Chapter 6
Talent Spotting in Crowd Prediction

Pavel Atanasov and Mark Himmelstein

Keywords Forecasting · Prediction · Crowdsourcing · Skill assessment

1 Introduction

Since Francis Galton's classic demonstration (1907), wisdom-of-crowds research has largely focused on methods for eliciting and aggregating estimates, while treating the skill of individual forecasters as exogenous. For example, Mannes et al. (2014) define the wisdom-of-crowds effect as the tendency for the average estimate to outperform the average individual forecaster. Davis-Stober et al. (2014) generalize this definition to include any linear combination of estimates and randomly chosen forecaster as the comparison point.

In this chapter, we summarize a complementary line of research that has thrived over the last decade—the search for skilled forecasters. The general idea is that accounting for individual forecasting skill is valuable in maximizing crowd accuracy. Research on superforecasting at the Good Judgment Project (GJP, Mellers et al., 2015a, b) has demonstrated identifying and cultivating highly skilled forecasters is a crucial lever in maximizing crowd wisdom. More recent work has shown that the skill of the forecasters making up the crowd may be more important to aggregate accuracy than the choice of elicitation or aggregation methods (Atanasov et al., 2022b). Many aggregation methods are flexible enough to incorporate performance weights (Atanasov et al., 2017; Hanea et al., 2021).

These superforecasters were famously identified using a single measure: performance rank at the end of each forecasting season, which generally lasted approximately 9 months and featured over 100 questions. Performance was measured using

P. Atanasov (✉)
Pytho LLC, Brooklyn, NY, USA
e-mail: pavel@pytho.io

M. Himmelstein
Department of Psychology, Fordham University, Bronx, NY, USA

M. Seifert (ed.), *Judgment in Predictive Analytics*, International Series in Operations Research & Management Science 343,
https://doi.org/10.1007/978-3-031-30085-1_6

Brier scores in prediction polls and market earnings in prediction markets. Given sufficient resources, a tournament designer—one who is tasked with collecting and scoring predictions—would be well-advised to follow this strategy: start with thousands of forecasters, pose 100 or more questions that resolve in the subsequent 6–9 months, and after all outcomes are known, pick the top 2% performers.

However, not all settings lend themselves to such an extensive pursuit of forecasting talent. For example, a forecasting tournament may feature questions that do not resolve for many years. Alternatively, tournament designers may lack the resources to wait several months or pose 100+ imminently resolvable questions to provide sufficiently reliable performance scores. In this chapter, we describe methods for identifying skilled forecasters in such time- or resource-limited environments.

We seek to make three main contributions to the research literature. First, we take stock of skill identification research, with a focus on ideas from the last decade, and propose an organizing schema for the various skill predictor measures.[1] It consists of five categories: accuracy-related, intersubjective, behavioral, dispositional, and expertise-based measures. Second, we provide a quantitative summary of effect sizes in pre-existing studies, expressed as correlation coefficients between predictor measures and performance outcome measures. Third, we address measurement challenges inherent to cross-study comparisons by conducting a new analysis of GJP data in which we recreate a subset of the predictor measures across the five categories, following the analytical framework originally developed in Atanasov et al. (2020b).

This chapter does not aim to develop a unified theory of what makes a great forecaster. Our main goal is descriptive: to summarize existing ideas, measures and evidence. However, we hope that this review will prove helpful developing deeper theoretical understanding of drivers of forecasting skill. The relationships we examine are correlational, not causal, so the recommendations stemming from this work are primarily relevant to tournament organizers—who collect and organize forecasts—and less relevant to forecasters working to improve their craft.

This chapter relates to several research strands within the judgment and decision making literature. Our outcome of interest, predictive skill, is conceptually related to decision making competence (Bruine de Bruin et al., 2007) and may share common correlates. Separately, research based on the Brunswikian lens model has examined the accuracy of expert predictive judgments, using regression-style models as benchmarks (Blattberg & Hoch, 1990; Stewart et al., 1997; Seifert et al., 2015). Our focus here is on relative forecaster skill. The ideas tested here may serve as a starting point for the task of identifying forecasters who outperform or add value to model-based estimates.

At a practical level, improved understanding of talent spotting measures can help tournament designers in at least three ways. First, skill measures can be used in

[1] We refer to measures correlating with skill as predictors or correlates. To avoid confusion, we refer to individuals engaged in forecasting tasks as forecasters.

forecast weighting aggregation schemas, as discussed in this volume (Collins et al., 2022). Second, skill measures allow tournament designers place forecasters into smaller, selective, high-performing crowds (Mannes et al., 2014; Goldstein et al., 2014), such as superforecasting teams (Tetlock & Gardner 2016). Lastly, accurate forecasters tend to also excel at other quantitative and logical challenges (Mellers et al., 2017), so talent identified in forecasting tasks may be beneficially deployed in other analytically challenging contexts.

1.1 Definition of Skill

When we use the term *talent spotting,* we do not mean to suggest that excelling in forecasting tournaments is mostly a matter of inborn talent. Rather, we use talent as a synonym of *consistently demonstrated forecasting skill*: the tendency to display consistently strong relative performance on forecasting tasks. Consistency in this context means that that performance is assessed across many questions, which reduces the importance of luck. In prediction polls, the method that is of primary focus here, accuracy, is usually measured using proper scoring rules such as the Brier scoring rule (Brier, 1950). Scores are usually modified to account for the varying difficulty of questions. The modifications range in complexity, from z-score normalization to item-response theory (IRT) modeling. In prediction markets, the core performance measure is market earnings. Performance is generally compared against peers—forecasters recruited in similar ways and assigned to equivalent experimental conditions.

Put simply, this chapter is focused on answering the question: "Who is good at prediction?" We note the subtle difference between this performance-focused question, and related research that focuses on the question of *expertise* in prediction or foresight (Mauksch et al., 2020). We treat expertise and forecasting skill as two conceptually and empirically distinct concepts (Burgman et al., 2011). Namely, we use expertise measures as correlates of forecasting skill, not as outcome measures. Experts can be identified by reviewing resumes, while skilled forecasters are those who demonstrate strong performance in settings where accuracy is rigorously tracked, such as forecasting tournaments. The extent to which expertise relates to forecasting skill is an empirical question.

1.2 Five Categories of Skill Correlates

We classify all measures in the literature in five categories: accuracy-related, intersubjective, behavioral, dispositional, and expertise-related. The order of the categories presented here roughly corresponds to the strength of their relationship to forecasting skill.

A*ccuracy-related* measures all measures that rely on ground-truth resolutions (e.g., determination of whether or not a predicted event actually occurred). Accuracy is calculated based on the distance between a forecaster's estimates and ground-truth resolutions. Such measures fall in the general rubric of correspondence. The simplest versions rely directly on proper scoring rules (Gneiting & Raftery, 2007) such as quadratic Brier scores (Brier, 1950), and logarithmic scores (Good, 1952). Cross-forecaster comparison involves skill-based variants of proper scores, or standardized proper scores (e.g., Mellers et al., 2014). Item-response theory (IRT) models offer a more sophisticated approach, yielding estimates of two parameters: question discrimination ability and forecaster skill (Bo et al., 2017; Himmelstein et al., 2021; Merkle et al., 2016). Calibration and discrimination examine different facets of accuracy and can be obtained using Brier score decomposition (Murphy & Winkler, 1987).

Other measures in this category do not focus on accuracy directly but still rely on resolution data. Augenblick and Rabin (2021) develop a measure of Bayesian updating that gauges if time-series forecasts are characterized by insufficient or excessive volatility. Finally, Budescu and Chen (2015) proposed the use of contribution scores, which measure the extent to which including a person's forecast in an aggregate improves or reduces the aggregate accuracy.

Intersubjective measures do utilize ground-truth resolution data, which makes them suitable for settings in which outcomes are unverifiable or verification is delayed. Instead, individual forecasts verified based on their relation to consensus estimates, i.e., aggregate responses by peers. Measures like *proper proxy scoring rules* (Witkowski et al., 2017) can be applied to simple probability forecasts without requiring additional reports. In the canonical implementation, forecasts are scored by their squared distance from the aggregate (consensus) forecasts obtained from the same crowd, but other proper measures (e.g. logarithmic, spherical) can be used instead of squared distance. Surrogate scoring rules are conceptually similar, but rely on a model of forecast generation (Liu et al., 2020). Similarity-based measures (Kurvers et al., 2019) discretize probabilistic forecasts and calculate proportional agreement.

Other intersubjective measures depend on additional reports submitted by forecasters. For example, a forecaster may be asked to report her estimate of peer responses (e.g., the proportion of peers who select a given option, or the mean response across all peers). These methods include peer prediction (Miller et al., 2005), the Bayesian Truth Serum (Prelec, 2004), the Robust Bayesian Truth Serum (Witkowski & Parkes, 2012), and minimal pivoting (Palley & Soll, 2019). We do not discuss results from these elicitation mechanisms due to the additional report requirement.

Behavioral predictors measure what forecasters do on a forecasting platform. We distinguish among six sub-categories: activity, belief updating, extremity, coherence, rationale properties and question selection. Activity measures indicate how engaged forecasters are with the task, and include the number of forecasts, number of questions predicted, number of logins, time spent on forecasting platform, and news links clicked. It is generally expected that more active forecasters will perform better.

Belief updating measures describe how forecasters update their forecasts over time. Atanasov et al. (2020b) distinguish between three measures: frequency (how often updates occur per question), magnitude (how large is the average update in absolute terms) and confirmation propensity (how often forecaster re-enter their most recent forecast). Rationale-based measures rely on text analysis of the rationales that forecasters write on the platform. We also treat probabilistic extremity (how close a given probability forecast is to 0 or 1) as a behavioral measure of confidence, which we distinguish from self-reported expertise assessments used to assess calibration (see section on expertise below). Probabilistic coherence scores reflect the extent to which actual forecasts differ from logically and probabilistically coherent sets (Fan et al., 2019; Karvetski et al., 2013). Forecasters' choices about which questions to answer and which ones to skip may also be used as skill signals (Bennett & Steyvers, 2022).

Dispositional predictors are generally collected before or after the forecasting tournament, and involve psychometric tests. These aim to measure stable individual differences in fluid intelligence, thinking styles, and personality. Measures that closely relate to fluid intelligence include numeracy, cognitive reflection, matrix completion, number series completion, verbal and analytical aptitude. Thinking styles measures capture concepts such as active open-mindedness, need for cognition, need for closure and fox-hedgehog tendencies. Personality-type measures include the Big 5: conscientiousness, openness to experience, neuroticism, extraversion and agreeableness.

Expertise-based measures relate to the forecasters' knowledge and experience in the subject matter domain. *Demonstrated* knowledge is often measured using multiple-choice tests. Mellers et al. (2015a, henceforth Mellers et al. 2015b) describes knowledge scores as measures of crystallized intelligence. Such tests can also assess meta-knowledge, i.e., calibration, which relate the confidence expressed versus the rate of accurate responses (e.g., see the classical method, Cooke, 1991; Aspinall, 2010). *Biographical* expertise measures can generally be found on forecasters' resumes. These include education level, field of study, professional activities, publications and media appearances. Many of these measures were first described in Tetlock's (2005) book *Expert Political Judgment*. *Self-reported* expertise *measures* gauge how confident forecasters feel about their knowledge on a topic or about their predictive skills more generally.

This chapter consists of two studies. In Study 1, we review existing literature with the goal of providing a broad overview of skill identification measures. We first summarize all ideas in more detail, following the five-category structure, then report the correlation coefficients between prediction measures and accuracy outcomes. We do not provide a formal meta-analysis, mainly because the wide range of research designs makes such estimates tricky to aggregate or compare. Study 2 aims to address this comparability issue and provide more in-depth coverage: we reconstruct a subset of measures across each category and assess their correlations with accuracy, both in-sample and out-of-sample, using the same data and a uniform analytical framework originally developed and described in Atanasov et al. (2020b).

2 Study 1

2.1 Study 1: Methods

2.1.1 Literature Search

Articles of interest featured at least one of two elements: (a) new descriptions of predictive skill identification methods and measures, and (b) new empirical analyses featuring new or previously described measures. Relevant articles were identified in a four-step process.

First, we identified an initial set articles which we had read, reviewed or co-authored over the past decade. Second, we conducted literature searches, featuring search terms in two categories: (a) forecaster, forecasting, prediction, prediction, foresight, tournament; (b) talent, skill, performance, accuracy, earnings. Additional search terms included the award numbers for IARPA's Aggregative Contingent Estimation (ACE) and Hybrid Forecasting Competition (HFC) forecasting tournaments. Third, relevant articles that referenced or were referenced in the set compiled in the first two sets were added. Finally, we added several sources identified by peers. All in all, we identified over 40 individual measures from over 20 manuscripts from the above sources.

2.1.2 Outcome Variables

The core outcome variable in most studies was based on the Brier score (Brier, 1950). Although other proper scoring rules were mentioned in the literature, in practice, nearly all studies featured a version of the quadratic Brier scoring rule. We define one variant, mean standardized mean of daily Brier scores (MSMDB) in detail as it is used in both Study 1 and Study 2.

For any given forecast on a given day, the Brier score is the squared difference between probabilistic forecast and the ground truth, coded as 1 if event in question does occur, and 0 otherwise.

$$DB_{f,q,d} = 2\left(p_{f,q,d} - y_q\right)^2 \tag{6.1}$$

The daily Brier (DB) score for forecaster f, on question q on date d *is* twice the squared distance between the probability forecast p and the ground-truth outcome y (coded as 1 if event occurs, 0 otherwise). This result is a score that ranges from 0 (perfect accuracy) to 2 (reverse clairvoyance), with a 50% binary forecast earning a DB of 0.5. Mean daily Brier score is obtained by averaging Daily Brier scores across days within a question.

$$MDB_{f,q} = \frac{\sum_{d=1}^{D_q} DB_{f,q,d}}{D} \tag{6.2}$$

Standardized MDB (SMDB) is calculated as the difference between the forecaster's Mean Daily Brier score and the Mean of Mean Daily Brier scores across forecasters in a given condition, divided by the standard deviation of MDBs across these forecasters.

$$SMDB_{f,q} = \frac{MDB_{f,q} - \overline{MDB}_q}{SD\left(MDB_{f,q}\right)} \tag{6.3}$$

Accuracy across subsets of questions, *s*, SMDB scores can be averaged into a Mean SMDB (MSMDB) score. Variants of MSMDB are used in Mellers et al. (2014) and Atanasov et al. (2020b).

$$MSMDB_{f,s} = \frac{\sum_{q=1}^{Q_f} SMDB_{f,q}}{D} \tag{6.4}$$

Normalized Brier Score: $NBS_{f,\,q,\,d}$ refers to the *normalized accuracy* of the forecast made by a forecaster for a given question on a given date. It is a transformation of the Brier Score (or, put another way, a linking function) to make $\epsilon_{f,\,q,\,d}$ approximately normally distributed when used as an outcome measure in models which rely the assumption of normally distributed residuals. This variant is used in Himmelstein et al. (2021) and Merkle et al. (2016), as well as in the calculation of IRT scores in Study 2. In the original formulation, higher scores denote better accuracy. We reverse-code NBS to maintain consistency, so that all accuracy measures denote worse accuracy for higher values.[2]

$$\boldsymbol{NB_{f,q,d} = probit}\left(\mathbf{1} - \sqrt{\frac{\boldsymbol{DB_{f,q,d}}}{\mathbf{2}}}\right) \tag{6.5}$$

Delta Brier Score is based on the difference between the Brier score of the consensus estimate on a given question at a given time, and a forecaster's individual estimate for this question at this time. This version was used in Karvetski et al. (2021). It is reverse-coded in the current analysis, so that higher values denote worse accuracy, consistent with other outcome measures in this chapter.

[2] Normalization doesn't account for question difficulties on its own, just transforms the distribution. So, when used as criterion variables, normalized scores are then standardized: $SMNB_{f,q} = \frac{MNB_{f,q} - \overline{MNB}_q}{SD\left(MNB_{f,q}\right)}$.

2.1.3 Predictors of Skill

Study 1 provides an overview of measures and features that predict the skill level of forecasters. In order to avoid repetition between Study 1 and Study 2, we describe all measures here. All measures are summarized in Table 6.1 and detailed below, following the five-category structure.

2.1.3.1 Accuracy-Related

Raw Brier Score, Standardized Brier Score, Normalized Brier Score and Delta Brier Score measures also serve as outcome variables and are defined above. See Eqs. (6.1)–(6.5).

Item Response Theory Models: Item Response Theory (IRT) is a psychometric method for estimating latent traits, often latent ability or skill levels based on an objective measure, such as a standardized test or assessment (Embretson & Reise, 2013). Standardized tests, such as the SAT or GRE, are transformed based on an IRT estimation procedure.

The essential logic of IRT is that items on a psychological assessment instrument are not all created equal. Each item can carry unique diagnostic information people who answer it. For example, a very easy math problem may not be able to discriminate well between someone of moderate or high math ability, since either one would be very likely to get the item correct. However, it might be very well suited for discriminating between two different people of relatively low math ability, who would each have some chance of both getting the item right or wrong. Item Response Theory takes advantage of this by simultaneously estimating item-specific parameters that identify an item's unique diagnostic properties, as well as person-specific parameters that represent estimated ability levels.

Recent research has found that item response theory methods can be used to estimate the latent skill of forecasters based on the accuracy of their individual forecasts (Bo et al., 2017; Himmelstein et al., 2021; Merkle et al., 2016). This approach operates under the assumption that different forecasting problems convey different information about forecasters based on how accurate their forecasts are. Like standardized Brier scores, IRT assessment accounts for differences in question difficulty. It also allows questions to vary in discrimination—achieving a good raw score on some questions may be very informative about how good an individual forecaster is, while scores on other questions may not yield strong signal. There are additional features of the IRT approach that are especially appealing for assessing forecasting skill. Most crucially, IRT models are flexible enough to adjust for potential confounders. We describe one version of the model, which accounts for the role of time, in Appendix.

Calibration and *Discrimination* measures are facets of Brier score decomposition originally defined by Murphy and Winkler (1987). We use the operationalization in the context of individual forecasters in GJP, as described in Atanasov et al. (2020b).

Table 6.1 Description of forecasting skill identification measures

Measure	Brief description	Reference (Lead Author Year)
1. Accuracy-related		
Raw Brier Score	Strictly proper scoring rule, the squared distance between probability forecasts and ground-truth outcomes	Brier (1950)[a], Mellers et al. (2014)
Log Score	Log(p), where p is the probability estimate placed on the correct outcome	Good (1952)
Standardized Brier Score	Z-score transformed version of the raw Brier score; adjusts for question difficulty	Mellers et al. (2014), Atanasov et al., (2020b)
Normalized Brier Score	Transformed Brier score, see Appendix.	Himmelstein et al., (2021)
Calibration	Correspondence between predicted probabilities and observed base rates, Brier score decomposition component	Murphy & Winkler, 1987[a], Mellers et al. (2014)
Discrimination	Confidence of correct vs. incorrect forecasts, a.k.a. resolution, sharpness, Brier score decomposition component	Murphy & Winkler (1987)[a], Mellers et al. (2014)
Item Response Theory Models	Model-based estimate of forecaster skill, accounting for differences among questions	Himmelstein et al., (2021), Bo et al., (2017), Merkle et al., (2016)
Delta Brier	Difference in accuracy between an individual forecast and a contemporaneous consensus forecast	Karvetski et al., (2021)
Contribution Score	Difference in aggregate accuracy when a given individual is included vs. excluded from the aggregate	Budescu & Chen (2015)[a]
Excess volatility	Comparison of "measures of movement and uncertainty reduction given a Bayesian's changing beliefs over time." final beliefs correspond to the ground-truth outcome	Augenblick & Rabin (2021)[a]
2. Intersubjective		
Proper Proxy Scoring	Proxy scores are based on the distance between individual and consensus estimates; the latter are assumed to be unbiased	Witkowski et al., (2017)[a]
Surrogate scoring	Uses "noisy ground truth to evaluate quality of elicited information." Unlike proxy scoring, the noisy ground truth variable here is assumed to be biased	Liu et al., 2020[a]
Decision similarity	"The average percentage agreement of [the binarized forecasts of] this individual with all other N − 1 individuals."	Kurvers et al., (2019)[a]
Reciprocal Scoring	Forecasters are asked to estimate the consensus forecast from a large group of accurate forecasters; scored are based on the squared distance from consensus.	Karger et al., (2021)[a]

(continued)

Table 6.1 (continued)

Measure	Brief description	Reference (Lead Author Year)
Bayesian Truth Serum	Forecasters are asked to report both their true beliefs and their estimate of consensus belief. "The expected score [is a] measure of how much endorsing an opinion shifts others' posterior beliefs about the population distribution."	Prelec (2004)[a], Witkowski & Parkes, (2012)[a]
3. Behavioral		
3A. Activity		
Number of forecasts	Total number of forecasts entered	Mellers et al. (2015a)
Questions answered	Total number of questions with at least one forecast	Mellers et al. (2015a)
Number of sessions	Number of times a forecaster initiated a web session by logging in to a forecasting platform	Atanasov et al., (2020b)
Time on platform	Number of sessions multiplied by median session duration	Mellers et al. (2015a), Atanasov et al., (2020b)
News article clicks	Number of times forecasters clicked on unique news articles served in the forecasting platform	Atanasov et al., (2020b)
Training completion	Optional training completion, binary indicator on whether or not a forecaster completed an optional training module	Joseph & Atanasov (2019)
3B. Belief updating		
Update frequency	Number of forecasts per question, log-transformed	Mellers et al. (2015a)[a]
Update magnitude	Average absolute distance between subsequent forecasts, excluding confirmations	Atanasov et al., (2020b)[a]
Update confirmation propensity	Proportion of forecasts that actively confirm preceding forecasts	Atanasov et al., (2020b)[a]
3C. Other features		
Incoherence metric of Euclidean distance between observed responses and the closest coherent responses		Predd et al., (2008)[a], Karvetski et al., (2013), Collins et al., (2021)
Absolute distance from ignorance priors, normalized for number of answer options		Tannenbaum et al., (2017)
Rate of skipping impossible questions, i.e., questions with no correct answers		Bennett & Steyvers, (2022)[a]
3D. Rationale properties		
Average number of words or characters per rationale		Many
Flesch reading score uses features such as word and sentence length to determine the grade level proficiency needed to understand text		Zong et al., (2020)
Structural topic models discover sets of words that tend to occur together		Horowitz et al., (2019)
Integrative complexity, focus on the past, focus on the future, figures of speech		Karvetski et al., (2021), Zong et al., (2020)

(continued)

Table 6.1 (continued)

Measure	Brief description	Reference (Lead Author Year)
4. Dispositional		
4A. Fluid intelligence		
Number series	Correctness of open responses to ten questions involving number series completion	Dieckmann et al., (2017)[a], Himmelstein et al., (2021)
Numeracy	Berlin numerary: Computer adaptive score based on number of correct responses on up to three mathematical problems; others: % correct answers on mathematical problems	Cokely et al., (2012),[a] Lipkus et al., (2001)[a], Peters et al., (2006)[a]
Cognitive reflection	Original test included three mathematical questions, for which the obvious answers are incorrect. Extensions featured extra questions following this model.	Frederick (2005)[a], Toplak et al., (2014), Mellers et al. (2015a)
Inductive pattern recognition	Raven's progressive matrices test involves choosing one of six possible images to complete a series	Bors & Stokes (1998), Arthur et al., (1999)
Analytical intelligence	Shipley's analytical intelligence scale of Shipley–2 abstraction test	Shipley et al., (2009)[a]
Fluid intelligence	Equally-weighted combination of standardized scores from the available measures above	Mellers et al. (2015a), Atanasov et al., (2020b)
4B. Thinking styles		
Actively open minded thinking	Self-reported scale assessing the tendency to actively seek disconfirming information and keep an open mind	Baron (2000)[a], Haran et al., (2013), Mellers et al. (2015a)
Need for cognition	A self-reported tendency to structure relevant situations in meaningful, integrated ways	Cacioppo & Petty, (1982)[a]
Hedgehog-Fox	Hedgehogs see the world through a single big idea, while foxes use many perspectives. Multi-item self-report scale	Tetlock (2005)[a]
4C. Personality		
Conscientiousness	Personality trait reflecting the tendency to be organized, responsible, organized, hard-working and goal-directed	Costa & McCrae (2008)
5. Expertise		
5A. Demonstrated knowledge		
Knowledge test accuracy	Number of correct responses on binary or multiple choice questions about politics	Mellers et al. (2015a)[a], Himmelstein et al., (2021)[a]
Knowledge calibration	Difference between a forecaster's average confidence (subjective probability that answers are correct) and the proportion of correct responses on a knowledge test	Mellers et al. (2015a)
Classical method	The score includes calibration and information (discrimination) components, based on forecasters' confidence interval estimates for continuous values	Cooke (1991)[a], Aspinall (2010)

(continued)

Table 6.1 (continued)

Measure	Brief description	Reference (Lead Author Year)
5B. Biographical		
Fame	Frequency of engagement in policy advising, consulting and/or media appearances	Tetlock (2005)
Education	Advanced degree	Tetlock (2005)
h-index	Bibliographic measure of manuscript and citation counts	Benjamin et al., (2017), Atanasov et al., (2020a)
5C. Self-rated		
Self-rated expertise	Self-rating on scale from 1-not at all expert to 5-extremely expert	Mellers et al. (2015a)

Note: [a]Denotes that a source where the measure was first defined or operationalized, rather than just tested

Contribution Scores: There are many ways to define predictive skill. Most ground-truth-based approaches, such as proper scores and IRT assessments, involve assessing the accuracy of individual forecasts. A complementary approach is to ask how much individual forecasters contribute to the overall wisdom of the crowd. Framed differently: if you remove a given forecaster from a given crowd, how much does the crowd gain or lose in predictive accuracy? This is known as the *contribution*-based approach to predictive skill assessment (Budescu & Chen, 2015; Chen et al., 2016).

The contribution-based approach has a property that is appealing and absent from other ground-truth-based approaches to talent evaluation. Consider 20 forecasters who are equally skilled—each is likely to have the same amount of error in each of their forecasts as the those of the others. The first 19 all make an identical forecast for a given problem, while the 20th makes one that is very different. Because the first 19 contain redundant information, that information may be given more emphasis than the information from the 20th just by virtue of repetition, even though the 20th forecaster may have access to signals that are very useful.

This example highlights that even independent forecasters are often relying on redundant information to make their judgments (Broomell & Budescu, 2009; Palley & Soll, 2019). Having a strong consensus may indicate an informative signal, or it may be that the information shared between judges is creating shared bias, and the crowd wisdom would be improved with greater diversity (Davis-Stober et al., 2014). Assessing the contribution of individual analysts to the aggregate crowd judgment is a way to tease out which analysts are providing redundant information, and which are providing more unique information. For more details on calculating contribution scores, see Appendix.

Excess Volatility: This measure was originally defined by Augenblick and Rabin (2021) and is based on a comparison of "measures of movement and uncertainty reduction given a Bayesian's changing beliefs over time." Put simply, the more extreme the first judgment in a time-series, the smaller the subsequent updates should be. Question resolution is treated as a movement to $p = 1$ for the correct answer, and $p = 0$ for all other answer options. Thus, forecasters whose last reported

estimates tend to be inaccurate would earn higher volatility scores than those who make the identically sized updates but report more accurate final estimates. Augenblick and Rabin (2021) operationalized the measure in the GJP context and reported that most forecasters consistently exhibited excess volatility, i.e., larger-than-optimal cumulative movements, given the forecasters' starting points. In the original formulation, negative scores denote insufficient volatility while positive scores denote excess volatility. In a sensitivity analysis, we use absolute deviations from optimal volatility, so forecasters straying far from the Bayesian standards in either direction receive higher scores.

2.1.3.2 Intersubjective

Proper Proxy Scoring Rules: Proxy scores are based on the distance between individual forecaster estimates and relevant consensus estimates (Witkowski et al., 2017). The proxy scoring variant utilizing squared distance can be defined as follows:

$$\boldsymbol{DPrS_{f,q,d} = 2\left(p_{f,q,d} - c_{q,d}\right)^2} \tag{6.6}$$

In the current formulation the daily proxy score DPrS for forecaster q on question q on date d is calculated as the squared distance between the probability forecast $p_{f,q,d}$ and a consensus forecast on this question at that time, $c_{q,d}$. The consensus is constructed as the aggregate of individual estimates elicited from the same group of forecasters. In our new analysis (Study 2), we use the aggregation algorithm used in Atanasov et al. (2017) to produce consensus estimates. It features subsetting of the 72% most recent forecasts, higher weights placed on more frequent updaters on a given question, and an extremizing constant of $a = 1.5$. These parameters were not optimized to produce maximally accurate estimates or serve as an optimal basis for proxy score calculation. Thus, the current analysis is likely conservative, as optimized algorithms for constructing consensus estimates may improve the performance of proxy scores.

The original application by Witkowski et al. (2017) is forecast aggregation, and the measure is validated in the GJP context, where forecasters receive feedback in terms of objective Brier scores. The underlying idea is that wisdom-of-crowds consensus estimates are more accurate than most individuals, so forecasters whose independent estimates heave closer to the consensus are likely to be accurate. The main assumption is that consensus estimates are unbiased. The original definition does not pose constraints on the relative timing of individual and consensus forecasts or who makes up the peer group. In our analyses for Study 2, we compare contemporaneous individual and consensus forecasts that are based on the same group (condition) of forecasters. Consensus estimates may be improved by relaxing the contemporaneity constraint, or by sourcing consensus estimates from a group of forecasters with superior track records, e.g., superforecasters. Neither of those

modifications were employed here, which again makes our analyses of proxy scores' skill-spotting performance conservative.

A related variation, the expected Brier Score (EBS), is the average Brier score across each possible outcome weighted by the probability the crowd assigns to those outcomes (Himmelstein et al., 2023b). Formally,

$$\boldsymbol{EBS_{f,q,d}} \sum_{e=1}^{E} \boldsymbol{c_{f,q,d,e} DB_{f,q,d,e}}$$

Where $c_{f,\,q,\,d,\,e}$ is the probability assigned by the crowd to event e and $DB_{f,\,q,\,d,\,e}$ is the Brier Score the forecast would obtain if event e is realized as the ground truth.

Surrogate Scoring Rules: Surrogate scoring (Liu et al., 2020) is based on the similar underlying idea that consensus estimates are useful as departure points. Surrogate scoring models, however, build in the assumption that consensus forecasts are biased, and use "noisy ground truth to evaluate quality of elicited information" (p. 854). In practice, this makes surrogate scoring somewhat more complex to apply, as it involves the additional step of modeling the bias of the consensus crowd.

Decision Similarity: In a probabilistic elicitation context, decision similarity is assessed as "The average percentage agreement of [the binarized forecasts of] this individual with all other N − 1 individuals." (Kurvers et al., 2019; p. 2). Binarization involves transforming forecasts above 50% to 1 (i.e., 100%), and forecasts below 50% to 0. Binarized forecasts are then compared to the combined estimates made by all other forecasters. The authors use the measure in the context of skill identification and weighting, and do not test the incentive aspects of this schema. The measure is originally developed for non-probability contexts, where forecasters submit simple yes/no reports. The information loss stemming from binarizing forecasts make this measure suboptimal in the context of Study 2.

Reciprocal Scoring: This method incentivized forecasters to estimate the consensus forecast from a group of peers or a separate group of historically-accurate forecasters. Reciprocal scoring was defined and tested by Karger et al. (2021) mainly as an incentive schema, but the authors discussed how reciprocal scores may also serve as a skill identification or weighting measure. Forecasters in the reciprocal scoring only reported one set of estimates, as opposed to separate reports of personal vs. consensus beliefs.

Bayesian Truth Serum: This method was originally developed by Prelec (2004) and also applies to both resolvable and unresolvable questions. Respondents are asked about their own best guess about the true answer, as well as their estimate of the crowd's average answer. Responses are aggregated using the Surprisingly Popular algorithm, which boosts the likelihood of responses that are listed as correct more often than expected, based on the forecasters' consensus estimates. The method has shown to produce superior accuracy on questions where the obvious answer is incorrect. Witkowski and Parkes (2012) develop a version of this mechanism that applies to small crowds without common prior beliefs. Reciprocal scoring and Bayesian Truth Serum are not analyzed in Study 2, as they require additional reports from forecasters that are not available in the full GJP dataset.

2.1.3.3 Behavioral

Behavioral measures are generally sourced in the course of normal forecasting activities. Unlike accuracy-related features, they do not rely on question resolutions, and unlike intersubjective features, they do not involve comparisons of individual and consensus estimates.

Activity: Such measures assess forecaster task engagement and vary based on the specific features available in a forecasting platform. Activities measured on the GJP platforms included: the total number of forecasts submitted over the course of a season, the number of unique questions a forecaster answers (by reporting a probability estimate), the number of unique sessions on the forecasting platform, the per-session average or the total time spent on the forecasting platform, and the number of clicks on news articles served by the platform (Mellers et al., 2015a; Atanasov et al., 2020b).

Belief Updating: At first, updating measures simply measured forecast frequency: the mean number of estimates that a forecaster places on a given question (Mellers et al., 2015b). This measure was used as to determine forecast aggregation weights, where forecasters who submitted a larger number of estimates on a given question received higher aggregation weights (Atanasov et al., 2017). Later treatments distinguished among three separate aspects of belief updating (Atanasov et al., 2020b). First, *update frequency* is defined as the number of unique forecasts per question, which exclude forecast confirmations; if a forecaster submits two identical estimates on the same question in immediate succession, the latter is not counted. Frequency is log-transformed to reduce skew. Second, *update magnitude* is defined as the mean absolute distance among non-confirmatory estimates for a forecaster on a question. Third, *confirmation propensity* is the average proportion of all forecasts that confirm immediately preceding estimates. The original operationalization utilized forecasts on the first answer option on a question. In the current version, forecasts for all answer options are used to calculate update magnitude and confirmation propensity.

Probabilistic Coherence: Karvetski et al. (2013) define an incoherence metric as the "Euclidean distance between observed responses and the closest coherent responses." Examples of incoherent forecasts include ones for which the total probability across all answer options sums up to more or less than 100%, or forecasts that violate the Bayes rule, e.g., feature a combination of conditional and unconditional estimates that cannot be reconciled. In GJP, the platform interface forced forecast values to sum to 100%, and conditional forecasts were not elicited in ways that enable coherence assessment. In settings where forecasting activities do not enable coherence assessments, trait coherence can also be measured separately using an assessment tool specifically designed to identify analysts whose responses tend to be coherent (Ho, 2020; Budescu et al., 2021).

Probabilistic Extremity: The measure is based on the absolute distance from ignorance priors, normalized for the number of answer options. For a binary question, a 50%/50% would yield an extremity score of 0, while a forecast of 0%/100% would yield the highest possible extremity score. In Study 2, extremity is assessed exclusively based on the first estimate submitted by a forecaster on a

question. Highly attentive forecasters tend to update toward the extremes as uncertainty is reduced over time, so aggregating extremity across all forecasts would yield a measure that partly reflects belief updating tendencies. Using only the first forecasts is meant to distinguish confidence from belief updating. Tannenbaum et al. (2017) used a closely related measure to assess how forecasters predict on questions that vary in levels of perceived epistemic versus aleatory uncertainty.

Rationale Properties: In addition to eliciting quantitative forecasts, tournament platforms also provide space for text-based rationales where forecasters can explain the reasoning and evidence underlying their predictions. In some conditions, forecasters work as members of a team (Mellers et al., 2014), and sharing rationales can help team members coordinate, challenge one another, and otherwise contribute to team accuracy. Outside of the team context, the incentives for rationales are less clear: forecasters may be motivated to post detailed rationales in order to improve the overall predictions of the crowd, their own reputation in the forecasting community. In certain contexts, such as the Hybrid Forecasting Competition (Morstatter et al., 2019), rationales may also be analyzed to determine payment.

In the GJP independent elicitation condition that is the focus of Study 2, no specific incentives are provided for writing rationales, so it is possible that forecasters in that condition wrote rationales mostly as notes for their own use. Because of the sparsity of rationales in this condition, Study 2 does not feature linguistic rationale properties.

Linguistic properties of rationales vary in complexity. The simplest analyses focus on the rationale length, measured by the number of words or number of characters. More sophisticated natural-language processing techniques (NLP) include bag-of-words and topic modeling, which analyze which words and phrases tend to co-occur together (Horowitz et al., 2019; Zong et al., 2020).[3] NLP techniques have also been used to measure latent psychological factors, such as forecasters' tendency to consider base rates, to focus on the future versus the past, and to engage in complex thought patterns (Karvetski et al., 2021). The practice of gauging complexity of thought by analyzing written text predates automated NLP techniques (e.g., Suedfeld & Tetlock, 1977).

2.1.3.4 Dispositional

In both the ACE and HFC tournaments, a variety of dispositional variables were determined to be valid correlates of forecasting accuracy (ACE: Mellers et al. 2015b; HFC: Himmelstein et al., 2021). A major practical benefit of these results is that dispositional traits are a dimension of talent spotting that can be assessed *a priori*. Unlike ground-truth based and behavioral information, or even

[3] The authors were members of the SAGE team in the Hybrid Forecasting Competition. Linguistic properties of rationales were among the features used in aggregation weighting algorithms. The SAGE team the achieved highest accuracy in 2020, the last season of the tournament.

intersubjective approaches, no information about actual forecasting behavior, let alone ground-truth resolutions, are required to assess dispositional information. As a result, dispositional data can give talent-spotters a head start in picking out likely high performers before any forecasting starts (Himmelstein et al., 2021).

Dispositional information can be assessed with a battery of surveys. These can be broken into two classes, objective and subjective, sometimes referred to as performance-based and self-report-based, or as cognitive and non-cognitive in the psychometric literature (Bandalos, 2018). Objective surveys are akin to tests: they include math and reasoning problems with objectively correct answers, and can be scored based on how many responses were correct. Subjective surveys include self-reports that reflect how people view themselves.

Fluid Intelligence and Related Measures

Numeracy is considered a measure of statistical reasoning ability and risk literacy. It is most often measured with the four-item Berlin Numeracy scale (Cokely et al., 2012). It was employed during both ACE and HFC.

The *Cognitive Reflection Test (CRT)* is a measure of people's ability to reason reflectively in the presence of intuitively appealing, but incorrect answers. The original three-item measure (Frederick, 2005) has been expanded into longer versions, often containing 6–8 items (Baron et al., 2015; Toplak et al., 2014). Two versions of the CRT were administered during ACE (Mellers et al., 2015a). An extended version was administered during two HFC seasons (Himmelstein et al., 2021).

Matrix Reasoning tasks have long been staples of cognitive assessments, such as IQ tests. Matrix reasoning tasks are rooted in visual pattern matching. A series of shapes are displayed which contain a pattern of some sort, with one figure in the series left blank. Participants then must determine which among several choices of shapes would fit the pattern. A classic matrix reasoning scale, Raven's progressive matrices (Bors & Stokes, 1998), was administered during ACE (Mellers et al., 2015a). A newer matrix reasoning task, based on randomly computer-generated problems (Matzen et al., 2010), was administered during HFC Season 1 (Himmelstein et al., 2021).

Number Series is a more recently developed nine item scale, which has received less standalone psychometric validation than some of the others. The number series task is similar in structure to matrix reasoning, except featuring numerical patterns instead of visual patterns (Dieckmann et al., 2017). People are shown a series of numbers which follow a particular pattern, with one number missing, and must determine the missing value. The scale was administered in both HFC seasons (Himmelstein et al., 2021).

Thinking Style Measures are usually based on forecasters' self-report responses about the ways in which they think, behave and process information. The following four measures are included here:

Actively Open-Minded Thinking measures willingness to reason and accept information that is contrary to one's beliefs is necessary to forecast objectively.

See Baron (2000), Stanovich & West (1997). There are several versions of the scale. In the ACE study, a 7-item version was used (Haran et al., 2013, Mellers et al., 2015a).

Foxes and *hedgehogs* are defined as two poles of intellectual heterogeneity in Tetlock (2005). Hedgehogs represent people who tend to be highly specific in their expertise, while foxes tend to be more eclectic. Tetlock (2005) also found that foxes tended to be less overconfident in their predictions. In ACE, participants were asked a single self-report item about whether they would classify themselves as foxes or hedgehogs (Mellers et al. 2015a), as well as a 10-item scale; the latter is used in Study 2.

Need for Closure, which is conceptually opposed to open mindedness, is considered a hindrance to forecasting talent. People who have a higher need for closure will tend to more easily accept conclusions that conform with their preconceptions, while people with less need for closure will tend to seek counterfactual information. In ACE, an 11-item need for closure scale (Webster & Kruglanski, 1994) was included as a potential correlate of forecasting skill in ACE (Mellers et al., 2015a).

Need for Cognition measures people's willingness to engage in effortful reasoning behavior (Cacioppo & Petty, 1982), and was included in HFC Season 1 (Himmelstein et al., 2021).

Personality: Conscientiousness is a facet in the Big-Five personality inventory (Costa & McCrae, 2008), which describes a person's tendency to be organized, responsible, hard-working and goal-directed. Among the five facets, conscientiousness stands out as the one with most "consistent relations with all job performance criteria for all occupational groups" (Barrick & Mount, 1991). This result motivated us to include conscientiousness in Study 2, despite the lack of studies reporting its relation to predictive accuracy.

2.1.3.5 Expertise-Related

Expertise-related measures focus on forecasters' level of expertise with potential relevance to a given domain. By our definition, an expert is someone who demonstrates knowledge in a topic, has relevant educational or professional experiences, or considers themselves an expert. However, as originally noted by Tetlock (2005), an expert is not necessarily more accurate on a given topic than a non-expert. Again, we do not treat expertise and forecasting skill as synonymous with one another.

Demonstrated Expertise is usually assessed using subject-matter knowledge questionnaires. Mellers et al. (2015a) report on political knowledge questionnaires used in GJP, while Himmelstein et al. (2021) describes similar measures used in HFC. Expertise was measured as the proportion of correct responses, and probabilistic knowledge calibration. For example, a forecaster who places an average confidence of 80% that their answers are correct, but has an actual accuracy rate of 70% is considered as overconfident, while someone with the same 70% accuracy rate who places an average confidence of 60% is considered underconfident. Cooke (1991) originally developed a related so-called "classical method", which involves

elicitation of confidence intervals for continuous quantities. Individuals are then scored based on their knowledge calibration and resolution/sharpness. Forecasters in GJP did not provide estimates on continuous quantities, so measures based on the classical method are not used in Study 2.

Biographical measures can be assessed based on information forecasters would put in their resumes, such as educational level, educational specialty, professional experience, publication record and media mentions. Tetlock (2005) also uses several related measures of professional fame, such as the frequency with which individuals engage in media appearances, government or private sector consulting.

Self-Rated Expertise is assessed by asking the forecaster if they consider themselves an expert on the subject matter related to an individual forecasting question or a set of forecasting questions. In GJP, expertise was elicited on a 5-point scale, from 1-Not at all Expert to 5-Extremely Expert.

2.2 Study 1: Results

We identified 89 correlation measures across 16 manuscripts. Manuscripts that did not report correlations between predictors and accuracy measures are not discussed here. We organize the results following the five-category structure outlined above. To keep results consistent, we reverse-code outcome measures for which high scores denote better accuracy. After the reversal, higher scores for all measures denote larger errors, and thus lower accuracy. The median absolute correlation coefficient among all measures was $r = 0.23$; among all non-accuracy-related measures, it was $r = 0.20$. Correlation coefficients are tabulated in Table 6.2 and summarized visually in Fig. 6.1.

2.2.1 Accuracy-Related

As expected, accuracy-related (predictor) measures were most closely correlated with other accuracy (outcome) measures. The correlation coefficients exhibited variation across studies. On the high end, Atanasov et al. (2020b) used data and set-up very similar to that in Study 2 and found a correlation of $r = 0.75$ between standardized mean daily Brier scores (MSMDB) on one set of question and the same measure on another. Their sample included n = 515 forecasters across 4 seasons of the Good Judgment Project who answered a mean of 43 questions (SD = 35, Median = 32) per forecaster. The sample was split randomly in two question subsets, yielding a mean of 21.5 questions per forecaster for each subset.

Atanasov et al. (2020b) calculated cross-sample correlation differently: they tracked GJP forecasters across seasons, assessing the correlation of end-of-season leaderboard ranks between Seasons 2 & 3 (S1), and Seasons 3 & 4 (S2). In prediction polls, leaderboard rankings were based on Brier scores. The study also tracked performance rankings in prediction markets, which were based on end-of-season

Table 6.2 Correlations with outcome measures, where higher values denote larger errors, i.e., worse accuracy. Pearson's r coefficients reported, unless otherwise noted

Outcome variable	Predictor	Correlation Coefficient	Source
1. Accuracy-related			
Normalized Brier	Normalized, out-of-sample	0.54	Himmelstein et al., (2021), S2
Normalized Brier	IRT, out-of-sample	0.53	Himmelstein et al., (2021), S2
Normalized Brier	Normalized, out-of-sample	0.36	Himmelstein et al., (2021), S2
Normalized Brier	IRT, out-of-sample	0.30	Himmelstein et al., (2021), S2
Standardized Brier	Standardized Brier, out-of-sample	0.75	Atanasov et al., (2020b)
Brier-score rank	Brier-score rank, out-of-sample	0.37	Atanasov et al., (2022b), S1
Brier-score rank	Brier-score rank, out-of-sample	0.44	Atanasov et al., (2022b), S2
Market earnings rank	Market earnings rank, out-of-sample	0.25	Atanasov et al., (2022b), S1
Market earnings rank	Market earnings rank, out-of-sample	0.18	Atanasov et al., (2022b), S2
Long-term calibration	Short-term calibration	0.53	Tetlock (2005)
Long-term discrimination[a]	Short-term discrimination[a]	0.44	Tetlock (2005)
Calibration, out-of-specialty area	Calibration, in specialty area	0.39	Tetlock (2005)
Discrimination, out of specialty area[a]	Discrimination, in specialty area	0.31	Tetlock (2005)
2. Intersubjective			
Brier	Mean distance proxy score	0.66	
Brier	Mean distance proxy score, out-of-sample	0.44	
Brier	Mean expected Brier score	0.66	
Brier	Mean expected Brier score, out-of-sample	0.43	
Accuracy, balanced[a]	Similarity	−0.56[b]	Kurvers et al., (2019), S1
Accuracy, balanced[a]	Similarity	−0.83[b]	Kurvers et al., (2019), S2
Accuracy, % correct[a]	Similarity	−0.84[b]	Kurvers et al., (2019), S3
Accuracy, % correct[a]	Similarity	−0.84[b]	Kurvers et al., (2019), S4
3. Behavioral			
Standardized Brier	Number of questions attempted	0.07	Mellers et al. (2015a)
Standardized Brier	Deliberation time	−0.30	Mellers et al. (2015a)

(continued)

Table 6.2 (continued)

Outcome variable	Predictor	Correlation Coefficient	Source
Standardized Brier	Optional training completion	−0.21	Joseph & Atanasov (2019)
Standardized Brier	Frequency	−0.32	Atanasov et al., (2020b)
Standardized Brier	Magnitude	0.49	Atanasov et al., (2020b)
Standardized Brier	Confirmation	0.03	Atanasov et al., (2020b)
Standardized Brier	Frequency, out-of-sample	−0.32	Atanasov et al., (2020b)
Standardized Brier	Magnitude, out-of-sample	0.45	Atanasov et al., (2020b)
Standardized Brier	Confirmation, out-of-sample	0.03	Atanasov et al., (2020b)
Standardized Brier	Frequency	−0.49	Mellers et al. (2015a)
Delta Brier[a]	Rationale word count	−0.12	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale comparison class	−0.18	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale integrative complexity	−0.17	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale dialectical complexity	−0.20	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale elaborative complexity	−0.11	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale tentativeness	−0.17	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale focus on the past	−0.13	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale focus on the future	0.09	Karvetski et al., (2021), S1
Delta Brier[a]	Rationale word count	−0.22	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale comparison class IC	−0.32	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale integrative complexity	−0.28	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale dialectical complexity	−0.28	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale elaborative complexity	−0.23	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale source count	−0.25	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale use of quotes	−0.12	Karvetski et al., (2021), S2
Delta Brier[a]	Rationale focus on the future	0.21	Karvetski et al., (2021), S2
Overconfidence	Balance in rationales	−0.37	Tetlock (2005)
Correspondence error	Coherence error	0.68	Tsai & Kirlik, (2012)
Brier	Coherence forecasting scale (9-item)	−0.39	Budescu et al., (2021)
Brier	Coherence forecasting scale (18-item)	−0.50	

(continued)

Table 6.2 (continued)

Outcome variable	Predictor	Correlation Coefficient	Source
Brier	Impossible question criterion	−0.46	Bennett & Steyvers, (2022)
4. Dispositional			
Normalized Brier[a]	Number series	−0.15	Himmelstein et al., (2021), S1
Normalized Brier[a]	Number series	−0.30	Himmelstein et al., (2021), S2
Brier	Number series	−0.34	Budescu et al., (2021)
Normalized Brier[a]	Berlin numeracy	−0.15	Himmelstein et al., (2021), S1
Normalized Brier[a]	Berlin numeracy	−0.28	Himmelstein et al., (2021), S2
Brier	Berlin numeracy	−0.28	Budescu et al., (2021)
Standardized Brier	Numeracy	−0.09	Mellers et al. (2015a)
Brier	Subjective numeracy	−0.20	Budescu et al., (2021)
Normalized Brier[a]	Cognitive reflection test	−0.20	Himmelstein et al., (2021), S1
Brier	Cognitive reflection test	−0.28	Budescu et al., (2021)
Standardized Brier	Cognitive reflection test	−0.15	Mellers et al. (2015a)
Standardized Brier	Extended CRT	−0.14	Mellers et al. (2015a)
Normalized Brier[a]	Matrix reasoning	−0.15	Himmelstein et al., (2021), S1
Standardized Brier	Raven's progressive matrices	−0.23	Mellers et al. (2015a)
Normalized Brier[a]	Actively open minded thinking	−0.15	Himmelstein et al., (2021), S1
Normalized Brier[a]	Need for cognition	−0.15	Himmelstein et al., (2021), S1
Normalized Brier[a]	Actively open minded thinking	0.00	Himmelstein et al., (2021), S1
Standardized Brier	Actively open minded thinking	−0.10	Mellers et al. (2015a)
Standardized Brier	Need for closure	−0.03	Mellers et al. (2015a)
Calibration[a]	Fox-hedgehog	−0.35	Tetlock (2005)
Standardized Brier	Fox-hedgehog	0.09	Mellers et al. (2015a)
5. Expertise-related			
Normalized Brier[a]	Political knowledge % correct	−0.11	Himmelstein et al., (2021), S1
Normalized Brier[a]	Political knowledge, overconfidence)	0.15	Himmelstein et al., (2021), S1
Normalized Brier[a]	Political knowledge, % correct	−0.10	Himmelstein et al., (2021), S2
Normalized Brier[a]	Political knowledge, overconfidence)	0.14	Himmelstein et al., (2021), S2

(continued)

Table 6.2 (continued)

Outcome variable	Predictor	Correlation Coefficient	Source
Standardized Brier	Political knowledge, % correct	−0.18	Mellers et al. (2015a), Measure 1
Standardized Brier	Political knowledge, % correct	−0.20	Mellers et al. (2015a), Measure 2
Normalized Brier[a]	Education	−0.13	Himmelstein et al., (2021), S1
Normalized Brier[a]	Education	−0.03	Himmelstein et al., (2021), S2
Standardized Brier	Co-investigator (Y = 1, N = 0)	0.09	Atanasov et al., (2020a)
Standardized Brier	h-index	0.00	Atanasov et al., (2020a)
Brier score	h-index	−0.15	Benjamin et al., (2017)
Overconfidence	Fame/in-demand	0.33	Tetlock (2005)
Brier score	Confidence	0.20	Benjamin et al., (2017)
Calibration[a]	Self-rated relevance of expertise	−0.09	Tetlock (2005)

Notes: [a]Measures with asterisks were reverse-coded to maintain consistency. Positive correlation coefficients denote that higher levels of predictor variables are associated with larger error, i.e., worse predictive performance. Predictor measures were reverse-coded in cases where predictors and outcome measures were the same (e.g., discrimination)
[b]Denotes a Spearman's rank correlation coefficient

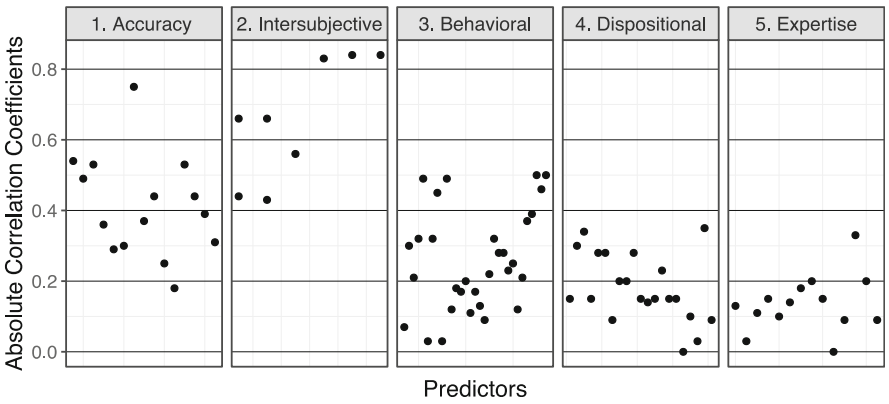

Fig. 6.1 Visual summary of absolute correlation coefficients for the five categories of predictors. Studies vary widely in design, outcome and predictor variables, so the figure aims to provide a general overview, not a detailed, self-sufficient summary of evidence. Average correlations are not reported. Horizontal axis coordinates for datapoints are random

market earnings. In prediction polls, leaderboard rank correlations were $r = 0.37$ between Seasons 2 and 3; $r = 0.44$ between Seasons 3 & 4. In prediction markets, rank correlations were lower: $r = 0.25$ between Seasons 2 and 3; $r = 0.18$ between Seasons 3 & 4. These results suggest that prediction polls' rankings based on Brier scores tend to be more reliable over time than earnings-based rankings in markets.

Himmelstein et al. (2021) used data from the Hybrid Forecasting Competition. Season 1 data was for 326 forecasters who were recruited openly on the web and who made at least 5 forecasts. Correlations were assessed between two sets of 94 questions, and the resulting cross-sample reliability was $r = 0.36$ for normalized Brier scores, slightly lower for IRT. Season 2 (uses data from n = 547 forecasters recruited through Amazon's Mechanical Turk. Cross-sample reliability was $r = 0.54$ for normalized Brier scores, again slightly lower for IRT. Tetlock (2005) reported correlations of calibration and discrimination measures in vs. outside of forecaster specialty area, and on long vs. short-term questions.

2.2.2 Intersubjective

Witkowski et al. (2017) introduced proxy scoring rules, which provide scores for individual forecasters based the distance between a forecaster's measure and the group's consensus. Instead of correlations, the validation takes the approach of pairing forecasters, comparing their proxy scores on one set of questions against their accuracy in a validation set. When the training set consists of 30 questions, the forecaster with the better proxy score achieves better accuracy scores in the validation set approximately in 65% of the comparisons. Thus, there appears to be a significant association between proxy and accuracy scores across question samples. The original analysis did not include correlation coefficients.

Himmelstein et al. (2023b) expanded on this work in a study in which 175 forecasters made predictions about 11 events related to politics, economics, and public health. Each forecaster forecasted each event five times at three-week intervals leading up to event resolution. Across all forecasts and time points forecasters' mean daily proxy scores (MDPrS) and mean expected Brier scores (MEBS) were significantly correlated with their actual MDB scores, r(EBS, MDB) = 0.66 and r(MDPrS, MDB) = 0.66. (Because each forecaster forecasted each question during each wave, Brier scores were not standardized.) To cross validate the results, the authors also split the 11 questions up into all 462 possible combinations of separate samples of 5 and 6 questions. They estimated cross-correlations between MDPrS and MEBS with MDB. The average out-of-sample correlations were r(MEBS, MDB) = 0.43 and r(MDPrS, MDS) = 0.44. The authors note that intersubjective scores were slightly more effective at discriminating poor performers than strong performers.

Liu et al. (2020) show that Surrogate Scoring Rules show slightly higher correlations with Brier and logarithmic scores in-sample than Proxy Scores across 14 datasets. Surrogate Scoring Rule also performed slightly better than Proxy Scoring in selecting top forecasters. However, correlation coefficients were not

reported. Kurvers et al. (2019) show that similarity scores strongly related to accuracy in-sample, with Spearman rank correlations ranging between $r_s = 0.56$ to and $r_s = 0.84$, the latter of which was based on GJP data. These data are included with the caveat that Spearman rank and Pearson correlation coefficients are not directly comparable.

2.2.3 Behavioral

Activity measures varied in their correlation with accuracy: Mellers et al. (2015a) reported that answering more questions was associated with worse SMDB accuracy ($r = 0.07$); while spending more deliberation time on the platform tended to correlate with better accuracy ($r = -0.30$). Using data from HFC, Joseph and Atanasov (2019) documented that when forecasters were given the option to review training materials, those who chose to complete training performed better than those who did not (Cohen's $d = 0.42$ for the full sample, converted to Pearson $r = -0.21$). Based on additional analyses and experimental data, they argued that the association between training and accuracy is primarily causal (training improves accuracy), and to a lesser extent a matter of self-selection (better forecasters choosing to engage in training).

Belief updating: More frequent updating corresponded to lower (better) MSMDB. The number of forecasts per question (update frequency) was moderately correlated with accuracy both in-sample and out-of-sample ($r = -0.32$ for both (Atanasov et al., 2020b). Mellers et al. (2015a) documented a stronger association between frequency and accuracy ($r = -0.49$). While both papers were based on GJP data, they used data from different seasons and slightly different selection criteria.

Update magnitude was among the strongest correlates of MSMDB accuracy measures, both in-sample ($r = 0.49$) and out-of-sample ($r = 0.45$), whereas incremental updaters tended to be more accurate than large-step updaters. Confirmation propensity was weakly correlated to accuracy on its own ($r = 0.03$, both in- and out-of-sample), but improved fit in multiple regression models that included frequency and magnitude. See Atanasov et al. (2020b).

Rationale Text Features: In his study of expert political judgment, Tetlock (2005) showed that forecasters who tended to produce forecast rationales with more balance (e.g., using terms like 'however' and 'on the other hand') tended to be less overconfident ($r = 0.36$). Similarly, a positive correlation was observed between integratively complex thought protocols and calibration ($r = 0.31$).

Karvetski et al. (2021) documented the relationship between linguistic properties of forecast rationales and accuracy in more recent forecasting tournaments, including ACE and the global forecasting challenge. In their study, the outcome is a Delta-Brier measure where higher scores denote better accuracy. We reverse-coded this measure, to maintain consistency with other Brier-based measures we review. In our reports, higher scores denote worse accuracy. Even simple word-count measures showed weak but consistent correlations ($r = -0.12$ to $r = -0.22$), as long-rationale writers tended to be more accurate. Frequent mention of reference classes was

among the strongest correlates of better accuracy scores ($r = -0.18$ to $r = -0.32$), as were measures of integrative ($r = -0.17$ to $r = -0.23$) and dialectic complexity ($r = -0.20$ to $r = -0.28$). Notably, using words and terms about the past related to better scores ($r = -0.13$), while using more terms about the future correlated with worse accuracy ($r = 0.09$ to $r = 0.21$).

In a separate analysis based on data from Good Judgment Open, Zong et al. (2020) also found that statements about uncertainty, and ones focused on the past relate to better accuracy, while statements focused on the future correspond with worse accuracy. Zong et al. (2020) also documented that absolute sentiment strength (positive or negative) tended to relate to worse accuracy. Usage of cardinal numbers and nouns predicted better accuracy, while frequency of verb usage predicted worse accuracy. In a second study, the authors analyze earnings forecast statements. The results regarding focus on the past and future were replicated; uncertainty terms correlated with worse accuracy. Zong et al., did not report correlation coefficients, only significance tests.

Coherence: Because coherence can be difficult to distill from the experimental designs common to forecasting research (i.e., forecasting tournaments), Ho (2020) developed an independent coherence assessment: the *coherence forecasting scale* (CFS). Himmelstein et al. (2023b) ran a performance test in which 75 forecasters each forecasted 11 questions at five different time points. The longer, 18-item version of the CFS was correlated with MDB, $r = -0.50$, while a shortened 9-item version exhibited slightly lower correlation, $r = -0.39$ (Budescu et al., 2021).

2.2.4 Dispositional

Fluid Intelligence: Out of all psychometric measures, fluid intelligence measures maintained the strongest and most consistent relationship with accuracy. Correlations for individual test measures generally fall in the range of $r = 0.15$ to $r = 0.30$. For example, the correlation of cognitive reflection test (CRT) with standardized Brier is $r = -0.15$ in Mellers et al. (2015a), while Himmelstein et al. (2021) document a correlation with normalized scores of $r = -0.20$, and Budescu et al. (2021) reported a correlation with raw Brier scores of $r = -0.28$. Similar patterns were observed for Number Series, Berlin Numeracy, and Matrix Reasoning. While each measure is conceptually distinct, individual measures appear to load well on a single underlying factor that Mellers et al. (2015a) called fluid intelligence.

Cognitive Styles: In contrast to fluid intelligence measures, which are objective (performance-based), thinking style measures rely on forecasters to self-reflect and report on their own proclivities. Among these measures, actively open-minded thinking was the only one that has been significantly linked with better accuracy: Mellers et al. (2015a) reported a small but significant correlation ($r = -0.10$) with standardized Brier scores; Himmelstein replicated this result in HFC Season 1 ($r = -0.15$), but not in Season 2 ($r = 0.00$). Tetlock (2005) documented a strong correlation between expert political forecasters' fox-hedgehog scores and their calibration

scores (r = 0.30), whereas foxier experts tended to be better calibrated. In the context of open forecasting tournaments, however, Mellers et al. (2015a) did not find a significant correlation between fox-hedgehog scores and standardized Brier scores (r = 0.09). Mellers et al., also showed that Need for Closure measure did not corelate to accuracy scores (r = 0.03).

2.2.5 Expertise-Related

Demonstrated Expertise: these measures focus on subject matter knowledge, e.g., political knowledge tests measuring how much forecasters know about topics covered in geopolitical forecasting tournaments. Mellers et al. (2015a) reported on two political knowledge measures collected across two seasons of GJP, where scores are based on the proportion of correct responses to multiple-choice questions. The reported correlations were $r = -0.18$ and $r = -0.20$. Using a similar measure (percentage correct responses to political knowledge questions), Himmelstein et al. (2021) reported somewhat lower correlations with normalized accuracy (r = −0.10 to r = −0.11). In addition to raw accuracy, Himmelstein et al., also computed calibration scores, where perfect calibration denotes that a forecaster's average confidence, expressed as probability (e.g., 60%), equals the proportion of correct responses. These calibration scores significantly correlated with normalized accuracy in both volunteer ($r = 0.14$) and MTurk ($r = 0.15$) samples.

Biographical: One measure of general expertise is education level. Himmelstein et al. (2021) show that in a volunteer sample (HFC Season 1), higher educational attainment significantly predicted normalized accuracy (r = −0.13), but that pattern did not hold among forecasters recruited on Mechanical Turk (r = −0.03, Season 2). In the life-sciences context, Atanasov et al. (2020b) showed trial co-investigators—physicians who worked on a specific trial—were slightly but not significantly less accurate than independent observers ($r = 0.09$) in predicting efficacy outcomes. This study also showed no correlation between bibliographic measure of research impact (h-index) and accuracy ($r = 0.00$). In a similar context Benjamin et al. (2017) reported a low but significant correlation between h-index and brier score accuracy (r = −0.15).

In the context of expert political judgment, Tetlock (2005) showed that experts' degree of fame, as measured by the experts' ratings of "how often they advised policy makers, consulted with government or business, and were solicited by the media for interviews." This fame measure correlated with overconfidence (r = 0.33), whereas more famous experts tended to be more overconfident. A similar correlation with overconfidence (r = 0.26) was observed for an alternative measure of fame, based on the number of media mentions. Finally, a self-rating of media contact frequency (0–never to 7–every week) had a low correlation with calibration (r = −0.12).

Self-Rated Expertise: We did not find published results of confidence or expertise self-ratings in the recent forecasting tournaments literature. In EPJ, Tetlock (2005) collected self-ratings of forecasters' relevance of expertise and found that the

correlation between these ratings and calibration was not statistically significant ($r = 0.09$).

2.3 Study 1 Discussion

The summary of measures is a reasonable starting point, and the correlation coefficients summarized in Table 6.2 and Fig. 6.1 provide a general sense of how skill identification measures relate to accuracy in the individual studies we summarize. However, these coefficients are not directly comparable across studies. This is partly because individual studies vary in the types of stimuli (forecasting questions) and forecaster samples. It is possible that some questions are better than others at measuring underlying skill. In fact, that possibility is a central motivation for developing IRT models. Separately, correlation coefficients may vary across samples. As a hypothetical example, imagine that a tournament only accepts forecasters with IQ scores above 140. Such a tournament will likely yield low correlations between IQ and accuracy. While it is possible to adjust statistically for such restricted-range effects (e.g., Bland & Altman, 2011), any such adjustments often rely on information that is unavailable to tournament designers, e.g., because some measures are not normed for a given population. Studies also used a variety of different outcome variables for defining accuracy.

Even within a study, where question types and forecaster samples are held constant, estimated correlation coefficients will increase with the *number* of forecasting questions used to estimate the accuracy outcome measure. Increasing the number of questions boosts the outcome measure's reliability, and thus its potential correlation with other measures. For example, a study that assesses accuracy on 100 questions will yield a larger correlation coefficient between, say, IQ and accuracy, relative to a study that uses the same IQ measure, but assesses accuracy based on a subset of 30 forecasting questions. In Study 2, we aim to address these measurement challenges by directly comparing a subset of skill identification measures within the same analytical framework.

3 Study 2

3.1 Study 2: Methods

3.1.1 Good Judgment Project Data

All Study 2 analyses are based on the data and analytical framework described in Atanasov et al. (2020b). We provide a brief overview. The ACE tournament featured 481 forecasting questions over four seasons. Questions lasted approximately 3 months on average (Median = 81 days, M = 110 days, SD = 93). Our sample

consists of N = 515 participants (forecasters) who made at least two forecasts on at least 10 forecasting questions over one or more seasons. These forecasters worked independently, not as members of forecasting teams. These forecasters made at least one forecast on over one hundred questions on average (M = 113, SD = 73), and made at least two forecasts on forty-three questions (M = 43, SD = 35). Forecasters made an average of two forecasts per question (M = 2.0, SD = 1.6).

Forecasters were scored based on mean daily Brier scores, as described above. Instead of standardizing scores, performance across questions of varying difficulty was equalized through imputation. First, once a forecaster placed an estimate, their forecasts were carried over across days until an update. Second, a forecaster placed an estimate after the first day that a question was open, their scores were imputed as the median daily Brier score across all their peers in a given condition. Third, if a forecaster skipped a question altogether, they received the median overall Brier score for their condition on a question. Overall, Brier scores were displayed on a leaderboard, which featured only peers within a condition. The top 2% forecasters in each condition were invited to work as superforecasters in the following season.

3.1.2 Cross-Validation and Outcome Variable Definition

For each forecaster included in the analysis, we randomly divided all questions they answered in two random question subsets. Let's call then A and B. Each subset consists of approximately half of the questions on which each forecaster placed an estimate. The subsets are randomly split for each forecaster, so that even if two forecasters answered the same set of questions, these questions will most likely be split differently When skill predictors in one subset are correlated with MSMDB (see Eq. 6.4) accuracy in the same subset (A and A, or B and B), we refer these as in-sample. When skill predictors in one question subset are correlated with forecasting accuracy in the other, we refer to these analyses as out-of-sample.

3.1.3 Predictor Selection

We used three main criteria to select predictors for further testing in Study 2: importance, data availability and fidelity. First, with regard to importance, we focused on skill predictors that demonstrated significant associations with accuracy in the literature, in either univariate or multivariate analyses. Second, data for some measures was either unavailable or insufficient. Insufficiency was the reason for excluding linguistic rationale data, as forecasters in our sample made independent forecasts and were not incentivized to write detailed rationales. Data availability also eliminated measures that require additional forecast reports from forecasters, such as Bayesian Truth Serum.

Third, fidelity concerns centered on our ability to reproduce the measures in the context of the current study. These involved contribution scores, as well as inter-subjective measures such as surrogate scores. For all of these, our initial examination

led to the assessment that the measures would be difficult to reproduce, as small details in decisions about the adaptation of the methods to our analytical framework may have large impacts on results. The fidelity criterion is admittedly subjective, and we see the benefits of including these measures in future research.

3.1.4 Statistical Tests

The core univariate analyses focus on the Person's *r* correlation coefficient between each predictor and mean standardized Brier scores (MSMDB). The univariate correlation analyses provide a useful starting point for examining the value of various predictors of skill. To provide useful recommendations for skill spotting in forecasting tournaments with limited resources, we need to go a step further. Namely, we need to understand which measures add the most value in the presence of others.

To address this need, we fit a series regularized LASSO regression models. Regularization involves penalizing complexity in model building, so that predictors are only included with non-zero coefficients if the improvements in fit overcome the penalty. The models we report follow ten-fold cross validation; these models prioritize sparsity, and are based on "the value of λ that gives the most regularized model such that the cross-validated error is within one standard error of the minimum" (Hastie et al., 2021, p. 5). All predictors were standardized before entry into the model, to distributions with mean zero and standard deviation of one. We report two runs for each model, one for each subset of questions. We include at least one predictor measure from each category: accuracy-related (out-of-sample mean standardized Brier scores), intersubjective (proxy scores), behavioral (update frequency, update magnitude, forecast extremity), dispositional (fluid intelligence composite scores, AOMT), expertise (knowledge test scores, advanced degree, self-assessed expertise ratings).

3.2 Study 2: Results

3.2.1 Correlational Analyses

We first report the univariate correlation coefficients with our core measure of accuracy: Mean Standardized Mean Daily Brier (MSMDB) scores. Results are organized according to the categories describe above. For question-specific measures, we report in-sample and out-of-sample correlations with accuracy. In-sample correlations are calculated for predictors and outcomes (MSMDB) assessed on the same set of questions.

We also report cross-sample reliability of each measure, also in the form of Pearson *r* coefficient. For the full sample of $n = 515$ forecasters, absolute values above $r = 0.10$ are statistically significant at $\alpha = 0.05$, for a two-tailed test, and those

Table 6.3 Correlations with accuracy (MSMDB) and reliability for predictors in Study 2

Predictor	Correlation with accuracy		Cross-sample reliability
	In-sample	Out-of-sample	
Standardized Brier (SB)	NA	NA	0.74
Debiased Brier	0.96	0.69	0.69
SB, first forecast	0.85	0.62	0.68
SB, last forecast	0.84	0.68	0.83
IRT forecaster score	0.30	0.22	0.89
Calibration	0.51	0.36	0.67
Discrimination	0.71	0.57	0.74
Excess volatility	0.40	0.30	0.85
Proper proxy, all forecasts	0.69	0.60	0.81
Proper proxy, first forecast	0.57	0.52	0.77
Number of questions	0.25	0.25	NA
Update magnitude, abs. Distance	0.51	0.45	0.75
Update frequency	−0.31	−0.32	0.98
Confirmation propensity	0.03	0.03	0.86
Extremity, first forecast	0.19	0.15	0.91
Fluid IQ composite all	NA	0.27	NA
Fluid IQ composite free	NA	0.28	NA
Political knowledge score	NA	−0.10	NA
AOMT	NA	0.10	NA
Fox-hedgehog scale	NA	0.08	NA
Conscientiousness	NA	0.13	NA
Education (advanced degree = 1)	NA	0.01	NA
Expertise self-ratings	0.00	0.00	0.97

above r = 0.12, are statistically significant at α = .01. To avoid repetition, we do not report p-values. We do report sample size only for predictors that are not available in the full sample. The cross-sampling procedure differs somewhat from Atanasov et al. (2020b), as the results reported here are based on two sampling iterations. Thus, the results reported here occasionally differ slightly (by r = 0.01 or less). Correlations with MSMDB and reliability coefficients are reported in Table 6.3. Cross-correlations among predictors are presented in Appendix Table 6.6.

3.2.1.1 Accuracy-Related Measures

The cross-sample reliability of standardized Brier scores (MSMDB) was $r = 0.74$. Notably, standardized Brier scores for the last estimate a forecaster made on a question had higher cross-sample reliability ($r = 0.83$) than those for first forecast (r = 0.68). This may be because last-forecast accuracy relates to updating effort, a reliable individual difference. IRT estimates exhibited a low correlation with

MSMDB ($r = 0.30$), suggesting that the two are distinct measures of skill. Notably, IRT estimates demonstrated high cross-sample reliability ($r = 0.89$).

In terms of Brier score decomposition, discrimination was more strongly correlated with overall MSMDB than calibration. Discrimination and MSMDB were strongly negatively correlated both in-sample ($r = -0.71$) and out-of-sample ($r = -0.57$), while calibration error and MSMDB were positively correlated in-sample ($r = 0.51$) and out-of-sample ($r = 0.36$), as expected. Discrimination ($r = 0.74$) and calibration error ($r = 0.67$) exhibited similar levels of cross-sample reliability. Overall, forecasters' discrimination scores were more strongly related to accuracy than their calibration scores. This result was consistent with a pattern where most forecasters are relatively well-calibrated, and the best forecasters mostly distinguish themselves through superior discrimination.

The Augenblick-Rabin measure of volatility exhibited high cross-sample reliability ($r = 0.85$) but was only moderately correlated with accuracy in-sample ($r = 0.40$), and out-of-sample ($r = 0.30$), where forecasters who produced time-series with more excess volatility tended to be less accurate. The core version of this measure was coded such that the results indicated that a forecaster exhibiting insufficient volatility would be expected to be more accurate than one exhibiting optimal levels of volatility, who in turn would be expected to be more accurate than one producing excessively volatile forecast series. As a sensitivity analysis, we calculated a different version of this measure in which we calculated absolute deviations from optimal volatility levels at the forecaster level, treating errors of excess volatility as equivalent to errors of insufficient volatility. Curiously, this absolute-distance-from-optimal-volatility measure had lower correlations with accuracy, both in-sample ($r = 0.29$), and out-of-sample ($r = 0.23$).

3.2.1.2 Intersubjective Measures

Proper proxy scores calculated based on all forecasts by a person on a question were highly correlated with MSMDB, both in-sample ($r = 0.69$) and out-of-sample ($r = 0.60$), whereas forecasters who tended to place independent estimates closer to the consensus were generally more accurate than those who strayed from the consensus. Even when proxy scores were calculated only based on the first forecast made by a forecaster on a question, the correlations remained very high in-sample ($r = 0.57$) and out-of-sample ($r = 0.52$). First-forecast proxy scores are useful as they are available as soon as a question is posed and several forecasters have placed their initial estimates. Proxy scores exhibited cross-sample reliability similar to that of accuracy: $r = 0.81$ for all-forecast proxy scores, and $r = 0.77$ for first-forecast proxy scores. Among predictors that could be calculated without the need for ground-truth question resolutions, proper proxy scores yielded the highest out-of-sample correlations with MSMDB. These results highlight the promise of

intersubjective measures in talent spotting, especially in settings where forecaster selection decisions must take place before questions resolutions are known.[4]

3.2.1.3 Behavioral Measures

Behavioral measures comprised the widest and most diverse category in the literature. We distinguished four sub-categories: general activity measures, belief updating, probabilistic confidence, linguistic properties of forecast rationales and coherence. The GJP user interface forced within-forecast coherence, so such a measure is not included here. Our analysis follows Atanasov et al. (2020b) in focusing on independently elicited forecasts, where forecasters had no incentive to write detailed rationales, so we do not include linguistic rationale properties.

Among activity measures, the number of questions a forecaster attempted was a predictor of worse performance, correlating with higher standardized Brier scores ($r = 0.25$) both in-sample and out-of-sample. In other words, forecasters who answered more questions registered worse accuracy. Cross-sample reliability of number of questions was not assessed as the sample was constructed by splitting questions into equal categories. In contrast, the number of questions with forecast updates was weakly correlated with better accuracy ($r = -0.10$).

Among belief updating measures, update frequency (the number of non-confirmatory forecasts per question), was the measure with the highest cross-sample reliability ($r = 0.98$), showing that individual differences in how often forecasters update are stable across questions. Update frequency was moderately correlated with accuracy both in-sample and out-of-sample ($r = -0.31$ and $r = -0.32$). Absolute update magnitude between forecast updates was also reliable ($r = 0.75$), and relatively highly correlated with MSMDB both in-sample ($r = 0.51$) and out-of-sample ($r = 0.45$). The positive signs denote that that forecasters who updated in small-step increments tended to register better (lower) accuracy scores. Confirmation propensity was highly reliable ($r = 0.82$), but it had a low correlation with MSMDB: $r = 0.03$ in-sample and $r = 0.03$ out-of-sample.

Probability extremity, the absolute distance between forecasts and the ignorance prior, was assessed based on the first estimate for a forecaster on a question. Extremity was negatively correlated with MSMDB, both in-sample ($r = -0.19$), and out-of-sample ($r = -0.15$), denoting that forecasters who tended to make more extreme (confident) probabilistic estimates tended to have better accuracy scores. Examination of correlations across predictors provides an interpretation for this result: forecasters' who exhibited higher probabilistic confidence tended to earn better discrimination scores ($r = 0.32$), but did not earn significantly worse calibration-error scores ($r = 0.05$).

[4] We do not offer complete coverage of intersubjective measures, including surrogate scores and similarity measures, but given our current results, further empirical investigation seems worthwhile.

3.2.1.4 Dispositional Measures

The strongest psychometric predictor of MSMDB accuracy was a fluid intelligence. Our composite measure was calculated as an equal-weight combination of available standardized scores on Berlin Numeracy, Cognitive Reflection, Raven's Progressive Matrices and Shipley's Analytical Intelligence test (Cronbach's alpha = 0.62). This Fluid IQ measure was negatively correlated with MSMDB (n = 409, $r = -0.27$). The first two measures (Berlin Numeracy and CRT) are freely available, while the last two are commercially available for a fee. A combination of the freely available measures yielded lower in reliability (Cronbach's alpha = 0.43), but similar correlations with MSMDB (n = 408, $r = -0.28$). Thus, it does not appear that the available-for-purchase fluid intelligence measures add value in terms of predicting MSMDB measures of accuracy. Actively open minded thinking (AOMT) measure had moderately low internal reliability (Cronbach's alpha = 0.64). AOMT scores yielded a marginally significant correlation with SMDB (n = 379, $r = -0.10$).

Fox-hedgehog measure scores (Cronbach's alpha = 0.31) were positively but not significantly correlated with SMDB (n = 311, $r = 0.08$). The positive sign indicates that forecasters who rate themselves as hedgehogs tend to have worse accuracy. Conscientiousness measure was high in internal reliability (Cronbach's alpha = 0.81) and scores were positively correlated with SMDB (n = 311, $r = 0.13$), indicating that forecasters who rated themselves as more conscientious tended to perform worse.

3.2.1.5 Expertise Measures

Demonstrated Expertise: Political Knowledge (PK) test scores were reliable across GJP Seasons 1, 2 and 3 (Cronbach's alpha = 0.75). The combined PK test scores were marginally correlated with SMDB (n = 409, r = −0.10). Zooming in on individual tests, PK scores from Season 2 (n = 263, r = −0.24) yielded somewhat higher correlations with SMDB than did PK scores in Season 1 (n = 281, r = −0.18) and Season 3 (n = 323, r = −0.10). The forecaster sample of PK test completers differs across seasons, making cross-season comparisons less direct. For each of Season 1 and Season 2, the overall PK scores (measured as the number of correct responses) were as good or better predictors of accuracy than calibration and discrimination measures based on the same tests.

Biographical: The most general measure of demonstrated expertise was education, coded as binary variable indicating whether the forecaster had obtained an advanced (post-Bachelor) degree or not. This binary indicator was uncorrelated to SMDB (r = 0.01).

Self-Rated Expertise: Self-ratings of forecasters' own relevance of expertise in the question domain were highly reliable across question sets, indicating that some forecasters tended to exhibit consistently higher confidence in their own expertise than others ($r = 0.97$). However, expertise self-ratings were completely uncorrelated

with accuracy (SMDB) in-sample (n = 404, $r = 0.00$), and out-of-sample (n = 404, $r = 0.00$).

3.2.2 Multivariate LASSO Models

We constructed a set of LASSO models with only out-of-sample predictors and without any accuracy-related measures. Such models mirror a setting in which forecasters have registered dozens of predictions, but no accuracy data are available yet. LASSO models tend to produce zero coefficients for some predictors, meaning that they do not improve fit enough to overcome the overfitting penalty. We show results for two model runs, A and B. In model run A, predictors are calculated for question subset 1, and then used to predict accuracy on subset 2. In model run B, the direction is reversed. This pattern is equivalent to two-fold cross validation and both sets of coefficients are shown to provide a look into the variability of model fits across questions sub-samples. All predictors are Z-score transformed (i.e., standardized) to enable more direct coefficient comparison. Table 6.4 reports the coefficients of the final model results. Predictors with non-zero coefficients are estimated to improve fit enough to offset the overfitting penalty, which does not necessarily mean that the coefficients would be statistically significant in a conventional ordinary-least squares model.

In Table 6.4, column A, we report the coefficients in the final model specification in order of decreasing absolute value: first-forecast proxy scores (b = 0.137), first-forecast extremity (b = −0.060), update frequency (b = −0.031), update magnitude (b = 0.004), fluid intelligence composite score (b = 0.004). All other coefficients were zero, as the predictors did not improve fit enough to overcome the LASSO penalty. In the converse run B, all coefficients were somewhat similar, except absolute update magnitude which was notably larger (b = 0.040).

Several notable patterns emerged. First, intersubjective proxy scores made the strongest out-of-sample predictor of accuracy in our set. Second, forecast extremity was the second strongest predictor, a result that would not be obvious from examining univariate correlations. Third, belief updating measures remained relevant. Coefficients for frequency were more consistent than those of magnitude. The most likely explanation for this pattern is that proxy scores were highly correlated with update magnitude ($r = 0.66$), but weakly correlated with update frequency ($r = -0.10$). See Appendix Table 6.6.

In Table 6.4, Columns C and D, we report the results models including out-of-sample accuracy (MSMDB) and excess volatility, both of which depend on resolution data, as well as out-of-sample and in-sample measures that do not rely on resolution data. These model specifications mirror a setting in which the tournament has been running for long enough to accuracy data on approximately one half of questions. More informally, these specifications follow an exploratory approach where we err on the side of over-inclusion of predictors, and rely on the regularization to reduce the risk of overfitting.

Table 6.4 LASSO regression models predicting SMDB accuracy measures. Non-zero coefficients do not imply statistical significance in an OLS model

	Out-of-sample, no accuracy data		In and out-of-sample, accuracy data	
Predictors	A	B	C	D
Intercept	−0.072	−0.082	−0.069	−0.080
Out-of-sample				
Accuracy, full MSMDB	NA	NA	0.056	0.081
Accuracy, first forecast Brier	NA	NA	0	0
Accuracy, last forecast Brier	NA	NA	0.059	0.029
IRT skill	NA	NA	0	−0.009
Excess volatility	NA	NA	0	0
Proxy score, first forecast	0.137	0.098	0	0
Update magnitude	0.004	0.040	0	0
Update frequency	−0.031	−0.033	0	−0.016
Confirmation propensity	0	0	0	0
Extremity, first forecast	−0.060	−0.042	0	0
Political knowledge score	0	0	0	0
Fluid IQ score	−0.004	−0.012	0	0
AOMT	0	0	0	0
Education, advanced degree	0	0	0	0
Expertise self-rating	0	0	0	0
In-sample				
Proxy score, first forecast	NA	NA	0.109	0.116
Update frequency	NA	NA	0	0
Update magnitude	NA	NA	0.017	0.014
Extremity, first forecast	NA	NA	−0.051	−0.042
All other predictors[a]	NA	NA	0	0

Note: [a]Only in-sample predictors with non-zero coefficients are shown. Other in-sample predictors are omitted due to space considerations

In the model run reported in column C, the only out-of-sample predictors with non-zero coefficients were overall MSMDB ($b = 0.056$) and last-forecast standardized Brier score (b = 0.059). Among in-sample predictors, the largest absolute coefficients were for first-forecast proxy scores ($b = 0.109$), first-forecast extremity ($b = -0.051$), followed by update magnitude ($b = 0.017$). The second sampling run, reported in column D, produced similar results, with one notable exception: out-of-sample IRT forecaster parameter had a non-zero regression coefficient ($b = -0.009$).

3.3 Study 2: Discussion

In summary, even when out-of-sample accuracy data on dozens of questions was available, in-sample intersubjective and behavioral measures still added value in

identifying skilled forecasters. More specifically, forecasters whose independent initial estimates were both relatively close to the consensus (yielding better proxy scores) and were relatively extreme, as well as those who updated in frequent, small steps, tended to be most accurate. At that point, none of the other predictors such as psychometric scores, other behavioral measures or self-reported confidence provided enough marginal value in improving fit to warrant inclusion into the model.

4 General Discussion

4.1 Research Synthesis

Our main objective was to summarize the existing evidence on measures for identifying skilled forecasters who tend to perform consistently better than their peers. Our review catalogued over 40 measures in a growing body of research from a wide range of academic fields, including psychology, judgment and decision making, decision science, political science, economics and computer science. The wide range of ideas, measures and naming conventions poses challenges to summarizing all in one place, but makes this summary more useful in enabling learning and synergy across disciplinary boundaries.

While not the result of a formal meta-analysis, the median absolute correlation coefficient among non-accuracy-related measures ($r = 0.20$) provides a rough but useful baseline for researchers conducting power analyses for studies about new skill-identification measures. More importantly, the current research helps us confirm or update views about the strongest correlates of prediction skill. Among the five categories, accuracy-related measures were, unsurprisingly, most highly correlated to the outcome measures, which were also based on accuracy. Put simply: predictive accuracy is reliable. Posing dozens of rigorously resolvable questions and scoring individuals on their accuracy on those questions remains the undisputed gold standard in skill spotting.

The results of Study 2 provide an upper limit of cross-sample reliability for accuracy measures of approximately $r = 0.74$ across random sub-samples of questions. As Atanasov et al. (2022b) noted, test-retest reliability across *seasons* tends to be lower, at approximately $r = 0.45$. In other words, skill assessments become less reliable with time (see Himmelstein et al., 2023a, this volume, for an in-depth discussion of temporally driven issues in judgmental forecasting). While relative accuracy appears to be consistent across questions and over time, the limits to reliability also relate to our expectations of the predictive fit of any measures, whether they are based on accuracy or not. It is difficult to predict the future values of any measure better than by using past values of the same measure.

The relatively low correlation IRT-model based skill estimates with of our accuracy measure highlights the importance of specific details in measurement definition, such as imputation, time-trends and transformations. In open tournaments, where forecasters generally answer a small proportion of available questions,

simpler measures, such as standardized Brier scores may be most practical. IRT model skill estimates may be most useful in settings where most forecasters answer the majority questions, avoiding sparse-matrix data issues. These models also show potential in adjusting for potential confounders, such as timing effects, and understanding the diagnostic properties of different types of forecasting questions.

In many real-world settings, gold-standard accuracy data are not available. Among the other categories, intersubjective measures demonstrated the strongest correlation withs accuracy. In Study 2, proxy scores based on the forecasters' initial estimates on questions provided stronger predictor fit that any other non-accuracy measure. Given this result, we see the study of intersubjective measures as an especially promising avenue for future research. Additional research may focus on improving intersubjective measures by maximizing the accuracy of consensus estimates that are used as proxies. For example, tournament designers must choose which forecasters are included in the consensus (e.g., superforecasters or less selective crowds), how consensus is updated over time, and how individual estimates are aggregated. It appears likely that more accurate consensus estimates will make for more effective proxies, but more research is needed to examine potential edge cases.

Promising applications of intersubjective measures include skill identification and incentive provision (Karger et al., 2022). At the same time, as Himmelstein et al. (2023b) point out, intersubjective measures that relate an individual's estimates to the consensus may be limited in their utility in spotting accurate forecasters with unique views. Intersubjective measures may be most helpful in identifying a small group of individuals whose aggregate estimates tend to be as accurate as those generated by a larger crowd. This is a useful property. To spot outstanding forecasters, intersubjective measures may need to be complemented by others.

Our analysis underscores the importance of behavioral measures. Building on Atanasov et al. (2020b), we showed that update frequency and magnitude add value in identifying accurate forecasters, even in the presence of accuracy-related and intersubjective measures. Frequent, small-increment updaters tend to generate accurate predictions across questions. Probabilistic extremity also appears useful in spotting accurate forecasters, especially as a complement to intersubjective measures. This finding may be specific to the construction of our proxy estimates, e.g., if we had applied stronger extremization in the aggregation algorithm producing the proxy estimates, forecaster extremity may have added less or no value. Our Study 2 analysis focused on independent forecasters who were not incentivized to write detailed rationales, so we did not analyze linguistic features of rationales. However, strong results across multiple previous studies (Horowitz et al., 2019; Zong et al., 2020; Karvetski et al., 2021) show that that such can be very helpful in spotting consistently accurate forecasters in settings where inter-forecaster communication is encouraged.

Among dispositional measures, performance-based scores related to forecasters' fluid intelligence were by far the most useful in assessing forecaster skills. As we showed, combinations of freely available measures can provide a useful starting point for spotting consistently accurate forecasters; fluid intelligence measures'

correlations with accuracy ranged up to $r = 0.3$. Thinking-styles measures, generally based on self-reports, registered relatively low correlations with forecasting skill. The measure with the highest correlation was actively-open minded thinking, and even for that, the range of correlations was between r = 0.10 and 0.15.

Other thinking-style measures yielded low and generally not statistically significant correlations. One notable example is the fox-hedgehog scale. Tetlock's (2005) seminal research on expert political judgment highlighted a version of this measure as a key correlate of accuracy among geopolitical experts in his multi-decade research study. The result that foxy forecasters tend to be better than their hedgehog-like peers is well known among researchers and forecasters. However, this finding did not replicate in the 2011–2015 ACE tournament. More specifically, Mellers et al. (2015a, p. 7) included fox-hedgehog scale, along with need for closure and actively open-minded thinking and concluded that: "Only one of the measures, actively open-minded thinking, was significantly related to standardized Brier score accuracy." Our current analysis, which included two additional years of GJP data, replicated this null relationship.

Popular science accounts of crowd prediction are still catching up to this evidence. Epstein (2019), for example, noted that Tetlock and Mellers' approach in GJP was to "identify a small group of the foxiest forecasters."[5] Foxiness was not actually used for the selection of superforecasters, nor in the weighting schemas for tournament-winning aggregation algorithms. While it is plausible that the measure is still useful in identifying relatively accurate subject matter experts, it is not predictive in an open forecasting tournament environment. This measure may also serve as an example of a broader concern about self-report measures: when a measure becomes well-known, it loses some of its predictive validity as survey respondents learn which responses will make them look good.[6]

One classic result from Tetlock (2005) that appears valid in our context is the notion that biographical measures of expertise are not effective at identifying consistently accurate forecasters. The literature review in Study 1 included several such measures education level and h-index, and most studies did not show strong correlations between biographical expertise measures and skill. In Study 2, we showed that forecaster self-reports about their own expertise were completely uncorrelated with accuracy.

Our results underscore a methodological challenge to researchers: seek ways to assess forecaster tendencies through their behaviors, and rely less on their self-reports. For example, if you seek confident forecasters, track the extremity of their

[5] We have notified Epstein of this. As a result, he shared plans to edit the sentence in future editions of Range.

[6] Readers who have been exposed to research on forecaster skill identification through general media or popular science outlets may find some of our findings surprising. For example, a recent admittedly non-scientific poll of 30 twitter users by one of us (Atanasov) revealed that the plurality (40%) of respondents thought active open mindedness was more strongly correlated with accuracy than update magnitude, fluid intelligence or subject matter knowledge scores. Fewer than 20% correctly guessed that the closest correlate of accuracy was update magnitude.

estimates and ignore their expertise self-ratings. If you seek open-minded forecasters, pay more attention to the frequency of their updates than to their responses on open-mindedness questionnaires. These two examples are consistent with our results. The challenge lies in creatively constructing behavioral measures suitable to new contexts.

4.2 Use Cases

We illustrate the real-world use of skill spotting measures with two vignettes, summarized in Table 6.4. Both involve forecasting tournaments consisting of hundreds of participants with dozens of questions. In the first vignette, the tournament takes place within a large corporation. All participants are employees of the firm. The questions focus on outcomes relevant to the firm, such as product launch dates, sales, popularity of product features (Cowgill & Zitzewitz, 2015), or clinical trial development milestones (Atanasov et al., 2022a). Questions resolve within weeks or months. The company runs the tournament to inform its strategy and operations but also to uncover analytical talent.

Given the short-term questions, accuracy-related measures become relevant quickly, and can thus add much value. At the start of the tournament, intersubjective and behavioral measures can be very helpful in assessing aggregation weights to individual forecasters. Most dispositional measures will likely be of little utility, as human-resource regulations may constrain the use of IQ-related tests, while self-report measures tend to have low predictive validity. Expertise information may be available from forecasters' biography and record at the company, but such information is not very useful in uncovering skilled forecasters. The most useful expertise measures will likely be knowledge tests with a calibration component, which tend to have moderate predictive validity.

The second vignette involves a public tournament focused on existential risk. The tournament is open for anyone to participate. Questions range in duration from several months to one hundred years. Due to the long-time horizon of most questions, accuracy-related measures do not provide sufficient skill signals early on. Intersubjective measures may prove especially useful here in terms of skill identification, as well as means of providing feedback and incentives to forecasters (Karger et al., 2022; Beard et al. 2020). Behavioral measures can also add value in the short to medium-run, mostly as inputs to aggregation weights. In open tournaments, the range of allowable dispositional measures expands, as fluid intelligence measures can be included, subject to IRB approval. Such measures may even be used as initial screening tools e.g., if there are thousands of interested forecasters but sufficient resources to administer or pay only a subset. Dispositional measures can also provide signals for aggregation algorithms, addressing the "cold start" problem. Over time, as data from intersubjective and behavioral measures accumulates, the relative value of dispositional measures will likely diminish. Expertise measures are

Table 6.5 Predicted value-added for each category of measures in two application scenarios

Skill identification measure category	Predicted value added	
	Corporate tournament: short-term questions, small teams	Open tournament: long-term questions, large crowds
1. Accuracy-related	Highest	Low
2. Intersubjective	High	Highest
3. Behavioral	High	High
4. Dispositional	Moderate	Moderate
5A. Expertise: Knowledge tests	Moderate	Moderate
5B. Expertise: Others	Low	Low

again of limited usefulness, except for knowledge tests with a calibration component (Table 6.5).

4.3 Limitations and Future Directions

We must acknowledge several limitations of the current research. First, while we aim to provide comprehensive summary of measures and empirical relationships, it is possible that we have missed some measures, especially ones older than 10 years, as well as new measures in unpublished studies. Relatedly, most measures included in our Study 1 review could not be practically included in our own analysis (Study 2), because of data availability, contextual differences or sensitivity to key assumptions. More comprehensive follow-up studies simultaneously testing multiple ideas will likely be beneficial, and consistent with the recent trend of "megastudies" in behavioral science (Milkman et al., 2022).

Second, the main statistical test utilized in most empirical analyses, the Pearson correlation coefficient, is designed to capture linear relationships. As such, we did not attempt to capture any non-linearities. For example, our research does not allow us to assess if a specific measure is particularly well suited for distinguishing among skill levels near the bottom or near the top of the distribution. Item-response theory (IRT) models are designed for this purpose. Such models are most useful in data-rich environments*, i.e.,* cases where most forecasters have answered most questions, and resolution information is available. Short of that, follow-up research should address non-linearities by zooming in on forecaster sub-sets or using more advanced statistical techniques, such as quantile regression models.

Third, most of the evidence summarized here is based on forecasting tournaments in which forecasters are asked and incentivized to produced maximally accurate forecasts, with the prospect of ground-truth verification. Different patterns may emerge in settings where forecasters produce unincentivized predictions (Dana et al., 2019), or are held accountable for process rather than accuracy-related outcomes (Chang et al., 2017). Finally, this research relies heavily on data from

forecasting tournaments focused on geopolitics and economics. The body of research focused on public health and life sciences outcomes is growing (Benjamin et al. 2017; Atanasov et al., 2020a, 2022a; McAndrew et al., 2022; Sell et al., 2021), but the evidence base on correlates of individual skill outside of geopolitics and economics remains relatively thin. Future research should examine if subject matter expertise is more or less closely related to forecasting skill in other domains.

4.4 Conclusion

Individual forecasting performance is largely a function of skill, as some forecasters perform consistently better than others. In the presence of plentiful historical accuracy information across dozens of questions, accuracy track records constitute the overwhelming gold-standard in talent spotting. In settings where such information is not available, however, we show that researchers have plenty of options for gauging predictive skill. Unfortunately, most measures that seem intuitively attractive at first sight are not very effective. Asking forecasters about their expertise, or about their thinking patterns is not useful in terms of predicting which individuals will prove consistently accurate. Examining their behaviors, such as belief updating patterns, as well as their psychometric scores related to fluid intelligence offer more promising avenues. Arguably the most impressive performance in our study was for registered intersubjective measures, which rely on comparisons between individual and consensus estimates. Such measures proved valid as predictors of relative accuracy. As our research focus moves away from large crowds of amateurs staring at oxen to smaller, more selective crowds, we need better maps to navigate through a peculiar terrain sown with broken expectations. This chapter aims to provide the most complete rendition of such a map.

Acknowledgments We thank Matthias Seifert, David Budescu, David Mandel, Stefan Herzog and Philip Tetlock for helpful suggestions. All remaining errors are our own. No project-specific funding was used for the completion of this chapter.

Appendix: Methodological Details of Selected Predictors

Item Response Theory Models

In forecasting, one such confounder is the timing in which forecasts are made. In forecasting tournaments, forecasters make many forecasts about the same problems at various time points. Those who forecast problems closer to their resolution date have an accuracy advantage which may be important to account for in assessing their talent level (for more detail, see Himmelstein et al., this issue). IRT models can be extended so that their diagnostic properties change relative to the time point at which

a forecaster makes their forecast. One such model is given below (Himmelstein et al., 2021; Merkle et al., 2016).

$$NB_{f,q,d} = b_{0,q} + \left(b_{1,q} - b_{0,q}\right)e^{-b_2 t_{f,q,d}} + \lambda_q \theta_f + \epsilon_{f,q,d}$$

The three b parameters represent how an item's difficult changes as time passes: $b_{0,\,q}$ represents an item's maximum difficulty (as time to resolution goes to infinity), $b_{1,\,q}$ an item's minimum difficulty (immediately prior to resolution), and b_2 the shape of the curve between $b_{0,\,q}$ and $b_{1,\,q}$ based on how much time is remaining in the question at the time of the forecast ($t_{f,\,q,\,d}$). The other two parameters represent how well an item discriminates between forecasters of different skill levels (λ_q) and how skilled the individual forecasters are (θ_f). As the estimate of forecaster skill, talent spotters will typically be most interested in this θ_f parameter, which is conventionally scaled so that it is on a standard normal distribution, $\theta_f \sim N(0, 1)$, with scores of 0 indicating an average forecaster, -1 a forecaster that is 1 SD below average, and 1 a forecaster that is 1 SD above average.

One potential problem with this model is that, in some cases, the distribution of Brier Scores is not well behaved. This typically occurs in cases which have many binary questions, so that the Brier score is a direct function of the accuracy assigned to the correct option. In such cases, the distribution of Brier scores can be multi-modal, because forecasters will tend to input many extreme and round number probability estimates, such as 0, .5, and 1 (Bo et al., 2017; Budescu et al., 1988; Merkle et al., 2016; Wallsten et al., 1993). To accommodate such multi-modal distributions, one option is to discretize the distribution of Brier scores into bins and reconfigure the model into an ordinal response model. Such models, such as the graded response model (Samejima, 1969), have a long history in the IRT literature.

Merkle et al. (2016) and Bo et al. (2017) describe examples of ordinal IRT models for forecasting judgment. However, the former found that the continuous and ordinal versions of the model were highly correlated (r = .87) in their assessment of forecaster ability level, and that disagreements tended to be focused on poor performing forecasters (who tend to make large errors) than high performing forecasters.

Contribution Scores

To obtain contribution scores for individual forecasters, it is necessary to first define some aggregation method for all of their judgments for each question. The simplest, and most common form of aggregation would just be to obtain the mean of all probabilities for all events associated with a forecasting problem. The aggregate probability (AP) for each of the c events associated with a forecasting question across all forecasters would be

$$AP_{q,c} = \frac{\sum_{f=1}^{F} p_{q,c,f}}{F}$$

And the aggregate Brier score (*AB*) would then be

$$AB_q = \sum_{c=1}^{C} \left(AP_{q,c} - y_q\right)^2$$

Based on this aggregation approach, defining the contribution of individual forecasters to the aggregate is algebraically straightforward. We can define $AP_{q,\,c,\,-f}$ as the aggregate probability with an individual forecaster's judgment removed as

$$AP_{q,c,-f} = \frac{(F)\left(AP_{q,c}\right) - p_{q,c,f}}{F-1}$$

And the aggregate Brier score with an individual forecaster's judgment removed as

$$AB_{q,-f} = \sum_{c=1}^{C} \left(AP_{q,c,-f} - y_q\right)^{2.}$$

Finally, we define a forecaster's average *contribution* to the accuracy of the aggregate crowd forecasts as

$$C_f = \frac{\sum_{q=1}^{Q} AB_q - AB_{q,-f}}{Q}$$

C_f is a representation of how much information a forecaster brings to the table, on average, that is both *unique* and *beneficial*. It is possible that a forecaster ranked very highly on individual accuracy might be ranked lower in terms of their contribution, because their forecasts tended to be very similar to the forecasts of others, and so they did less to move the needle when averaged into the crowd.

Both weighting members of the crowd by average contribution scores, as well as selecting positive or high performing contributors, have been demonstrated to improve the aggregate crowd judgment (Budescu & Chen, 2015; Chen et al., 2016). The approach is especially appealing because it can be extended into a model that is dynamic, in that it is able to update contribution scores for each member of a crowd as more information about their performance comes available; it requires relatively little information about past performance to reliably estimate high performing contributors; and it is cost effective, in that is able to select a relatively small group of high performing contributors who can produce an

Table 6.6 Correlation matrix for measures in Study 2. Pearson correlation coefficients reported. Below-diagonal values are assessed in-sample, above diagonal values are calculated out-of-sample. Diagonal values highlighted in gray are cross-sample reliability coefficients. The bottom five measures are not question specific, so out-of-sample correlation coefficients or cross-sample reliability coefficients are not relevant

Measure	Standard. Brier (SB)	Debiased Brier	SB, 1st Forecast	SB, Last Forecast	IRT Model	Excess Volatility	Proper Proxy, All Forecasts	Proper Proxy, 1st Forecast	N. of Questions	Update Magnitude, Abs. Dist.	Update Freq.	Confirm. Prop.	Extremity 1st Forecast	Mean Expert.
Standardized Brier (SB)	0.74	0.69	0.62	0.68	-0.22	0.30	0.58	0.49	0.25	-0.32	0.45	0.03	-0.15	0.00
Debiased Brier	0.96	0.69	0.58	0.64	-0.15	0.32	0.59	0.50	0.22	-0.29	0.45	0.03	-0.07	0.02
SB, 1st Forecast	0.85	0.82	0.68	0.49	-0.15	0.39	0.56	0.53	0.21	-0.09	0.44	0.09	-0.22	0.02
SB, Last Forecast	0.84	0.80	0.64	0.83	-0.29	0.18	0.51	0.38	0.22	-0.44	0.35	0.01	-0.17	-0.06
IRT Model	-0.30	-0.23	-0.24	-0.34	0.89	0.42	0.15	0.21	0.05	0.16	0.23	0.27	0.60	0.17
Excess Volatility	0.40	0.43	0.48	0.26	0.40	0.85	0.63	0.67	0.12	0.04	0.46	0.35	0.38	0.15
Proper Proxy, All Forecasts	0.69	0.69	0.68	0.60	0.11	0.55	0.81	0.75	0.22	-0.18	0.58	0.23	0.28	0.13
Proper Proxy, 1st Forecast	0.60	0.60	0.67	0.44	0.15	0.61	0.93	0.77	0.18	-0.10	0.58	0.25	0.31	0.15
Number of Questions	0.25	0.23	0.22	0.24	0.05	0.11	0.21	0.18	0.87	-0.15	0.29	-0.11	0.00	-0.10
Update Magnitude, Abs. Dist.	-0.31	-0.28	-0.07	-0.44	0.16	0.05	-0.17	-0.10	-0.16	0.96	-0.29	0.25	-0.03	0.05
Update Frequency	0.51	0.52	0.51	0.35	0.21	0.51	0.63	0.63	0.31	-0.29	0.75	0.00	0.22	0.02
Confirmation Propensity	0.03	0.02	0.09	0.02	0.27	0.34	0.23	0.25	-0.07	0.24	0.00	0.86	0.25	0.32
Extremity, 1st Forecast	-0.19	-0.11	-0.28	-0.20	0.63	0.38	0.27	0.30	0.01	-0.05	0.21	0.26	0.91	0.24
Mean Expertise	0.00	0.01	0.01	-0.06	0.18	0.15	0.13	0.14	-0.11	0.05	0.02	0.32	0.23	0.97
Non-Question Specific Measures														
Fluid IQ Composite All	-0.27	-0.24	-0.29	-0.22	0.03	-0.25	-0.23	-0.23	-0.05	-0.09	0.04	-0.33	0.07	-0.15
Political Knowledge Score	-0.10	-0.06	-0.01	-0.17	0.13	0.06	-0.07	-0.05	-0.08	-0.04	0.14	0.00	0.01	0.03
AOMT	-0.10	-0.08	-0.05	-0.13	0.08	0.08	-0.04	-0.01	-0.14	-0.01	0.08	-0.01	0.08	-0.03
Fox-Hedgehog Scale	0.08	0.07	0.05	0.09	-0.05	0.00	0.03	0.02	0.07	0.07	0.00	0.00	-0.01	-0.04
Conscientiousness	0.13	0.12	0.14	0.12	0.03	0.12	0.15	0.15	0.16	0.20	-0.10	0.02	-0.05	0.01

aggregate judgment that matches or exceeds the judgment of larger crowds in terms of accuracy (Chen et al., 2016).

The advent of contribution assessment was initially designed with a particular goal in mind: to improve the aggregate wisdom of the crowd (Budescu & Chen, 2015; Chen et al., 2016). One might challenge as slightly as a slightly narrower goal than pure talent spotting. It is clearly an effective tool for maximizing crowd wisdom, but is it a valid tool for assessing expertise? The answer appears to be yes. Chen et al. (2016) not only studied contribution scores as an aggregation tool but tested how well contribution scores perform at selecting forecasters known to have a skill advantage through various manipulations known to benefit expertise, such as explicit training and interactive collaboration.

References

Arthur, W., Jr., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the raven advanced progressive matrices test. *Journal of Psychoeducational Assessment, 17*(4), 354–361.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463*(7279), 294–295.

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

Atanasov, P., Diamantaras, A., MacPherson, A., Vinarov, E., Benjamin, D. M., Shrier, I., Paul, F., Dirnagl, U., & Kimmelman, J. (2020a). Wisdom of the expert crowd prediction of response for 3 neurology randomized trials. *Neurology, 95*(5), e488–e498.

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B., & Tetlock, P. (2020b). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes, 160*, 19–35.

Atanasov, P., Joseph, R., Feijoo, F., Marshall, M., & Siddiqui, S. (2022a). *Human forest vs. random forest in time-sensitive Covid-19 clinical trial prediction*. Working Paper.

Atanasov, P., Witkowski, J., Mellers, B., & Tetlock, P. (2022b) *Crowdsourced prediction systems: Markets, polls, and elite forecasters*. Working Paper.

Augenblick, N., & Rabin, M. (2021). Belief movement, uncertainty reduction, and rational updating. *The Quarterly Journal of Economics, 136*(2), 933–985.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26.

Beard, S., Rowe, T., & Fox, J. (2020). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures, 115*, 102469.

Benjamin, D., Mandel, D. R., & Kimmelman, J. (2017). Can cancer researchers accurately judge whether preclinical reports will reproduce? *PLoS Biology, 15*(6), e2002212.

Bennett, S., & Steyvers, M. (2022). Leveraging metacognitive ability to improve crowd accuracy via impossible questions. *Decision, 9*(1), 60–73.

Bland, J. M., & Altman, D. G. (2011). Correlation in restricted ranges of data. *BMJ: British Medical Journal, 342*.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science, 36*(8), 887–1009.

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment & Decision Making, 12*(2), 90–103.

Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*(3), 382–398.

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

Broomell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika, 74*(3), 531–553.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*(5), 938–956.

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281–294.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*(2), 267–280.

Budescu, D.V., Himmelstein, M & Ho, E. (2021, October) Boosting the wisdom of crowds with social forecasts and coherence measures. In *Presented at annual meeting of Society of Multivariate Experimental Psychology (SMEP)*.

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., Fider, F., Rumpff, L., & Twardy, C. (2011). Expert status and performance. *PLoS One, 6*(7), e22998.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131.

Chang, W., Atanasov, P., Patil, S., Mellers, B., & Tetlock, P. (2017). Accountability and adaptive performance: The long-term view. *Judgment and Decision making, 12*(6), 610–626.

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis, 13*(2), 128–152.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision making, 7*(1), 25–47.

Collins, R. N., Mandel, D. R., Karvetski, C. W., Wu, C. M., & Nelson, J. D. (2021). The wisdom of the coherent: Improving correspondence with coherence-weighted aggregation. *Preprint available at PsyArXiv*. Retrieved from https://psyarxiv.com/fmnty/

Collins, R., Mandel, D., & Budescu, D. (2022). Performance-weighted aggregation: Ferreting out wisdom within the crowd. In M. Seifert (Ed.), *Judgment in predictive analytics*. Springer [Reference to be updated with page numbers].

Cooke, R. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Costa, P. T., Jr., & McCrae, R. R. (2008). *The revised neo personality inventory (NEO-PI-R)*. Sage.

Cowgill, B., & Zitzewitz, E. (2015). Corporate prediction markets: Evidence from Google, Ford, and Firm X. *The Review of Economic Studies, 82*(4), 1309–1341.

Dana, J., Atanasov, P., Tetlock, P., & Mellers, B. (2019). Are markets more accurate than polls? The surprising informational value of "just asking". *Judgment and Decision making, 14*(2), 135–147.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision, 1*(2), 79–101.

Dieckmann, N. F., Gregory, R., Peters, E., & Hartman, R. (2017). Seeing what you want to see: How imprecise uncertainty ranges enhance motivated reasoning. *Risk Analysis, 37*(3), 471–486.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Epstein, D. (2019). *Range: How generalists triumph in a specialized world*. Pan Macmillan.

Fan, Y., Budescu, D. V., Mandel, D., & Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis, 16*, 197–217.

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

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature, 75*(7), 450–451.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378.

Goldstein, D. G., McAfee, R. P., & Suri, S. (2014, June). The wisdom of smaller, smarter crowds. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (pp. 471–488).

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological), 1952*, 107–114.

Hanea, A. D., Wilkinson, D., McBride, M., Lyon, A., van Ravenzwaaij, D., Singleton Thorn, F., Gray, C., Mandel, D. R., Willcox, A., Gould, E., Smith, E., Mody, F., Bush, M., Fidler, F., Fraser, H., & Wintle, B. (2021). Mathematically aggregating experts' predictions of possible futures. *PLoS One, 16*(9), e0256919. https://doi.org/10.1371/journal.pone.0256919

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision making, 8*(3), 188–201.

Hastie, T., Qian, J., & Tay, K. (2021). *An introduction to glmnet*. CRAN R Repository.

Himmelstein, M., Atanasov, P., & Budescu, D. V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment & Decision Making, 16*(2), 323–362.

Himmelstein, M., Budescu, D. V., & Han, Y. (2023a). The wisdom of timely crowds. In M. Seifert (Ed.), *Judgment in predictive analytics*. Springer.

Himmelstein, M., Budescu, D. V., & Ho, E. (2023b). The wisdom of many in few: Finding individuals who are as wise as the crowd. *Journal of Experimental Psychology: General.* Advance online publication.

Ho, E. H. (2020, June). *Developing and validating a method of coherence-based judgment aggregation*. Unpublished PhD Dissertation. Fordham University, Bronx NY.

Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Resnick Samotin, L., Roberts, M., Chang, W., Mellers, B., & Tetlock, P. (2019). What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *The Journal of Politics, 81*(4), 1388–1404.

Joseph, R., & Atanasov, P. (2019). *Predictive training and accuracy: Self-selection and causal factors*. Working Paper, Presented at Collective Intelligence 2019.

Karger, E., Monrad, J., Mellers, B., & Tetlock, P. (2021). *Reciprocal scoring: A method for forecasting unanswerable questions*. Retrieved from SSRN

Karger, J., Atanasov, P., & Tetlock, P. (2022). *Improving judgments of existential risk: Better forecasts, questions, explanations, policies*. SSRN Working Paper.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis, 10*(4), 305–326.

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2021). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting, 38*(2), 688–704.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., Zalaudek, I., Carney, P. A., & Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances, 5*(11), eaaw9011.

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

Liu, Y., Wang, J., & Chen, Y. (2020, July). Surrogate scoring rules. In *Proceedings of the 21st ACM Conference on Economics and Computation* (pp. 853–871).

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

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods, 42*(2), 525–541.

Mauksch, S., Heiko, A., & Gordon, T. J. (2020). Who is an expert for foresight? A review of identification methods. *Technological Forecasting and Social Change, 154*, 119982.

McAndrew, T., Cambeiro, J., & Besiroglu, T. (2022). Aggregating human judgment probabilistic predictions of the safety, efficacy, and timing of a COVID-19 vaccine. *Vaccine, 40*(15), 2331–2341.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & 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. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1.

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

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multitask, multi-benchmark study. *Judgment and Decision making, 12*(4), 369–381.

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision, 3*(1), 1–19.

Milkman, K. L., Gandhi, L., Patel, M. S., Graci, H. N., Gromet, D. M., Ho, H., Kay, J. S., Lee, T. W., Rothschild, J., Bogard, J. E., Brody, I., Chabris, C. F., & Chang, E. (2022). A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences, 119*(6), e2115126119.

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science, 51*(9), 1359–1373.

Morstatter, F., Galstyan, A., Satyukov, G., Benjamin, D., Abeliuk, A., Mirtaheri, M., et al. (2019). SAGE: A hybrid geopolitical event forecasting system. *IJCAI, 1*, 6557–6559.

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

Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science, 65*(5), 2291–2309.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407–413.

Predd, J. B., Osherson, D. N., Kulkarni, S. R., & Poor, H. V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis, 5*(4), 177–189.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science, 306*(5695), 462–466.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 34*, 1–97.

Seifert, M., Siemsen, E., Hadida, A. L., & Eisingerich, A. B. (2015). Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management, 36*, 33–45.

Sell, T. K., Warmbrod, K. L., Watson, C., Trotochaud, M., Martin, E., Ravi, S. J., Balick, M., & Servan-Schreiber, E. (2021). Using prediction polling to harness collective intelligence for disease forecasting. *BMC Public Health, 21*(1), 1–9.

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2 manual*. Western Psychological Services.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*(2), 342–357.

Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes, 69*(3), 205–219.

Suedfeld, P., & Tetlock, P. (1977). Integrative complexity of communications in international crises. *Journal of Conflict Resolution, 21*(1), 169–184.

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2017). Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Science, 63*(2), 497–518.

Tetlock, P. E. (2005). *Expert political judgment*. Princeton University Press.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning, 20*(2), 147–168.

Tsai, J., & Kirlik, A. (2012). Coherence and correspondence competence: Implications for elicitation and aggregation of probabilistic forecasts of world events. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, pp. 313–317). Sage.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*(2), 176–190.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6), 1049–1162.

Witkowski, J., & Parkes, D. (2012). A robust bayesian truth serum for small populations. *Proceedings of the AAAI Conference on Artificial Intelligence, 26*(1), 1492–1498.

Witkowski, J., Atanasov, P., Ungar, L., & Krause, A. (2017) Proper proxy scoring rules. In *Presented at AAAI-17: Thirty-First AAAI Conference on Artificial Intelligence.*

Zong, S., Ritter, A., & Hovy, E. (2020). Measuring forecasting skill from text. *arXiv preprint arXiv:2006.07425*.