale (also see Table 1). For the average forecaster, 19% of all forecasts were confirmations. Confirmations are treated as distinct from initial forecasts and updates.

Superforecasters, as discussed in Mellers, Stone, Murray, et al. (2015) and Tetlock and Gardner (2015), are only included in our core analysis in the seasons before attaining superforecaster status, which was granted to the top 2% of performers at the end of each season based on Brier score. We perform a separate sensitivity analysis using data from superforecasters after they attained superforecaster status and were assigned to work in teams (see Appendix Table A2.2). A total of N = 175 superforecasters are included, and the sensitivity analysis includes all forecasts submitted after they attained superforecasting status. Superforecasters submitted a mean of 5.1 forecasts per question (vs. 2.0 in the core sample), and had an average absolute update magnitude of 0.11 (vs. 0.20 in the core sample).

## 4. Results

The analyses below are organized as follows: (a) reliability of updating measures; b) associations between update measures and forecasting skill; (c) predictability of update patterns from other behavioral and psychometric measures; and (d) effects of training.

**Table 1**
Forecasting behavior and updating characteristics in N = 515 forecasters.

| Study sample characteristics | Mean (SD) | Median (IQR†) |
|---|---|---|
| Number of questions with updates* | 43 (35) | 32 (17, 59) |
| Average update frequency per question* | 2.0 (1.6) | 1.6 (1.3, 2.1) |
| Average absolute update magnitude* | 0.20 (0.10) | 0.19 (0.14, 0.24) |
| Forecast confirmations, % of forecasts | 0.19 (0.18) | 0.18 (0.06, 0.26) |

*Notes:* * Forecast confirmations are distinct from updates and are not included in calculations. † Inter-quartile range.

### 4.1. Reliability and uniqueness

We use Pearson correlation coefficients across question sets (out of sample) as measures of test–retest reliability. We also use out-of-sample correlation coefficients to measure the predictiveness of different updating patterns for Brier score. Standardized Brier scores, where lower values denote better accuracy, had a test–retest reliability of $r = 0.75$ across questionsets. All updating measures had higher reliability coefficients than Brier score. Update frequency had the strongest reliability of $r = 0.98$, suggesting that the tendency to update more or less often was a highly stable individual difference. Update magnitude had a test–retest reliability of $r = 0.79$, denoting that forecasters who made small or large updates in one set of questions tended to do the same in the other set. Appendix Fig. A1 depicts the distribution of update magnitude in our sample. The reliability of update confirmations was $r = 0.83$.

To study the relationship between updating measures, we use correlation coefficients within a question set sample (in sample), allowing us to estimate whether they capture different aspects of updating behavior. Magnitude and frequency were negatively correlated; forecasters who updated more often tended to take smaller steps. The relationship was weak to moderate ($r = -0.26$, $p < .001$). Update frequency correlated positively with confirmation propensity ($r = 0.51, p < .001$); forecasters who made more frquent updates were also tended to confirm their prior forecasts. Update magnitude did not correlate with the propensity for confirming one's beliefs ($r = 0.02$, $p > .10$), as forecasters who confirmed their beliefs more often showed no tendency to make smaller updates. This provides evidence for a psychological distinction between the decisions of *whether* and *how much* to update one's beliefs. Test-retest (i.e., out-of-sample for the same variable) and cross-variable (in and out-of-sample) correlation coefficients are shown in Table 2.

### 4.2. Belief updating and accuracy

Tests of our hypotheses consisted of ordinary least squares (OLS) regression models, all of which had the same dependent variable: relative accuracy of forecasters for out-of-sample questions, as measured by mean standardized Brier scores. Update frequency, magnitude and confirmation propensity were predictors. The specifications also accounted for covariates, such as psychometric scores and out-of-sample accuracy scores. All models included training assignment as a covariate.

The first model, shown in Table 3, column A, tested the relationship of frequency of forecasts per question, magnitude of updates, and confirmation propensity with forecasting accuracy. This model investigated all three hypotheses. To facilitate comparison of effect sizes and regression coefficients, frequency, magnitude and confirmation measures were standardized.

Results were consistent with Hypothesis 1, which stated that update frequency would be positively related to accuracy, i.e., negatively related to Brier scores (b = −0.079, $t = -6.86$, $p < .001$). Consistent

**Table 2**
In and out of sample Pearson correlation coefficient matrix across N = 515 forecasters. Median coefficients based on 100 resamples shown.

| Variable | In vs. out of Sample | Brier Score | Frequency | Magnitude | Confirmation |
|---|---|---|---|---|---|
| Brier Score | In | 1 | | | |
| Frequency | In | −0.32 | 1 | | |
| Magnitude | In | 0.49 | −0.26 | 1 | |
| Confirmation | In | 0.03 | 0.51 | 0.02 | 1 |
| Brier Score | Out | 0.75 | | | |
| Frequency | Out | −0.32 | 0.98 | | |
| Magnitude | Out | 0.45 | −0.25 | 0.79 | |
| Confirmation | Out | 0.04 | 0.51 | 0.02 | 0.83 |

**Table 3**
Predictors of forecaster accuracy. Results are based on OLS models in which measures based on one set of questions are used to predict accuracy in a different set of questions. All continuous variables, independent and dependent, are standardized. Lower scores denote better accuracy. Bounds of 95% confidence intervals are shown in brackets.

| DV: Standardized Brier Score | A. Base Model | B. A + Psychometrics | C. A + Accuracy in Training Set |
|---|---|---|---|
| Intercept | −0.050<br>[−0.079, −0.021] | −0.069<br>[−0.102, −0.036] | −0.069<br>[−0.091, −0.047] |
| Frequency | −0.079<br>[−0.101, −0.057] | −0.060<br>[−0.084, −0.036] | −0.037<br>[−0.055, −0.019] |
| Magnitude | 0.098<br>[0.074, 0.122] | 0.105<br>[0.081, 0.029] | 0.024<br>[0.006, 0.042] |
| Confirmation propensity | 0.028<br>[0.006, 0.050] | 0.010<br>[−0.017, 0.037] | 0.020<br>[0.004, 0.036] |
| Training | −0.063<br>[−0.104, −0.022] | −0.099<br>[−0.144, −0.054] | −0.015<br>[−0.046, 0.016] |
| Fluid intelligence | | −0.064<br>[−0.086, −0.042] | |
| Political knowledge | | −0.001<br>[−0.025, 0.023] | |
| AOMT | | −0.012<br>[−0.034, 0.010] | |
| Accuracy in training set | | | 0.173<br>[0.155, 0.191] |
| N | 515 | 382 | 515 |
| Adj. $R^2$ | 0.28 | 0.38 | 0.57 |

with Hypothesis 2, larger update magnitude was associated with higher Brier scores, i.e., worse forecasting accuracy (b = 0.098, $t$ = 8.69, $p$ < .001). And as predicted by Hypothesis 3, confirmations were associated with worse accuracy (b = 0.028, $t$ = 2.62, $p$ = .009). In short, more accurate forecasters tended to make frequent, smaller updates and rarely confirmed their previous forecasts. Training was associated with greater accuracy after accounting for updating patterns. This specification provided the basis of the sensitivity analyses listed in the Appendix Section A2.

To provide context for the coefficients in Table 3, Column A, we map standardized to raw values, for both Brier scores and update measures. An untrained forecaster with mean scores on all updating measures would have received a raw Brier score of 0.36. Ceteris paribus, if that forecaster had update frequency that was one standard deviation (1 SD) below average (1.1 forecasts per question versus mean of 1.7), her model-predicted raw Brier scores would be 0.39, while a forecaster with 1 SD above average frequency (2.8 forests per question) would have an estimated Brier score of 0.33. If update magnitude were one SD below or above the mean (0.10 or 0.31 absolute magnitude versus a mean of 0.20), predicted Brier scores would be 0.33 and 0.39, respectively. If confirmation propensity were one SD below or above the mean (1% or 37% confirmation rate versus 19% mean), predicted Brier scores would be 0.35 or 0.37, respectively. Undergoing training would reduce estimated Brier scores from 0.36 to 0.34.

Then we examined the predictive value of updating measures in the presence of psychometric measures. See Table 3, column B. Fluid intelligence was associated with accuracy (b = −0.064, $t$ = −5.29, $p$ < .001), while political knowledge and actively open-minded thinking were not (p > .10 for both). Frequency and magnitude were related to accuracy when accounting for those covariates. Of all the psychometric and behavioral predictors of accuracy in this model, update magnitude was the strongest (b = 0.105, $t$ = 8.47, $p$ < .001).

Finally, we assessed whether updating measures were associated with accuracy even if one's track record was included as a predictor. (See Table 3, Column C.) We used standardized Brier scores in the training set as predictors of standardized Brier scores in the validation set of questions. The relationship between accuracy in training and validation sets was strong ($b$ = 0.173, $t$ = 18.24, p < .001). Frequency (b = −0.037, $t$ = 4.14, $p$ < .001), magnitude ($b$ = −0.024, $t$ = −2.60, $p$ = .010), and confirmation propensity ($b$ = 0.020, $t$ = 2.43, $p$ = .015) were also associated with accuracy. Across the 100 split-sample iterations, 100% yielded negative coefficients for frequency, while 97% yielded positive coefficients for magnitude, and 96% yielded positive coefficients for confirmation propensity. These 100 iterations are not independent, so frequency counts reported above should be interpreted with caution.

Additional tests are listed in Appendix Section A2. Table A2.1 shows the results when we relax or tighten forecaster inclusion criteria regarding the number of questions with updates. Table A2.2 shows that the core results directionally replicated the base model in a sample of superforecasters working in teams. Table A2.3 provides a sensitivity analysis using an alternative measure of accuracy: mean-debiased rather than standardized Brier scores; and an alternate measure of magnitude: squared distance rather than absolute distance. Table A2.4 breaks down performance by question resolution outcome: status quo vs. change and time-sensitive vs. others. All of these analyses yield results that are consistent with the base model: magnitude was significantly associated with standardized Brier scores in all cases, frequency was significantly associated with accuracy in all cases except the superforecaster analysis, and confirmation propensity was associated with accuracy, except for the least selective independent forecaster sample, the superforecaster sample, and when questions were broken down by type and outcome.

### 4.3. Early versus late forecasts

The high accuracy of frequent, incremental updaters could be associated with highly accurate initial forecasts or with accuracy improvements attributable to the updates. To distinguish between these possibilities, we examined the accuracy of first vs. last forecasts (see Appendix Table A3.1). Small updates were a marker of initial accuracy. A one standard deviation decrease in magnitude corresponded to a 0.12 decrease in mean standardized Brier scores (b = 0.121, t = 10.94, p < .001). Magnitude was a weaker, but still significant predictor of last-forecast accuracy (b = 0.044, t = 3.64, p < .001). Forecasters appeared to demonstrate sufficient metacognitive skill to gauge how much they needed to update their beliefs. Those with greater initial accuracy needed—and made—smaller updates.

Greater frequency of updating was the best predictor of last-forecast accuracy. An increase of one standard deviation in frequency corresponded to a 0.18 decrease in mean standardized Brier scores ($b$ = −0.181, $t$ = 14.72, $p$ < .001). However, frequency was unrelated to initial accuracy ($b$ = 0.008, $t$ = 0.67, $p$ > .20). The propensity to confirm prior predictions was associated with worse accuracy of both initial forecasts ($b$ = 0.030, $t$ = 2.80, $p$ = .005) and final forecasts ($b$ = 0.045, $t$ = 3.90, $p$ = .001). In summary, forecasters who made relatively accurate initial judgments tended to make smaller belief updates, while frequent updaters got more accurate over time.

Fig. 2 illustrates the relationship between frequency, magnitude, and accuracy. Forecasters were separated into four categories using a median-split on frequency and magnitude of updating. The median frequency was 1.6 forecasts per question, and the median magnitude was 19 percentage points. Magnitude and frequency were correlated, so approximately twice as many forecasters were placed in the "large, infrequent" (N = 164) and "small, frequent" (N = 164) categories as in the "large, frequent" (N = 93) and "small, infrequent" (N = 94) categories. We divided forecasts based on whether they were made in the beginning, middle, or end of the forecasting period. For example, if a question was open for 90 days, we would separately score forecasts made from days 1 to 30, 31 to 60, and 61 to 90. Mean scores on a hold-out set of questions are presented, with no regression adjustments.

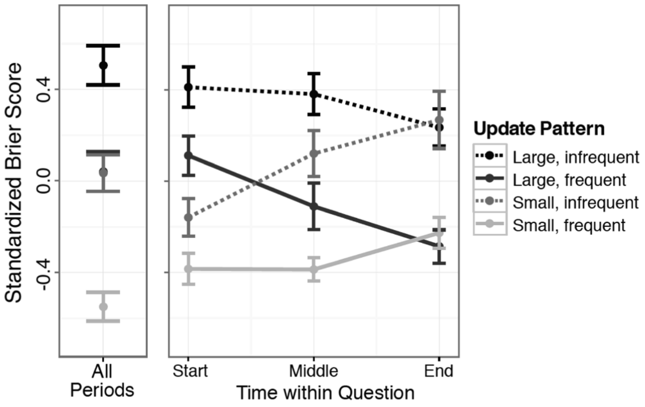


**Fig. 2.** Standardized Brier scores by update pattern and period within question. Forecasters were placed in four categories by median splits on frequency and magnitude in a training sample of questions. Mean standardized Brier scores based on a different set of questions are shown. Scores are divided by time period within question. Calipers denote one standard error of the mean in each direction.

Across all periods, forecasters who made small, frequent adjustments achieved Brier scores that were approximately 0.5 standard deviations better than the average forecaster. Forecasters who made large, infrequent updates were 0.5 standard deviations worse than average. Differences in full-period accuracy were larger than those in any subperiod. This is partly due to the lower variance in raw Brier scores for the full period, which accentuates differences in standardized (variance-adjusted) Brier scores.

The regression results for first versus last forecast analysis (see Appendix Table A3.1) were directionally similar to the analysis of early, middle, and late period forecasts (see Appendix Table A3.2), except that the association between magnitude and accuracy was approximately null in the late period. Magnitude and frequency were also associated with accuracy in an analysis of forecast-level Brier scores, adjusting for forecast order and timing (see Appendix Table A3.3 and Fig. A3). In sum, magnitude was more strongly associated with accuracy early on, while frequency was more strongly linked to accuracy in the later periods of questions.

### 4.4. Calibration and discrimination

What advantages do small, frequent updaters have over other forecasters? We performed Brier score decomposition analysis to determine how forecasters with different update patterns vary with regard to calibration and discrimination, following the same categorization and cross-validation strategy used to produce Fig. 2. Four questions were excluded from the analysis, due to lack of Brier score data among at least one of the four groups. Forecasters from all four categories covered the remaining questions, so the uncertainty score is the same across categories. Results are shown in Table 4. Large, infrequent updaters registered the highest (worst) raw Brier scores, with the highest levels of calibration error and worst (lowest) discrimination scores. Large, frequent updaters were approximately tied with small, infrequent updaters in terms of raw Brier scores, with the former outperforming in terms of discrimination but slightly underperforming in terms of calibration. Small, frequent updaters performed the best in the validation set of questions, registering the lowest calibration errors and the best discrimination scores. For calibration plots, see Appendix Section A4, including Fig. A4.

**Table 4**
Brier score decomposition for individual forecasters by updating category.

| Group | Raw Brier Score | Calibration Error | Discrimination | Uncertainty |
|---|---|---|---|---|
| Large, infrequent | 0.432 | 0.023 | 0.174 | 0.583 |
| Large, frequent | 0.374 | 0.014 | 0.224 | 0.583 |
| Small, infrequent | 0.394 | 0.010 | 0.200 | 0.583 |
| Small, frequent | 0.292 | 0.002 | 0.293 | 0.583 |

### 4.5. Update magnitude simulation

We have shown that forecasters who updated incrementally tended to be more accurate. But did forecasters, on average, under- or overreact to new information? Would forecasters have benefitted from debiasing their estimates after elicitation? To answer these questions, we constructed simulated forecast streams with smaller or larger-than-actual update magnitudes. This procedure was illustrated in Fig. 1.

We produced forecast streams with 30% smaller-than-actual and 30% larger-than-actual belief updates, filling in the intermediate steps in 10 percentage-point increments. For both the factual and counterfactual forecast streams, we then scored the accuracy of the last forecast made by a forecaster on a given question, i.e., the forecast they would have produced after all forecast revisions. Accuracy was assessed in terms of absolute Brier scores and indexed to the accuracy of actual forecasts, which was set to 100%. Indexed scores above 100% indicated worse accuracy of counterfactual forecasts relative to actual ones, and vice versa. For example, an indexed Brier score of 103% denotes that the simulated forecasts yielded 3% higher (worse) Brier scores than the actual forecasts. Scores shown in Fig. 3 represent simple means of Brier scores across forecasts, and do not account for clustering across forecasters or questions.

Fig. 3 shows that actual forecasts (denoted by 100% on the horizontal axis) resulted in nearly optimal accuracy. Increasing update size by 20–30% produced an accuracy boost, reducing Brier score by 0.3% on average across all forecasters, a relatively small improvement. For comparison, the mean Brier score for forecasts made by incremental updaters (M = 0.22) was approximately 30% lower than that for large-increment updaters (M = 0.29). Thus, selecting forecasters based on small update magnitude would have produced an approximately 100 times stronger accuracy-boosting effect than increasing update size ex-post (30% versus 0.3% Brier score decrease).

On the other hand, reducing update increments across the board worsened accuracy, increasing Brier scores by up to 4%.Incremental updaters would have benefited by 0.7% from larger updates. We calculated the optimal update magnitude transformation and found that 65% of participants would have received better scores on their last forecasts if they had made larger updates, 25% would have benefited from smaller updates, and 10% would not have benefited from any of the adjustment levels we tested. The median forecaster would have

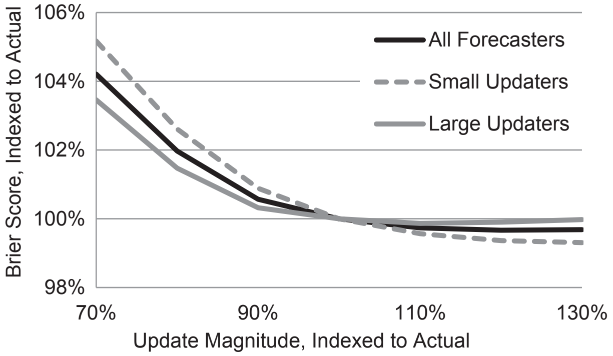


**Fig. 3.** Last-forecast accuracy of simulated forecast streams with smaller-than-actual and larger-than-actual belief updates. Black line shows results for all forecasters, dotted-gray and solid-gray lines show results for forecasters with smaller and larger than median update magnitude, respectively. Lower indexed Brier scores denote better accuracy.

achieved greater accuracy from a 10% increase in update magnitude. Increment updaters were more likely to benefit from magnitude increases, but the relationship between magnitude and optimal transformation was not strong enough to justify applying custom transformations. Thus, it appeared that forecasters showed a slight bias toward underreaction to new information, but correcting such bias would have resulted in very small improvements.

Augenblick and Rabin (2017) provide an alternative measure of underreaction versus overreaction, based on a comparison of the initial extremity and subsequent time-series movements of forecasts. This comparison showed that forecast streams suffered from excess volatility on average; only incremental updaters produced forecast streams with appropriate volatility levels (see Appendix Section A5).

### *4.6. Belief updating, effort, and psychometric profiles*

What distinguishes frequent updaters from infrequent updaters, and incremental updaters from large-increment updaters? Linking the psychometric profiles of forecasters to their updating propensities may give us insights. We conducted regressions with update measures as the dependent variable and activity and psychometric measures as predictor variables.[9] Activity measures indicate the extent to which belief updating patterns were associated with effort.

Update frequency was strongly associated with activity measures. Frequent updaters tended to attempt more questions, log in more often, and click on more news links. Frequent updaters appeared to be active information gatherers based on available activity measures. Update magnitude had negative or null associations with activity measures. Incremental updaters tended to spread their activity across a small number of questions, and log in to the forecasting platform less often. There was no association between magnitude and news-click activity. Higher confirmation propensity was strongly associated with more sessions (i.e., number of logins), but not with other activity measures. Overall, apart from its association with update frequency, smaller magnitudes were not correlated with higher levels of activity. For full regression results, see Appendix Section A6, including Table A6.

In the analysis of psychometric measures, higher updating frequency was associated with higher scores on political knowledge and actively open-minded thinking (AOMT), but it was unrelated to fluid intelligence. In contrast, lower updating magnitudes were associated with higher fluid intelligence, but magnitude was unrelated to political knowledge and AOMT scores. The only significant predictor of belief confirmation was fluid intelligence. Moreover, forecasters with high fluid intelligence scores were less prone to confirm their forecasts.

### *4.7. Training effects on updating and accuracy*

Training was associated with more frequent updates and greater accuracy (Mellers et al., 2014). We replicated this result, as trained forecasters updated 15% more often ($M = 1.9$, $SD = 1.3$ vs. $M = 2.2$, $SD = 1.8$, $t = 1.79$, $p = 0.075$). Here we ask whether update magnitude is associated with training. Across four years of the tournament, trained forecasters made updates that were on average 12% smaller ($M = 0.22$, $SD = 0.11$ for untrained vs. $M = 0.19$, $SD = 0.09$ for trained forecasters, $t = 2.52$, $p = 0.012$).

We conducted mediation analyses to see whether updating patterns accounted for the effects of training on accuracy. The analysis employed a causal mediation approach, as implemented in the mediation package in R statistical software (Tingley, Yamamoto, Hirose, Imai, & Keele, 2014). In Fig. 4, the indirect effect of frequency accounted for 17% of the total effect of training on accuracy (proportion mediated = 0.17, 95% CI (−0.02, 0.35)). The benefits of training were partially channeled through more frequent updates. The indirect effect of magnitude was stronger, accounting for 29% of the total effect of training on accuracy (proportion mediated = 0.29, 95% CI (0.09, 0.55)). Training caused forecasters to update their beliefs in smaller increments, which in turn boosted accuracy. Confirmation propensity did not mediate the training effect (proportion mediated $< 0.01$). See Appendix Section A7, including Tables A7.1–A7.3.

## 5. Discussion and conclusion

### *5.1. Two paradoxes*

In our core tests of forecasters working independently, we showed that small, frequent updates were strongly and robustly associated with greater accuracy. These results directionally replicated among elite forecasters working in teams. Two perhaps counterintuitive patterns accompanied these top-line results.

First, the tendency to confirm one's prior forecasts was associated with worse performance. This might seem inconsistent with simple brain-as-computer intuitions about forecasting. For example, if a machine model provides the same probability values on two subsequent weeks, one possibility would be that positive and negative inputs cancelled each other out, yielding a forecast update that rounds to zero. A more conservative model would produce smaller updates and more frequent confirmations. A more aggressive model could produce larger updates and fewer confirmations. In contrast, we found that decisions on *whether* and *how much* to adjust forecasts were unrelated to one another. Belief confirmation propensity was qualitatively different from the tendency to make small updates. In fact, both fewer belief confirmations and smaller updates correlated with higher fluid intelligence and higher accuracy. Thus updating may be best modeled using a mixture distribution, separately estimating the probability of a non-zero update and update magnitude.

The other counterintuitive aspect relates to the interpretation of these results: Forecasters who made small belief revisions were highly accurate, and training effectively reduced update magnitude, thereby boosting accuracy. However, these results should not be interpreted to mean that simply advising forecasters to make small updates would improve their accuracy. Forecasting training did not issue explicit, general-purpose advice favoring small updates. And our simulations showed that simply reducing update magnitude would have degraded accuracy. It appears that the most accurate forecasters did not update in small increments by mechanically throttling down update increments; instead, their revisions reflected generally accurate meta-judgments that large updates were unnecessary. In other words, forecasters generally demonstrated well-calibrated trust in their previous forecasts, a tendency that persisted across forecasting questions. A consistent pattern of small updates was the key factor that differentiated those who were initially accurate from those who were not.

We considered three ways in which training could have reduced magnitudes and improved accuracy. First, we instructed forecasters to ground their estimates in stable historical base rates. This advice could have diminished the relative weight forecasters placed on new, inside-view cues.[10] This explanation is consistent with the observation that small updates were associated with more accurate initial forecasts.

[9] Activity measures were not used as predictors in models focused on accuracy, such as those shown in Table 3, for two reasons. First, news link click data were only available for Seasons 2 and 3 of the tournament. Second, activity measures, such as news link clicks and logins were not question specific, so they could not be meaningfully incorporated in analyses that utilize cross-validation across questions.

[10] As an illustration of the way incorporating base rates dampens update size, compare FiveThirtyEight's 2016 Presidential Election forecasting models, polls-plus and polls-only variants (Silver, 2016). The polls-plus incorporated stable cues such as economic indicators and produced smaller updates than the polls-only model.

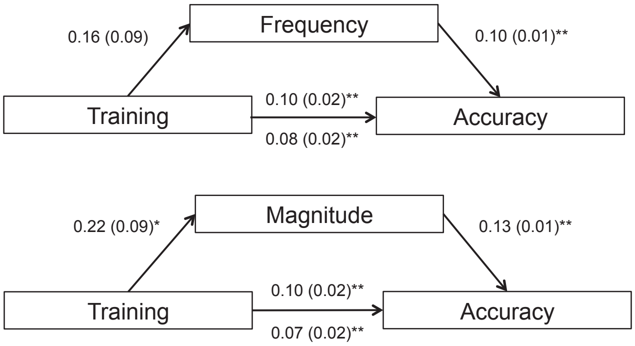


**Fig. 4.** Mediation models showed that frequency and magnitude account for 17% and 29% of the effect of training on accuracy, respectively. Lower scores denote better accuracy. Standardized regression coefficients are shown, with standard errors in parentheses.

Another way in which training could have influenced updating and accuracy was that training encouraged forecasters to seek more information. They may have picked up subtler signals that less careful forecasters missed. To illustrate, let's compare two hypothetical forecasters, one who reads a newspaper from start to finish, and another who reads only the front page. Most news stories do not fit on the front page, so the first forecaster is likely to pick up more signals about future developments. But some of the mid-page news may be of limited predictive value. Front-page news on any topic is, in contrast, less frequent but usually more informative of current or future developments. The in-depth reader is thus likely to update in frequent, but smaller steps. This depth-of-information-search link was not supported by our analysis: frequency and magnitude were only moderately correlated, explaining less than 8% of the variance in one another. In addition, incremental updaters were not espcially active information consumers, and they-derived their accuracy advantages primarily from their superior initial estimates.

Finally, training materials instructed forecasters to combine information from multiple sources, i.e., to average probability value estimates based on different sources. This guidance might have helped forecasters synthesize evidence rather than attending to a single cue and ignoring all others. More generally—and more speculatively—forecasting training may have improved forecasters' ability to hold opposing ideas in their heads while still retaining the ability to function—a "test of first-rate intelligence," in F. Scott Fitzgerald's words (2009). Indeed, the effect of forecasting training on accuracy was equivalent to that of a fluid intelligence boost of approximately 1.5 standard deviations (in Table 3, Column B, compare coefficients for training and fluid intelligence), and nearly one third of the training effect was mediated by smaller update magnitude.

### 5.2. Updating and effort

The current analysis of forecast updating and accuracy addresses an often-raised concern: that superior forecasting performance is mostly a matter of hard work and has nothing to do with unique skill; that accuracy scores reveal which forecasters are working hard, not necessarily which ones are thinking in the right way. And because forecasting tournament research is often conducted on volunteers who have leeway to choose how hard to work, results might not generalize to settings in which effort is strongly incentivized.

The connection between belief updating and effort is non-trivial because effort and engagement are often difficult to assess reliably. Even professional forecasters are frequently inattentive (Andrade & Le Bihan, 2013), so understanding who pays attention to the task is useful. But not all possible measures related to attention and effort are associated with accuracy: the number of questions participants attempted did not correlate with average accuracy. Frequent belief updates indicate a specific way in which forecasters choose to engage and invest effort, which was highly reliable and a valid predictor of accuracy.

The other two forms of updating, confirmation propensity and magnitude, told a different story. Better forecasters made fewer confirmations. Although greater confirmation propensity indicated more effort, the extra work of confirming one's beliefs was associated with less accuracy. Incremental updating was positively correlated with some activity measures and negatively with others. Incremental updaters outperformed without putting in more apparent effort than large-increment updaters.

Our results provide evidence for two distinct mechanisms of association between updating and accuracy. Update frequency was associated with *the quantity of information* forecasters processed in three ways. Frequent updaters had higher levels of activity, including observed information consumption. They had higher starting levels of political knowledge, indicating they had processed or retained more information about relevant political facts before their tournament season entry. Finally, they scored relatively higher in open-mindedness, implying higher willingness to seek and process new information. By contrast, update magnitude may have captured *the quality of information processing.* Small updaters did not exhibit higher levels of activity and information consumption than large updaters. Instead, incremental updaters were distinguished by their higher fluid intelligence, exposure to training, and higher accuracy of initial judgments.

### 5.3. Situational context

The relationship between frequent updating and higher accuracy seemingly contradicts results from the multiple cue probability learning literature (e.g., Castellan, 1973), in which frequent updating may be driven by responses to irrelevant cues. While updating in response to pseudo-diagnostic cues is a threat in our context as well, the average independent forecaster made only two forecasts per question—one initial estimate, and one update. At such a low rate of updating, the benefits of bringing valid, new information presumably outweighed the risks of attending to irrelevant cues. In our sensitivity analysis of high-update-frequency superforecasters, the association between frequency and accuracy was notably weaker (See Appendix Section A2). Thus, as average update frequency increases, the marginal benefit of additional revisions may be reduced or even reversed.

Our results showed that underreaction was a larger threat to accuracy than overreaction, consistent with Edwards (1968) notion of Bayesian conservatism. The forecasting tournament task allowed forecasters to learn through practice, making it conceptually similar to the learning-from-experience paradigm employed by Edwards.

Fildes, Goodwin, Lawrence, and Nikolopoulos (2009) found that experts making small adjustments to model forecasts made statistical predictions worse, while larger adjustments improved accuracy. We point out two potential explanations for the divergence between these results and ours. First, our forecasters had the choice to update their own past estimates, not those produced by a model. The forecasters' insight into their own past reasoning may help them develop better-calibrated trust in past estimates than in a model's forecast, avoiding biases against trusting external advice (e.g., Yaniv & Kleinberger, 2000; Dietvorst, Simmons, & Massey, 2015). Second, while the forecasters in our study competed purely on accuracy, experts in Fildes et al. (2009) operated in a professional setting where the desire to signal competence and attention to detail may have watered down the motivation to maximize accuracy. As Fildes et al. noted, small adjustments might have served primarily to send such signals.[11]

[11] In a different context, politicians appear to believe that the public perceived incremental updating more favorably than large changes, and usually describe their views as having *evolved* rather than *changed* (Leibovitch, 2015).

### 5.4. Beyond forecasting tournaments

The current effort likely represents the largest empirical study of naturalistic belief updating. As such, the study setting was unusually rich in information: it included hundreds of forecasters who underwent psychometric assessments and produced thousands of verifiable probabilistic forecasts across hundreds of questions. Few real-world environments offer such rich data. We surmise that belief updating measures will be especially useful in low-information settings.

While this study focused on relative accuracy of individual human forecasters, belief updating measures may be informative in assessing the validity of predictive sources more generally, be they groups of forecasters, advisors, statistical models, or even wider fields of inquiry. As an example, consider nutritional guidelines. Some forms of traditional medicine have produced specific, albeit implicit, predictions about the influence of different foods on the human body. Such predictions have not been updated for hundreds of years. Such lack of belief revision raises doubts about the resulting recommendations: mankind has surely produced new knowledge about health effects of food that has not been incorporated. Conversely, modern nutritional science has gone through large swings in implicit beliefs and explicit guidelines about the relative risks of consuming saturated fat versus sugar (DiNicolantonio, Lucan, & O'Keefe, 2016). Such belief swings may indicate the relative paucity of reliable and valid evidence on this matter. While philosophers of science have posited that science advances primarily through revolutions (Kuhn, 1954), frequent, incremental opinion revisions in scientific communities may indicate a healthy combination of maturity and openness to new knowledge.

### 5.5. Limitations and future directions

A threat to the internal validity of studies in real-world settings is that the stimulus-generation mechanism is not under experimental control and the correct answer—that is, the correct probabilistic forecast on a given outcome at any given time—is inherently unknowable. We will never know whether Nate Silver was right that the likelihood of Hillary Clinton winning the 2016 Presidential Election, the day before the election, was approximately 70%. By contrast, laboratory tasks, such as those involving card decks or urns and balls, produce perfectly knowable answers and thus allow us to conduct more precise comparisons between normative and actual behavior. In our setting, the relationships between belief-updating propensities and forecasting outcomes could only be assessed in the aggregate by following hundreds of forecasters across many forecasting problems. Despite this limitation, we believe that the gains in external validity from real-world judgment tasks such as forecasting tournaments more than compensate for imperfect experimenter control.

Our ability to generalize the current findings is somewhat limited by the context of the task. Forecasting questions in the tournament relate to relatively complex global issues, and many of the informational cues available to forecasters were quite subtle in nature, such as new data on shifts in public mood or new readings of economic indicators. Thus, the current results may not generalize to forecasting environments in which large shifts are far more common (Massey & Wu, 2005). Examples of such settings include professional sports, where a single event can change the course of a game or even a season, or clinical drug trials, where data become publicly available in few but highly informative steps.

Within the political domain, it is possible that small updates were associated with forecasting accuracy in part because the underlying events developed gradually, and that political environments characterized by regime shifts would have produced different results. We cannot rule this out but consider it unlikely because the forecasting tournament covered a wide range of events, including ones in settings beset with high volatility and sudden developments, such as the Arab Spring and the Greek financial crisis. Our analyses of question subsets with status quo versus change outcomes, shown in Appendix Section A2, show that our results would hold even if a larger proportion of questions yielded non-status quo outcomes. Still, it would be useful to test if our results would replicate in more dynamic settings.

Separately, the scope diversity of geopolitical forecasting questions likely limits the influence of politically motivated reasoning in belief updating. For example, questions about leadership transition in Zimbabwe may fire up less political passion among our mostly U.S.-based forecasters than questions at the center of the U.S. political discourse. Thus, the current results should be placed in the context of a tournament, in which forecasters were generally motivated to improve their forecasting skills.

### 5.6. Conclusion

Our results offer evidence supporting Bezos' proposition that people who get things right tend to change their mind often. However, accurate forecasters are not flip-floppers: their tendency to change their mind frequently is matched by a propensity for gradual revision. The best forecasters seemingly experience the prediction task as a long sequence of slight surprises rather than a short string of hard collisions with reality. A pattern of frequent, small belief adjustments helps identify forecasters and decision makers who maintain an edge in a complex, turbulent world.

## 6. 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.

## Acknowledgements

The authors benefited from helpful comments from conference participants at the Judgment and Decision Making Society meeting and the Forecasting Workshop at the 18th ACM conference on Economics and Computation as well as from helpful discussions with Michael Bishop and Stephen Bennett. The authors also thank the Intelligence Advanced Research Projects Agency (IARPA) for their support. This research was supported by 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.

## Appendix A

### A.1. Distribution of update magnitude

The figure below describes depicts the frequency of forecast revisions of a given absolute magnitude. The most common revision is zero, which corresponds to a confirmation, followed by 10-percentage points and 5-percentage points. Each observation contains one or more updates for one forecaster on one question. If a forecaster makes more than one update, absolute magnitudes are averages across updates within a forecaster-question (see Fig. A1).

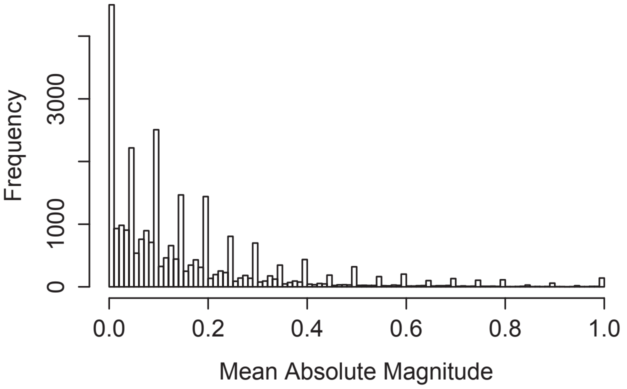


**Fig. A1.** Distribution of absolute update magnitude.

### *A.2. Sensitivity analyses for main hypothesis tests*

The analyses listed below in Tables A2.1–A2.4 characterize the association between updating and accuracy across different sets of forecasters and questions, as well as for different measures and update magnitude and accuracy. For more information on the core model specification, see manuscript Table 3.

Do frequent, small updates indicate superior performance even among the best forecasters? To find out, we performed the core analyses in superforecasters, all of whom received training and worked in teams of 12–15 individuals. In terms of overall Brier scores, the direction of coefficients replicated for frequency ($b = -0.053$, $t = -1.54$, $p = .123$), magnitude ($b = 0.100$, $t = 2.90$, $p = .004$), and confirmation propensity ($b = 0.041, t = 1.47, p = .143$). Only magnitude coefficients were significant at conventional levels. In summary, the associations of magnitude and frequency with overall performance were directionally replicated among the elite cohort of superforecasters; and small updates predicted stronger performance overall, while higher frequency was linked to better final-forecast accuracy.

The lack of statistical significance for frequency and confirmation propensity was mostly due to the smaller sample size, as the superforecaster sample was approximately one third the size of the core sample (N = 175 for superforecasters vs. N = 515 in the core sample). The regression coefficient for the frequency – accuracy association was approximately a third smaller in absolute terms ($b = -0.053$ in superforecaster sample vs. $b = 0.079$ in the core sample), perhaps pointing to diminishing returns of additional forecast updates. The strength relationship between magnitude and accuracy, as measured by the absolute value of regression coefficients, was almost identical in the superforecaster ($b = 0.100$) vs. the core sample ($b = 0.098$). The regression coefficient for the confirmation propensity – accuracy association was slightly higher among superforecasters ($b = 0.041$) than in the core sample ($b = 0.028$).

In summary, small magnitude, and to a lesser extent high frequency and lower confirmation propensity, were effective indicators of forecasting skill at the high-end of the skill continuum.

**Table A2.1**
Varying inclusion criteria based on minimum number of questions with updates.

| DV: Standardized Brier Score | At least 2 questions | At least 6 questions | At least 10 questions[1] | At least 20 questions |
|---|---|---|---|---|
| Intercept | −0.024 (0.013) | −0.039 (0.015)** | −0.050 (0.015)** | −0.073 (0.016)** |
| Frequency | −0.068 (0.009)** | −0.079 (0.011)** | −0.079 (0.011)** | −0.075 (0.013)** |
| Magnitude | 0.057 (0.010)** | 0.089 (0.011)** | 0.098 (0.012)** | 0.067 (0.012)** |
| Confirmation propensity | 0.015 (0.019) | 0.025 (0.010)* | 0.028 (0.011)** | 0.033 (0.012)** |
| Training | −0.081 (0.018)** | −0.070 (0.020)** | −0.063 (0.021)** | −0.073 (0.024)** |
| N | 832 | 598 | 515 | 352 |
| $R^2$ | 0.15 | 0.22 | 0.28 | 0.27 |

*Note:* * $p < 0.05$; ** $p < 0.01$.

1. Identical to base model shown in Table 3, column A of the manuscript. Duplicated here for ease of comparison.

**Table A2.2**
Sensitivity analysis on base model in a sample of superforecasters working in teams.

| DV: Standardized Brier Score | A. Base Model |
|---|---|
| Intercept | 0.079 (0.028)** |
| Frequency | −0.053 (0.034) |
| Magnitude | 0.100 (0.034)** |
| Confirmation propensity | 0.041 (0.027) |
| Training | |
| N | 175 |
| $R^2$ | 0.13 |

*Note:* * $p < 0.05$; ** $p < 0.01$.

**Table A2.3**
Sensitivity analysis: using mean-debiased Brier score[1] rather than standardized Brier score as an accuracy measure and squared distance rather than absolute distance as a measure of update magnitude.

| | A. Mean Debiased Score | B. Mean Standardized Score | C. Base Model[2] |
|---|---|---|---|
| Intercept | −0.017 (0.004)** | −0.044 (0.014)** | −0.050 (0.015)** |
| Frequency | −0.023 (0.004)** | −0.085 (0.011)** | −0.079 (0.011)** |
| Magnitude - Absolute | 0.029 (0.004)** | | 0.098 (0.012)** |
| Magnitude - Squared | | 0.073 (0.010)** | |
| Confirmation propensity | 0.010 (0.003)* | 0.030 (0.010)** | 0.028 (0.011)** |
| Training | −0.015 (0.007)* | −0.072 (0.020)** | −0.063 (0.021)** |
| N | 515 | 349 | 515 |
| $R^2$ | 0.25 | 0.33 | 0.28 |

*Notes:* * $p < 0.05$; ** $p < 0.01$.
1. Mean-debiased scores are calculated by subtracting the mean raw Brier score across all participants on a question from individual raw Brier scores.
2. The base model uses absolute magnitude as a predictor of standardized Brier scores. This specification is identical to the one shown in Table 3 in the manuscript, and is duplicated here for ease of comparison.

**Table A2.4**
Predicting accuracy separately by question type.

| DV: Standardized Brier Score | Status Quo & Timing | Non-Status Quo or No Timing |
|---|---|---|
| Intercept | −0.053 (0.020)** | −0.172 (0.029)** |
| Frequency | −0.053 (0.014)** | −0.045 (0.023)* |
| Magnitude | 0.087 (0.015)** | 0.028 (0.022) |
| Confirmation propensity | 0.015 (0.013) | 0.038 (0.022) |
| Training | −0.078 (0.026)** | −0.015 (0.042) |
| N | 482 | 165 |
| $R^2$ | 0.16 | 0.03 |

*Note:* * $p < 0.05$; ** $p < 0.01$.

Do small, frequent updates indicate accuracy advantages across a wide range of stimuli? We divided questions based on two properties. First, we differentiated between questions for which timing was essential for determining the outcome and those for which timing was not directly relevant. Timing is relevant to questions of the form “Will event X occur by date Y?” but not, for example, to the question “Which candidate will win election in country Z?” Second, we differentiated among questions with status quo outcomes, those that resulted in change outcomes, and cases when there was no relevant status quo. Status quo outcomes were cases in which the event did not occur. For elections, status quo pertained to selection of incumbent candidate, if available.

For questions where timing was pertinent and the outcome was status quo, forecasters would have generally benefited from updating with the passage of time. The question about Bashar al-Assad in the manuscript illustrates this point: if Bashar al-Assad is still in power six months after the question was posed and two weeks before the question was scheduled to close, that implies a very low probability of the event. On the other hand, mechanistic updating for the passage of time would hurt accuracy on questions that result in event occurrence (i.e., change outcome).

Status quo was not relevant to all questions. For those that could be categorized as status quo versus change, 77% resulted in status quo outcomes. Timing was relevant to 92% of questions. The combination of timing relevance and status quo (TRSQ) outcome characterized 73% of questions. We applied the same cross-sample validation procedure as in the core analysis. Updating behaviors on all available questions within a subsample were used to predict accuracy on each type of questions (timing-relevant & status quo outcomes vs. all others). Forecasters needed to answer at least 10 questions of a given type to be included in this analysis. While $N = 482$ forecasters met this inclusion criterion for the larger TRSQ question set, only $M = 165$ met this criterion for the other set.

Small update magnitude was predictive of better accuracy in both specifications, reaching significance at $p < .01$ for the TRSQ question set, but not in the other set. Frequency was significantly predictive of accuracy in the TRSQ question set, and directionally predictive in the other set. Confirmation propensity was not a significant predictor of accuracy in either question set, but yielded directionally consistent estimates.

### A.3. Accuracy over time and by forecast order

This section describes the relationship between updating behavior and accuracy for forecasts at different time-points: first versus last forecast on a question for a given forecaster; during early, middle and late periods within a question; and by forecast order.

What are frequent updaters’ main sources of edge? To examine this question, we calculated the accuracy of forecasts by order in which they were placed. More specifically, we categorized forecasts as first, second, third, or fourth for a given forecaster on a given question. The final “5 + ” category includes forecasts that were placed fifth or later. We use a randomly chosen half of questions (training set) for each forecaster to obtain the classification of low- versus high-frequency updater, and then calculated Brier scores for the other half of questions (validation set). Scores were averaged first within forecaster, and then across forecasters.

As seen in Fig. A3, frequent updaters tended to achieve better (lower) scores conditional on forecast order for the first four forecasts. Note however that this plot visually understates the accuracy advantage of frequent updaters for two reasons. First, frequent updaters place their second, third, and later forecasts earlier in a question duration, when the forecasting questions are more difficult to predict due to higher uncertainty. For example, a frequent updater may place their second forecast on a question on day 30 out of the 100-day duration, while the infrequent updater may place her second forecast on day 60. The analysis in Fig. A3 does not adjust for this. Second, update frequency is highly reliable across questions, so at any point in time, frequent updaters are, on average, further to the right on the updating curve. For example, 23% of all observations for frequent

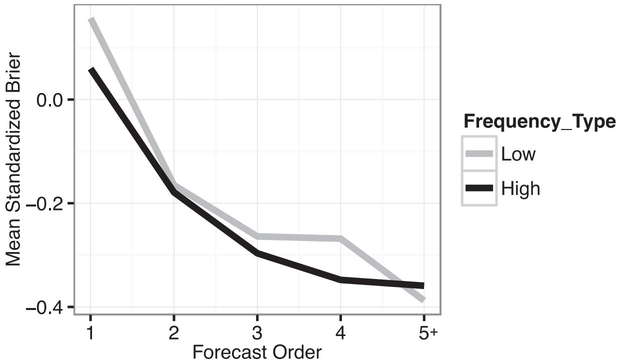


**Fig. A3.** Mean standardized Brier scores for validation set of questions by forecaster type (high versus low frequency of updating in the training set), and forecast order. The analysis excludes confirmations.

**Table A3.1**
Predicting accuracy for first and last forecast per question. Standard errors in parentheses.

| DV: Standardized Brier Score | A. First Forecast | B. Last Forecast |
|---|---|---|
| Intercept | −0.020 (0.015) | −0.131 (0.016)** |
| Frequency | 0.007 (0.011) | −0.181 (0.012)** |
| Magnitude | 0.121 (0.011)** | 0.044 (0.012)** |
| Confirmation propensity | 0.030 (0.011)** | 0.045 (0.012)** |
| Training | −0.039 (0.021) | −0.042 (0.023) |
| N | 515 | 515 |
| $R^2$ | 0.22 | 0.37 |

*Note:* * $p < 0.05$; ** $p < 0.01$.

**Table A3.2**
Predicting accuracy in the start, middle and end periods within a question (also see Fig. 2 in manuscript). Standard errors in parentheses.

| DV: Standardized Brier Score | A. Start Period | B. Middle Period | C. End Period |
|---|---|---|---|
| Intercept | −0.067 (0.017)** | −0.007 (0.023) | −0.050 (0.023)** |
| Frequency | −0.036 (0.013)** | −0.085 (0.018)** | −0.119 (0.017)** |
| Magnitude | 0.101 (0.013)** | 0.091 (0.017)** | −0.001 (0.017) |
| Confirmation propensity | 0.023 (0.012) | 0.071 (0.017)** | 0.067 (0.016)** |
| Training | −0.045 (0.024) | −0.023 (0.032) | −0.019 (0.032) |
| N | 515 | 515 | 515 |
| $R^2$ | 0.17 | 0.14 | 0.10 |

updaters fell in the “5+” category whereas less than 1% of observations are in this category for the infrequent updaters. The nested linear models reported in Table A2.3 show the predictors of individual forecast accuracy, accounting for within-question timing. A comparison between t-statistics for coefficients on “High-Frequency” forecaster type versus “Forecast Order” shows that forecaster type is at least as predictive of the accuracy of an individual forecast as the order in which it was placed (see Table A3.3).

**Table A3.3**
Predicting the accuracy of individual forecasters by the order in which a forecaster placed them; t-statistics listed in parentheses (also see Fig. A3 above).

| DV: Standardized Brier Score at Forecast Level | A. Without Magnitude | B. With Magnitude |
|---|---|---|
| Intercept | 0.381 (28.82) | 0.331 (19.00) |
| *Forecaster Characteristics* | | |
| High-Frequency | −0.114 (−6.32) | −0.078 (−5.15) |
| High-Magnitude | | 0.101 (4.29) |
| *Forecast Characteristics* | | |
| Forecast Order (1–5) | −0.012 (−4.28) | −0.011 (−4.19) |
| Time within Question (0–1) | −0.726 (−77.11) | −0.726 (−77.08) |
| N Forecasters | 515 | 515 |
| Forecaster Fixed Effects | Yes | Yes |
| AIC | 318,292.8 | 318,282.8 |

Notes:
(1) Forecaster characteristics assessed in training set of questions, accuracy assessed in the validation set.
(2) Forecast order excludes confirmations. Updates after the 5th one are imputed a value of 5.
(3) Time within question is scaled, so 0 denotes start date and 1 denotes closing date of question.

### *A.4. Forecaster calibration*

Calibration plots provide a visual depiction of forecast behavior, allowing us to characterize if forecasters in different categories are well-calibrated, under- or over-confident. Forecasts are divided in ten bins, each with approximately ten percentage points on the probability scale. Percentage values denote the weighted proportion of forecasts in each bin. For further discussion of calibration and discrimination, see manuscript Table 4 and related text description (see Fig. A4).

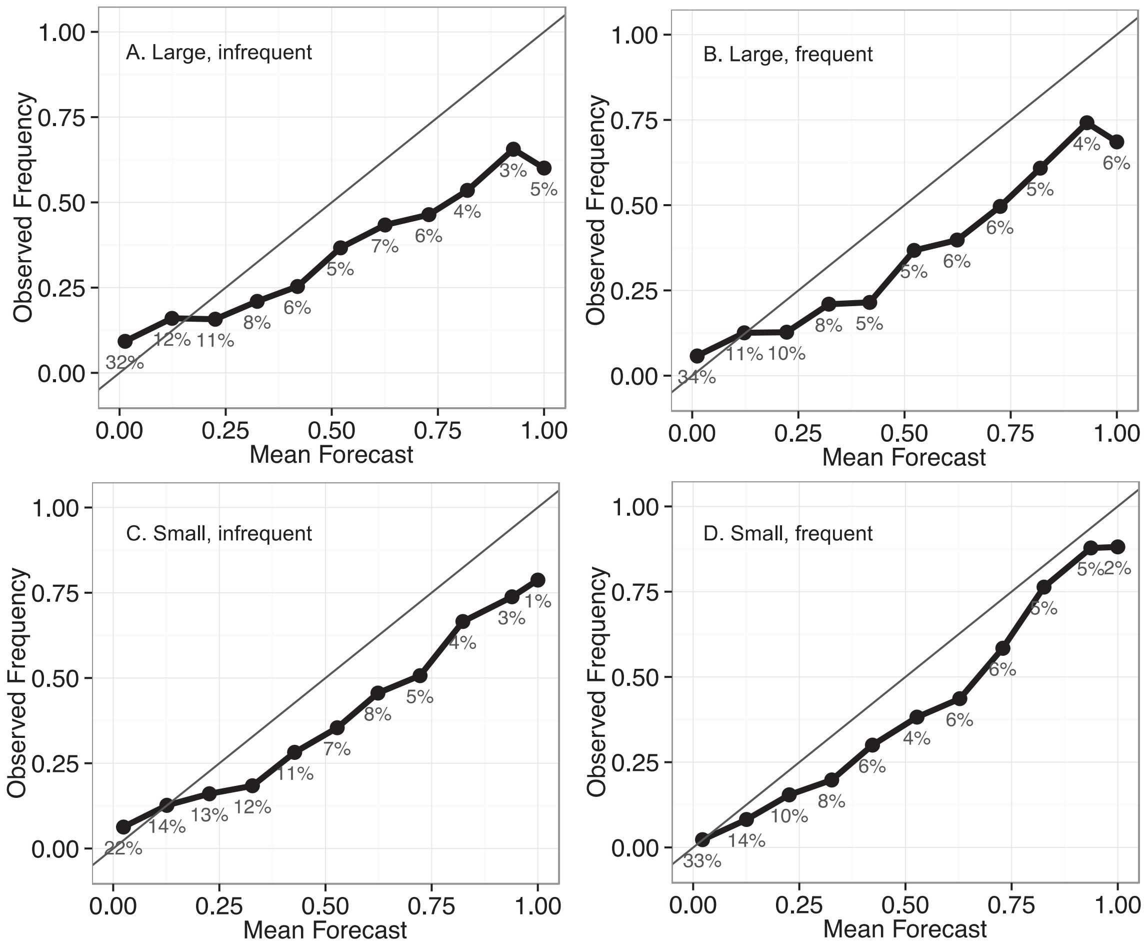


**Fig. A4.** Calibration plots for individual forecasts by updating pattern category.

### *A.5. Bayesian updating and volatility*

Method: Augenblick and Rabin (2017) described a model in which the volatility of a forecasting stream, calculated as the sum of squared deviations between prior and posterior beliefs, must equal the uncertainty in initial beliefs. This condition must be satisfied in expectation for a forecasting stream to adhere to the martingale property of Bayesian forecast streams. This is expressed as follows:

$$\sum_{t=1}^{T} (\pi_t - \pi_{t-1})^2 = \pi_0(1 - \pi_0) \tag{3}$$

where $\pi_t$ is a probability belief at time *t*, and $\pi_0$ is the initial belief. More extreme beliefs are associated with less uncertainty to be resolved. Less uncertainty in initial beliefs should correspond to less volatility in the subsequent updating stream.

To obtain a measure of adherence to the Bayesian standard, we can take the difference between volatility and uncertainty, so that positive values denote a forecast stream with excessive volatility, negative values denote insufficient volatility, and zero denotes that the forecast stream perfectly matches this Bayesian standard. We henceforth refer to this measure as excess volatility, which encodes the direction of positive values. We calculated excess volatility for each individual question that a forecaster attempted, and averaged this measure across all available questions for a given forecaster.

Result: We assessed the correlations between excess volatility and updating measures. As shown in the left panel of Fig. A5, there was little association between updating frequency and excess volatility. As a simple illustration, we performed a median split analysis, separating forecasters with higher-than-median vs. lower-than-median frequency. We then performed single-sample t-tests to compare excess volatility measures to the optimal level of zero. Frequent updaters—those making more than the median 1.6 forecasts per question—tended to produce forecast streams with excess volatility (excess volatility mean = 0.024, single sample *t* (*256*) = 4.03, $p < 0.001$). Infrequent updaters produced even higher levels of excess volatility (excess volatility mean = 0.044, *t (257)* = 8.16, $p < 0.001$). Incremental updaters, those with average absolute update magnitude below the median of 19%, produced forecasts that with only slightly elevated volatility levels (excess volatility mean = 0.006, *t (257)* = 1.91,

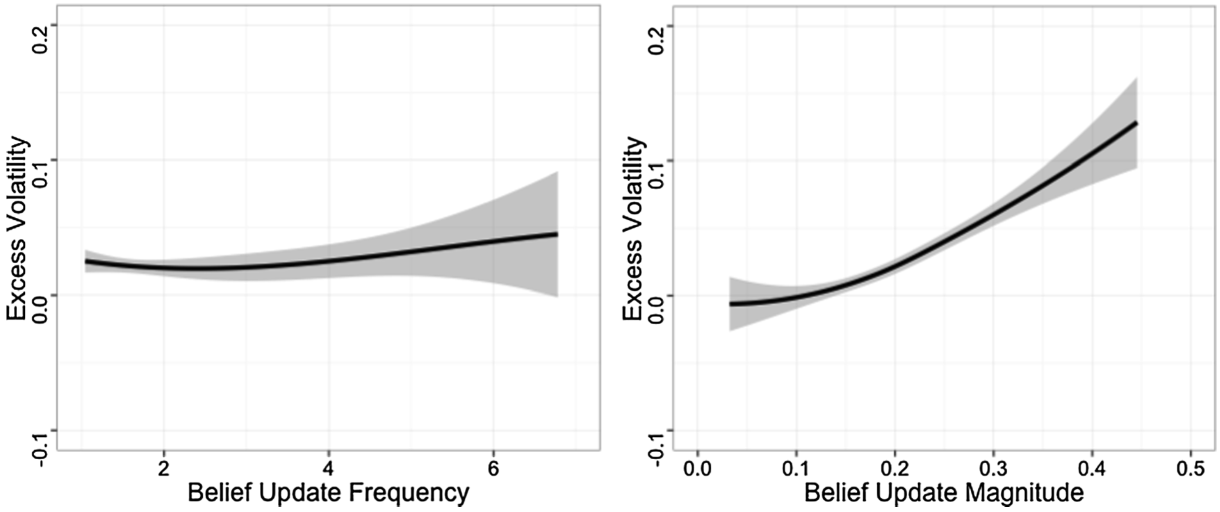


**Fig. A5.** Excess volatility of forecast streams for forecasters with varying belief updating frequency and magnitude. The Bayesian benchmark is denoted by zero excess volatility.

$p = 0.06$), while large-increment updaters had the highest average level of excess volatility (mean = 0.074, $t(256) = 7.78$, $p < .001$). Incremental updaters thus came closest to this standard of Bayesian updating.

### A.6. Relationship between updating and other behavioral/psychometric measures

In the analyses listed below, update frequency and magnitude across all questions are used as the dependent variable, while behavioral and psychometric predictors are utilized as predictors (see Table A6).

**Table A6**
Predictors of updating frequency and magnitude. All continuous variables, independent or dependent, are standardized prior to entry in model. Standard errors in parentheses.

| Predictor Set | Activity | | Psychometrics | |
|---|---|---|---|---|
| DV | Frequency | Magnitude | Frequency | Magnitude |
| Intercept | 0.023 (0.0050) | 0.058 (0.073) | −0.084 (0.069) | −0.113 (0.072)** |
| Frequency | | −0.608 (0.080)** | | −0.366 (0.049)** |
| Magnitude | −0.278 (0.036)** | | −0.356 (0.050)** | |
| Training | −0.030 (0.068) | −0.098 (0.100) | 0.058 (0.097) | −0.200 (0.098)* |
| Number of questions | 0.160 (0.039)** | 0.384 (0.054)** | | |
| Number of sessions† | 0.707 (0.041)** | 0.391 (0.083)** | | |
| Number of news links clicks†† | 0. 113 (0.040)** | −0.004 (0.060) | | |
| Fluid intelligence | | | −0.010 (0.048) | −0.139 (0.048)** |
| Political knowledge | | | 0.180 (0.048)** | 0.037 (0.050) |
| AOMT | | | 0.123 (0.048)* | 0.065 (0.049) |
| N | 294 | 294 | 380 | 380 |
| Adj. $R^2$ | 0.67 | 0.30 | 0.17 | 0.15 |

*Notes:* * $p < 0.05$, ** $p < 0.01$.
† Divided by the number of questions attempted by a forecaster.
†† Assessed only based on activity in Seasons 2 and 3 of the tournament, for which session and news link click data were available.

### A.7. Mediation analyses: training, updating and accuracy

The results of the mediation analyses below address the question if update measures mediate the relationship between training accuracy. Training condition was randomly assigned. In such cases, the approach is referred to as causal mediation analysis (see Tables A7.1–A7.3).

**Table A7.1**
Training, frequency and accuracy.

| DV: Standardized Brier Score | Estimate | 95% CI Lower | 95% CI Upper | p-value |
|---|---|---|---|---|
| Mediation Effect (ACME) | −0.017 | −0.033 | 0.001 | 0.07 |
| Average Direct Effect | −0.082 | −0.120 | −0.042 | < 0.01 |
| Total Effect | −0.098 | −0.140 | −0.055 | < 0.01 |
| Proportion Mediated | 0.167 | −0.015 | 0.353 | 0.07 |

*Note:* Based on N = 515 forecasters, 1000 model iterations.

**Table A7.2**
Training, magnitude and accuracy.

| DV: Standardized Brier Score | Estimate | 95% CI Lower | 95% CI Upper | p-value |
|---|---|---|---|---|
| Mediation Effect (ACME) | −0.029 | −0.052 | −0.001 | 0.01 |
| Average Direct Effect | −0.069 | −0.107 | −0.030 | < 0.01 |
| Total Effect | −0.098 | −0.140 | −0.054 | < 0.01 |
| Proportion Mediated | 0.293 | 0.089 | 0.548 | 0.01 |

*Note:* Based on N = 515 forecasters, 1000 model iterations.

**Table A7.3**
Training, confirmation propensity and accuracy.

| DV: Standardized Brier Score | Estimate | 95% CI Lower | 95% CI Upper | p-value |
|---|---|---|---|---|
| Mediation Effect (ACME) | 0.001 | −0.002 | 0.004 | 0.80 |
| Average Direct Effect | −0.098 | −0.140 | −0.055 | < 0.01 |
| Total Effect | −0.098 | −0.139 | −0.053 | < 0.01 |
| Proportion Mediated | −0.001 | −0.048 | 0.22 | 0.80 |

*Note:* Based on N = 515 forecasters, 1000 model iterations.

## Appendix B. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.obhdp.2020.02.001.

## References

Andrade, P., & Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics, 60*(8), 967–982.

Arrow, K. (1982). Risk perception in psychology and economics. *Economic Inquiry, 20*(1), 1–9.

Ashton, A., & Ashton, R. (1988). Sequential belief revision in auditing. *Accounting Review, 63*(4), 623–641.

Atanasov, P., Rescober, P., Stone, E., Swift, S., Servan-Schreiber, E., Tetlock, P. E., … Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science, 63*(3), 691–706.

Augenblick, N., & Rabin, M. (2017). *Belief movement, uncertainty reduction & rational updating.* Working Paper, University of California Berkeley.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233.

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

Birnbaum, M. H. (1983). Base rates in Bayesian inference: Signal detection analysis of the cab problem. *The American Journal of Psychology, 96*(1), 85.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151.

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

Castellan, N. J., Jr (1973). Multiple-cue probability learning with irrelevant cues. *Organizational Behavior and Human Performance, 9*(1), 16–29.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. E. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making, 11*(5), 509–526.

Chang, W., & Tetlock, P. E. (2016). Rethinking the training of intelligence analysts. *Intelligence and National Security, 31*(6), 903–920.

Chen, Y., Chu, C., Mullen, T., & Pennock, D. (2005). Information markets vs. opinion pools: An empirical comparison. *Proceedings of the 6th ACM conference on electronic commerce* (pp. 58–67). .

Clemen, R., & Winkler, R. (1999). Combining probability distributions from experts in risk analysis. *Risk Analysis, 19*(2), 187–203.

De Bondt, W., & Thaler, R. (1985). Does the stock market overreact? *Journal of Finance, 40*(3), 793–805.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126.

DiNicolantonio, J., Lucan, S., & O'Keefe, J. (2016). The evidence for saturated fat and for sugar related to coronary heart disease. *Progress in Cardiovascular Diseases, 58*(5), 464–472.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.). *Formal representation of human judgment* (pp. 17–51). New York: Wiley.

Ehrig, T. R. (2015). The revision of beliefs underlying organizational expectations. *Academy of management proceedings: Vol. 2015, No. 1*, (pp. 14988–). Briarcliff Manor, NY 10510: Academy of Management.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*(1), 3–23.

Fitzgerald, F. S. (2009). *The crack-up.* New York, NY: New Directions Publishing.

Fried, J. (2012). Some advice from Jeff Bezos. Accessed on Sep 12, 2019 at https://signalvnoise.com/posts/3289-some-advice-from-jeff-bezos.

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

Goldstein, D. G., McAfee, R. P., & Suri, S. (2014). The wisdom of smaller, smarter crowds. *Proceedings of the fifteenth ACM conference on economics and computation* (pp. 471–488). .

Goel, S., Reeves, D. M., Watts, D. J., & Pennock, D. M. (2010). Prediction without markets. *Proceedings of the 11th ACM Conference on Electronic Commerce* (pp. 357–366). ACM.

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

Harvey, N. (1988). Judgmental forecasting of univariate time series. *Journal of Behavioral Decision Making, 1*(2), 95–110.

Hogarth, R. M. (1991). A perspective on cognitive research in accounting. *Accounting Review, 66*(2), 277–290.

Jain, K., Bearden, J., & Filipowicz, A. (2013). Do maximizers predict better than satisficers? *Journal of Behavioral Decision Making, 26*(1), 41–50.

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55*(4), 582–590.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1–17.

Kuhn, T. S. (1954). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Larrick, R., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science, 52*(1), 111–127.

Leibovitch, M. (2015). You and I change our minds. Politicians 'evolve.' The New York Times Magazine, March 15, 2015. Accessed on September 13, 2019 at https://www.nytimes.com/2015/03/15/magazine/you-and-i-change-our-minds-politicians-evolve.html.

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

Lord, C., Lepper, M., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*(6), 1231.

Mannes, A., Soll, J., & Larrick, R. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276.

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and over-reaction. *Management Science, 51*(6), 932–947.

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

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S., Ungar, L., & Tetlock, P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied. 21*(1), 1–14.

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

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review, 115*(7), 1330–1338.

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall167.

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J., & Mertz, C. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review, 64*(2), 169–190.

Pittampalli, A. (2016). *Persuadable: How great leaders change their minds to change the world.* New York, NY: HarperCollins Publishers.

Redd, S. B. (2002). The influence of advisers on foreign policy decision making. *Journal of Conflict Resolution, 46*, 335–364.

Ries, E. (2011). *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses.* New York, NY: Crown Books.

Shiller, R. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review, 71*(3), 421–436.

Silver, N. (2016). Who will win the presidency? Accessed on September 12, 2019 at https://projects.fivethirtyeight.com/2016-election-forecast/.

Simon, H. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.

Tetlock, P., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology. 57*(3), 388–398.

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

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

Tingley, D., Yamamoto, T., Hirose, K., Imai, K., & Keele, L. (2014). mediation: R package for causal mediation Analysis. *Journal of Statistical Software, 59*(5), 1–38.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.

Witkowski, J., Atanasov, P., Ungar, L., & Krause, A. (2017). Proper proxy scoring rules. *Proceedings of the 31st AAAI conference on artificial intelligence* (pp. 743–749). .

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives, 18*(2), 107–126.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*(1), 1–13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83*(2), 260–281.

Yates, J. F., Price, P. C., Lee, J., & Ramirez, J. (1996). Good probabilistic forecasters: The 'consumer's' perspective. *International Journal of Forecasting, 12*, 41–56.

Zachary, R., Paulson, M., & Gorsuch, R. (1985). Estimating WAIS IQ from the Shipley Institute of Living Scale using continuously adjusted age norms. *Journal of Clinical Psychology, 41*(6), 820–831.