TO WHAT EXTENT IS FORESIGHT LUCK? IMPLICATIONS FOR COGNITIVE, EVOLUTIONARY, AND CHANCE THEORIES OF FORESIGHT

Benedikt Alexander Schuler
University of St.Gallen
Dufourstrasse 40a
9000 St. Gallen
Switzerland
benedikt.schuler@unisg.ch

ABSTRACT

This study assesses the relative plausibility of cognitive, evolutionary, and chance theories of foresight by quantifying the extent to which foresight is luck. Results from a regression-toward-the-mean approach applied to foresight estimates derived from a beta item response theory model using data from 265 forecasters who made 40,514 forecasts on 212 questions over two years in a forecasting tournament suggest that foresight is on average 32% luck. This means that the foresight of an individual forecaster regresses on average by 32% toward the mean foresight of all forecasters over one year. This regression toward the mean is stronger among top forecasters (35%) and forecasters with below average foresight (47%) but weaker for the average forecaster (23%). Despite this regression toward the mean, the top forecasters in year 1 did not display worse foresight than other forecasters in year 2. I discuss the implications of these findings for cognitive, evolutionary, and chance theories of foresight.

Keywords: foresight, cognition, evolution, luck, item response theory, regression toward the mean

INTRODUCTION

Theories in strategy differ in the extent to which they consider foresight—"the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time" (Schuler et al., 2024, p. 1)—to be luck. Cognitive theories suggest that strategists possess "boundedly rational foresight" (Gavetti et al., 2012, p. 9) that is grounded in their mental representations of the world rendering foresight mainly a skill (Csaszar, 2018; Csaszar & Laureiro-Martínez, 2018; Gavetti & Levinthal, 2000). Evolutionary theories (Levinthal, 2021; Murmann, 2003; Nelson & Winter, 1982) acknowledge that strategists may exhibit myopic foresight grounded in their cognitive preadaptation that enables them to discover opportunities serendipitously (Denrell et al., 2003; Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). Chance theories question the existence of managerial foresight by emphasizing the role of random processes rendering accurate foresight mostly luck (Denrell et al., 2015; Denrell & Fang, 2010; Liu, 2020; Liu & de Rond, 2016).

However, the relative plausibility of cognitive, evolutionary, and chance theories of foresight has received scant empirical attention. One explanation for this may be that unlike research on the relative importance of organizational and industry effects (McGahan & Porter, 1997; Rumelt, 1991; Wang, 2023), it is empirically challenging to directly examine the relative explanatory power of these theories in a single study. Such a study would require the measurement of the content and structure of mental representations of strategists (Csaszar & Laureiro-Martínez, 2018; Csaszar & Ostler, 2020), the cognitive preadaptation of strategists in relation to their environment and changes thereof over time (Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016), and an exhaustive set of variables controlling for alternative accounts of foresight as well as the comparison of alternative random processes to evaluate the role of chance in foresight

(Denrell et al., 2015). Despite these empirical challenges, it is important to assess the relative plausibility of cognitive, evolutionary, and chance theories of foresight because otherwise we may not fully understand to what extent foresight—and by extension the strategist—may be a source of competitive advantage resulting in superior organizational performance.

Therefore, the purpose of this paper is to assess the relative plausibility of cognitive, evolutionary, and chance theories of foresight by quantifying the extent to which foresight is luck. Although cognitive, evolutionary, and chance theories of foresight propose different explanations for foresight rendering a joint examination of these explanations unfeasible, these theories may be compared based on their implications for the extent of luck in foresight. Cognitive theories suggest that foresight is mainly a skill and only somewhat luck. Evolutionary theories consider foresight to be somewhat skill and mainly luck. Chance theories consider foresight to be mostly luck. Quantifying the extent to which foresight is luck may, therefore, provide important insights into the relative explanatory power of different theories of foresight. To this end, I apply a regression toward the mean approach (Liu, 2020; Samuels, 1991; Schmittlein, 1989) to foresight estimates derived from a beta item response theory model using data from 265 forecasters who made 40,514 forecasts on 212 questions over two years in a forecasting tournament. The results suggest that foresight is on average 32% luck. This means that the foresight of an individual forecaster regresses on average by 32% toward the mean foresight of all forecasters over one year. This regression toward the mean is stronger among top forecasters (35%) and forecasters with below average foresight (47%) but weaker for the average forecasters (23%). Despite this regression toward the mean, the top forecasters in year 1 did not display worse foresight than other forecasters in year 2.

This study contributes to the strategy literature by assessing the relative plausibility of cognitive, evolutionary, and chance theories of foresight. Although foresight may be somewhat

luck, it appears to be mainly skill. This is consistent with cognitive theories proposing that managers may exhibit boundedly rational foresight grounded in their mental representations of the world (Csaszar, 2018; Csaszar & Laureiro-Martínez, 2018; Gavetti & Levinthal, 2000). Furthermore, the fact that forecasters were able to accurately predict future states of the world three to six months and in some cases up to one year in advance on a wide range of questions suggests that this boundedly rational foresight was less myopic than assumed by evolutionary theories, albeit skillful foresight may still be partially driven by cognitive preadaptation (Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). At the same time, this result cautions against ascribing omniscience to strategists who displayed foresight historically. Although such strategists may continue to exhibit foresight, their foresight was partially rooted in serendipity (Denrell et al., 2003; Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016) and luck (Denrell et al., 2015; Denrell & Fang, 2010; Liu, 2020; Liu & de Rond, 2016). Taken together, this study suggests that the strategist's foresight may be a source of competitive advantage with potentially positive effects on organizational performance.

The remainder of this paper is structured as follows: First, I will briefly outline cognitive, evolutionary, and chance theories of foresight. Second, I will explain the regression-toward-the-mean approach, introduce a beta item response theory model to measure foresight, and describe the data and analytical approach of this study. Third, I will report the results of this paper. Last, I will discuss the theoretical and practical implications of this study, acknowledge its limitations, and articulate opportunities for future research.

THEORETICAL BACKGROUND

In the following, I will briefly outline cognitive, evolutionary, and chance theories of foresight. For each theory, I will first describe the nature of the strategist's foresight and why strategists may

or may not display foresight. Second, I will derive to what extent these theories consider foresight to be luck. Last, I will explain how the relative plausibility of cognitive, evolutionary, and chance theories of foresight can be assessed based on their different implications for the extent to which foresight is luck. Table 1 summarizes cognitive, evolutionary, and chance theories of foresight.

Insert Table 1 about here

Cognitive Theories of Foresight

Cognitive theories of foresight suggest that strategists possess "boundedly rational foresight" (Gavetti et al., 2012, p. 9) that is grounded in their mental representations of the world (Csaszar, 2018; Csaszar & Laureiro-Martínez, 2018; Gavetti & Levinthal, 2000). A mental representation is a "model of reality held in the mind of an individual, who can use this representation to generate predictions about reality" (Csaszar & Levinthal, 2016, p. 2031f.). Based on their predictions, strategists choose strategic decision alternatives they expect to satisfice organizational goals (March & Simon, 1958; Simon, 1947). In other words, mental representations guide strategists in their search of strategic initiatives that improve performance (Gavetti et al., 2005; Gavetti & Levinthal, 2000) by enabling them to predict future states of the world. Therefore, accurate mental representations underlying the strategist's foresight have been theorized to positively affect organizational performance (Csaszar, 2018; Gavetti, 2012).

From a cognitive perspective, foresight is considered to be mainly a skill as mental representations can be learned and adapted. Strategists may learn and use complex representations if they do not yet understand the drivers of organizational performance in an environment, whereas they may learn and use simple representations if they do (Csaszar & Ostler, 2020). Given the importance of foresight in strategic decision making, strategists should actively search for mental representations that enable them to predict future states of the world (Csaszar & Levinthal, 2016).

Strategists may learn accurate mental representations in strategy courses (Heshmati & Csaszar, 2024) that teach frameworks like Porter's Five Forces (Porter, 1979, 2008). By internalizing and adapting such external representations (Csaszar et al., 2024) for their specific organization, strategists may make better strategic decisions. That being said, the strategist's cognitive, informational, and time limitations (March & Simon, 1958; Simon, 1947) as well as the inherent randomness of the world (Packard & Clark, 2020; Schuler et al., 2024) mean that uncertainty about future cannot be fully resolved rendering foresight somewhat luck. Taken together, cognitive theories of foresight suggest that strategists possess boundedly rational foresight grounded in their mental representations of the world rendering foresight mainly a skill but also somewhat luck.

Evolutionary Theories of Foresight

Evolutionary theories (Levinthal, 2021; Murmann, 2003; Nelson & Winter, 1982) suggest that strategists may exhibit myopic foresight because they are cognitively preadapted to a future environment (Denrell et al., 2003; Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). Strategists are cognitively preadapted to a future environment if they possess deep domain knowledge that enables them to see unique opportunities that other strategists cannot see once the future environment realizes (Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). Thus, foresight represents a form of myopic, serendipitous exploitation (Busch, 2020, 2024; Denrell et al., 2003) of deep domain knowledge in a changed environment rather than the active anticipation of distant future states of the world (Levinthal, 2017, 2021). While evolutionary theories acknowledge that foresight may contribute to organizational performance serendipitously, organizational routines that are well adapted to the environment are considered to be the main driver of organizational performance (Levinthal, 2021; Murmann, 2003; Nelson & Winter, 1982).

From an evolutionary perspective, foresight is somewhat skill and largely luck. Foresight

is somewhat skill because strategists may deepen and broaden their domain knowledge to become cognitively preadapted to an increasing number of potential future environments (Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). Furthermore, strategists still need to attend to changes in the environment and exert effort to seize the opportunities in the environment that they see based on their cognitive preadaptation (Busch, 2020, 2024; Denrell et al., 2003). At the same time, foresight is mainly luck because it remains a form of myopic, serendipitous exploitation of changing environments that are mostly not under the control of the strategist but determine what deep domain knowledge is valuable (Levinthal, 2017, 2021). Overall, evolutionary theories of foresight suggest that strategists may exhibit myopic foresight because they are cognitively preadapted to a future environment rendering foresight somewhat skill but largely luck.

Chance Theories of Foresight

Chance theories suggest that true foresight may not exist because accurate predictions of future states of the world by strategists are due to chance (Denrell et al., 2015; Liu, 2020; Liu & de Rond, 2016). Strategists who accurately predict future states of the world may not exhibit foresight but were simply lucky (Denrell & Liu, 2012). Therefore, accurate predictions of future states of the world—especially of extreme events—may not be viewed as evidence of the strategist's foresight but rather as a form of bad judgment because such predictions systematically neglect the role of chance in the world (Denrell & Fang, 2010). Consequently, chance theories suggest that foresight only plays a minor role in explaining organizational performance. In fact, superior organizational performance itself—especially over long periods of time—is considered to be the result of random processes rather than the strategist's actions (Denrell, 2004; Denrell et al., 2013).

From a chance perspective, foresight is considered to be mainly luck as accurate

predictions of future states of the world are due to chance. Therefore, foresight should regress toward the mean in one of two systematic manners over time (Denrell et al., 2019; Denrell & Liu, 2012; Liu, 2020): First, strategists who accurately (inaccurately) predicted future states of the world initially may predict future states of the world less (more) accurately later (weak pattern of luck). Second, strategists who predicted future states of the world most accurately (inaccurately) initially may predict future states of the world less (more) accurately than moderately accurate (inaccurate) strategists later (strong pattern of luck). In other words, strategists who made extremely accurate (inaccurate) predictions regress more strongly toward the mean than strategists who made moderately accurate (inaccurate) predictions. Irrespective of whether foresight follows the weak or strong pattern of luck, chance theories imply that foresight fully regresses to the mean sooner or later and thus, is mostly luck.

Assessing the Relative Plausibility of Foresight Theories: To What Extent Is Foresight Luck?

Cognitive, evolutionary, and chance theories propose plausible explanations for foresight. However, the relative plausibility of cognitive, evolutionary, and chance theories of foresight has received scant empirical attention. While it is empirically challenging to examine the relative explanatory power of these theories in a single study directly, it may be assessed based on their differing implications for the extent to which foresight is luck. Therefore, one potential approach to better understand the relative plausibility of cognitive, evolutionary, and chance theories of foresight is to quantify the extent to which foresight is luck. In the remainder of this paper, I will apply a regression toward the mean approach to foresight estimates derived from a beta IRT model using data from 265 forecasters who made 40,514 forecasts on 212 questions over two years in a forecasting tournament to quantify the extent to which foresight is luck.

METHODS

In the following, I will first explain the regression-toward-the-mean approach to quantify the extent to which foresight is luck. Second, I will introduce a beta item response theory (IRT) model to measure foresight. Third, I will describe the data from the forecasting tournament. Finally, I will explain how I analyzed the data from the forecasting tournament using the beta IRT model and how I implemented the regression-toward-the-mean approach to quantify luck in foresight.

A Regression-Toward-the-Mean Approach to Quantify the Extent of Luck in Foresight

I used a regression-toward-the-mean approach to quantify the extent to which foresight is luck (Liu, 2020; Samuels, 1991; Schmittlein, 1989). To illustrate the intuition behind the regression-toward-the-mean approach, suppose we ask 100 strategists to forecast 10 questions about different future states of the world. After the true future states of the world realized, we assign each strategist one score representing their foresight at time 0. Then, we ask the same 100 strategists again to forecast 10 different questions about future states of the world. After the true future states of the world realized, we assign each strategist a second score representing their foresight at time 1.

If foresight is 100% skill, strategists receive the same foresight score at times 0 and 1 (e.g., forecaster A may receive a score of -1 at time 0 and 1, whereas forecaster B may receive a score of 1 at time 0 and 1). Assuming the scores at times 0 and 1 to be identically distributed (i.e., their shapes, means, and variances are identical) and standardized (i.e., mean = 0 and variance = 1), regressing the foresight score at time 1 on the foresight score at time 0 would yield an intercept of 0 and a regression coefficient of 1. Graphically, this corresponds to a 45 degree line intersecting the origin (i.e., $y = x$). In contrast, if foresight is 100% luck, strategists would receive different scores at times 0 and 1, and the scores would be unrelated to each other. In other words, the foresight score at time 0 would contain no information about the foresight score at time 1.

Assuming the scores at times 0 and 1 to be identically distributed and standardized, regressing the foresight score at time 1 on the foresight score at time 0 would yield an intercept of 0 and a regression coefficient of 0. Graphically, this corresponds to the x-axis. If both luck and skill play a role in foresight, strategists would receive different scores at times 0 and 1, but the scores would be related to each other. That is, the first score would contain information about the second score. Assuming the scores at times 0 and 1 to be identically distributed and standardized, regressing the foresight score at time 1 on the foresight score at time 0 would yield an intercept of 0 and a regression coefficient between 0 and 1. Graphically, this corresponds to a line between 45 and 0 degrees intersecting the origin (cf. Figure 1). Note that the regression coefficient corresponds to the correlation between the standardized foresight scores at times 0 and 1 in all these cases.

The regression coefficients (or the correlation between the foresight scores) in these examples may be interpreted as a measure of the percentage of skill involved in foresight. Conversely, one minus the regression coefficient may be interpreted as a measure of the percentage of luck involved in foresight. In other words, the regression toward the mean (Samuels, 1991) can be estimated by regressing the foresight score at time 1 on the foresight score at time 0. This assumes that the relationship between the foresight scores is linear.

However, if the foresight scores exhibit a non-linear (e.g., a quadratic or cubic) relationship, the role of luck in foresight differs depending on the foresight of forecasters at time 0. In these cases, it is worthwhile to explore the regression toward the mean in different subgroups of foresight at time 0 to assess the extent to which the regression toward the mean (i.e., the role of luck) is stronger or weaker in these subgroups (Liu, 2020). This can be done based on the reversion toward the mean, which generalizes the notion of the regression toward the mean (see Samuels, 1991 for a formal definition): Suppose we split the distribution of foresight scores at time 0 into an upper

and lower part based on a cutoff. Assuming that the cutoff lies above the mean foresight of the forecasters at time 0 (i.e., above 0 if the foresight scores are standardized), reversion toward the mean occurs if the mean foresight scores of forecasters in the upper part of the distribution at time 0 revert back toward the mean foresight of the forecasters at time 1 (i.e., toward 0 if the foresight scores are standardized). Analogously, assuming that the cutoff lies below the mean foresight of the forecasters at time 0 (i.e., below 0 if the foresight scores are standardized), reversion toward the mean occurs if the mean foresight scores of forecasters in the lower part of the distribution at time 0 revert back toward the mean foresight of the forecasters at time 1 (i.e., toward 0 if the foresight scores are standardized). This reversion toward the mean can be quantified by dividing the difference between the mean of the foresight scores of forecasters above (below) the cutoff at time 0 and the mean of the foresight scores of these forecasters at time 1 by the mean of the foresight scores of these forecasters at time 0 (cf. Figure 3). This yields the extent to which foresight is luck in percent (see data analysis section for a more formal description).

This regression-toward-the-mean-approach makes four assumptions: First, the foresight of the individual strategist is stable over time. In other words, the foresight of a strategists should not increase or decrease over time as such changes would be interpreted as evidence for luck. This assumption needs to be theoretically justified and is difficult to evaluate by statistical tests or plots. Second, the foresight scores must be identically distributed (Samuels, 1991). Third, the foresight scores must exhibit weak positive dependence—the means of the foresight scores must regress toward the mean of the foresight scores of all forecasters but not beyond (Samuels, 1991). This can be graphically evaluated, for example, by plotting the regression fitted to the data. Fourth, the measurement of foresight should be free from measurement error. Measurement errors attenuate the relationship between the foresight estimates (Fuller & Hidiroglou, 1978), which inflates the

estimate to which foresight is luck. Therefore, it is important and worthwhile to optimize the measurement accuracy of foresight to obtain the most accurate quantification possible for the extent to which foresight is luck. Item response theory models represent one way how foresight can be accurately measured. In the following, I will introduce a beta IRT model of probabilistic forecasts to measure foresight accurately.

A Beta Item Response Theory Model of Probabilistic Forecasts to Measure Foresight

This section introduces a novel beta item response theory (IRT) model of probabilistic forecasts to measure foresight (see Appendix A for an introduction into IRT models). Probabilistic item response theory (IRT) models assume that the responses of persons to items can be described as a function of their latent abilities, the item characteristics, and random factors (Cai et al., 2016; Carroll et al., 2016; Foster et al., 2017; Hambleton et al., 1991; Hambleton & Swaminathan, 2013; Van der Linden & Hambleton, 1997). The beta IRT model describes the probabilistic forecasts of persons on a set of questions over time as a probabilistic function of the latent foresight of persons, the difficulties of the questions, the discrimination of the questions, and time effects. Formally, the beta IRT model may be written as a multilevel model (Hox et al., 2017; Raudenbush & Bryk, 2002) with two levels, where the level-1 model can be defined as

$$
\begin{aligned}
Y_{iqt} &\sim Beta(\mu_{iqt}, \phi_{qt}) \\
\mu_{iqt} &= expit\left(\beta_{\mu 0q} + \beta_{\mu 1q} t_{iqt} + \lambda_q \theta_i\right) \\
\phi_{qt} &= \exp\left(\beta_{\phi 0q} + \beta_{\phi 1q} t_{qt}\right),
\end{aligned}
\tag{1}
$$

where the probabilistic forecast[1] on the true category Y_{iqt} of forecaster $i = 1, \dots, N$ ($N \in \mathbb{Z}^+$ the number of forecasters) on question $q = 1, \dots, Q$ ($Q \in \mathbb{Z}^+$ the number of questions) at time $t_q =$

[1] Extreme probabilistic forecasts of 1 and 0 are transformed to 0.999 and 0.001 to ensure that the number space of the probabilistic forecasts $f \in [0,1]$ matches the number space of the beta distribution $Beta(\mu, \phi) \in (0,1)$.

$1, 2, \ldots, T_q$ ($T_q \in \mathbb{Z}^+$ the number of days question q was forecastable) follows a beta distribution $Beta(\mu_{iqt}, \phi_{qt})$ with mean μ_{iqt} and precision ϕ_{qt}.

The mean μ_{iqt} is modeled as a probabilistic function of the foresight θ_i of forecaster i and the time variable t_{iqt} indicating at what time t forecaster i made their probabilistic forecast Y_{iqt} on question q, where $t_{iqt} = 0$ is the first day question q was forecastable. $\beta_{\mu 0q}$ is the difficulty of question q at time $t_{iqt} = 0$ indicating how easy it was to answer a question correctly at the outset, where larger $\beta_{\mu 0q}$ indicate a higher average probability forecast on the true category of question q; $\beta_{\mu 1q}$ refers to the time effect controlling for changes in the difficulty of question q over time; and λ_q represents the discrimination of question q indicating how well the question differentiates between different degrees of foresight, which can be reinterpreted as the effect of the latent ability θ_i on the forecast on the true category of question q. In other words, the item discrimination λ_q determines how useful the latent foresight θ_i is for answering item q correctly. The $expit(x) = \frac{\exp(x)}{1+\exp(x)}$ function ensures that the number space of the linear combination $\left(\beta_{\mu 0q} + \beta_{\mu 1q} t_{iqt} + \lambda_q \theta_i\right) \in \mathbb{R}$ matches the number space of $\mu_{iqt} \in (0,1)$. The precision ϕ_{qt} (i.e., the inverse of the variance) is modeled as a probabilistic function of the time variable t_{iqt}. $\beta_{\phi 0q}$ is the precision of question q at time $t_{iqt} = 0$; and $\beta_{\phi 1q}$ refers to the time effect controlling for changes in the precision of question q over time. The natural exponential function $\exp(x) = e^x$ ensures that the number space of the linear combination $\left(\beta_{\phi 0q} + \beta_{\phi 1q} t_{qt}\right) \in \mathbb{R}$ matches the number space of $\phi_{qt} \in \mathbb{R}^+$.

The second level of the beta IRT model describing the parameters varying across questions can be formally defined as

$$\beta_{\mu 0q} = \beta_{\mu 00} + u_{\mu 0q}$$

$$\beta_{\mu 1q} = \beta_{\mu 10} + u_{\mu 1q}$$

$$\lambda_q = \lambda_0 + v_q \tag{2}$$

$$\beta_{\phi 0q} = \beta_{\phi 00} + u_{\phi 0q}$$

$$\beta_{\phi 1q} = \beta_{\phi 10} + u_{\phi 1q}$$

where the random effects of the corresponding model parameters describing how each question q or person i differs from its corresponding mean parameter are assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{\mu 0q} \\ u_{\mu 1q} \\ v_q \\ u_{\phi 0q} \\ u_{\phi 1q} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{u\mu 0} & & & & \\ \tau_{u\mu 0,u\mu 1} & \tau^2_{u\mu 1} & & & \\ \tau_{u\mu 0,v} & \tau_{u\mu 1,v} & \tau^2_{v} & & \\ \tau_{u\mu 0,u\phi 0} & \tau_{u\mu 1,u\phi 0} & \tau_{v,u\phi 0} & \tau^2_{u\phi 0} & \\ \tau_{u\mu 0,u\phi 1} & \tau_{u\mu 1,u\phi 1} & \tau_{v,u\phi 1} & \tau_{u\phi 0,u\phi 1} & \tau^2_{u\phi 1} \end{pmatrix}\right] \tag{3}$$

and the second level of the beta IRT model describing the foresight parameter varying across forecasters can be formally written as

$$\theta_i = \theta_0 + z_i, \tag{4}$$

where the random effect pertaining to the level-2 model of the latent foresight parameter is assumed to be normally distributed with

$$z_i \sim \mathcal{N}(0, \tau_z^2), \tag{5}$$

where τ_z^2 is fixed to 1 to identify the model.

$\beta_{\mu 00}$, $\beta_{\mu 1q}$, λ_q, $\beta_{\phi 0q}$, $\beta_{\phi 1q}$, and θ_0 denote the means of the corresponding parameters across all items Q or all persons N, respectively. $\tau^2_{u\mu 0}$, $\tau^2_{u\mu 1}$, τ^2_v, $\tau^2_{u\phi 0}$, and $\tau^2_{u\phi 1}$ describe the variances of the corresponding random effects. $\tau_{u\mu 0,u\mu 1}$, $\tau_{u\mu 0,v}$, $\tau_{u\mu 0,u\phi 0}$, $\tau_{u\mu 0,u\phi 1}$, $\tau_{u\mu 1,v}$, $\tau_{u\mu 1,u\phi 0}$, $\tau_{u\mu 1,u\phi 1}$, $\tau_{v,u\phi 0}$, $\tau_{v,u\phi 1}$, and $\tau_{u\phi 0,u\phi 1}$ refer to the covariations between the

corresponding parameters.

The beta IRT model assumes that foresight varies across persons, whereas the difficulty, discrimination, precision, and the respective time effects vary across questions. The beta IRT model also assume the foresight of forecasters to be unidimensional. This means that any systematic differences in the persons' forecasts on the questions result from the same underlying ability like foresight (Hambleton & Swaminathan, 2013). Although the foresight of forecasters may in principle be multidimensional (i.e., forecasters possess different foresight in two or more domains like economics and politics), forecasting tournaments usually consist of questions designed to evaluate the foresight of forecasters in one specific domain. If the latent foresight parameter θ_i is unidimensional and the question responses are independent from each other conditional on the latent foresight parameter θ_i (i.e., the residuals of the questions in the model with a unidimensional foresight parameter θ_i are uncorrelated), the questions are assumed to be locally independent (Hambleton & Swaminathan, 2013). If the assumptions of unidimensionality and local independence are fulfilled, persons can be compared on the latent ability even if they answered different questions as each question measures the same foresight dimension.

While the difficulty and discrimination parameters consider the influence of question characteristics on the forecasts of persons, the time effects in the mean and precision models control for changes in the mean and precision parameters over time. Controlling for time effects is consistent with the idea that the quantity and quality of signals increases over time, which renders accurately forecasting future states of the world easier over time (Schuler et al., 2024). Although various time effects can be conceived (Bo et al., 2017; Merkle et al., 2016, 2017), research suggests that the time effects in the beta IRT model control for various signal trajectories (Schuler et al., 2024). The beta IRT model of probabilistic forecasts can estimate the foresight of

persons more accurately than strictly proper scoring rules and alternative IRT models (see Appendix B), which made it the ideal model to estimate foresight for the regression toward the mean approach.

Data

I applied the regression toward the mean approach to foresight estimates derived from the beta IRT model using data from 265 forecasters who made 40,514 forecasts on 212 questions over two years in a forecasting tournament. Forecasting tournaments are prediction polls in which participants make probabilistic forecasts on clearly defined questions about the future state of the real world at a specific point in time (Atanasov et al., 2017; Kapoor & Wilde, 2023; Tetlock et al., 2014; Tetlock & Gardner, 2016). For each question, the participants precisely quantify their beliefs about the future by allocating probabilities to a finite number of response categories representing an exhaustive set of answers to the question. For example, forecasters may forecast the question 'Will OpenAI release ChatGPT 5.0 by Dec 31st, 2025?' by allocating probabilities to the response categories 'Yes' and 'No.' Participants can also update their probabilistic forecasts over time until the question is resolved, which enables them to account for changes in the world. Questions are resolved when the true state of the world is known either before or on the resolution date. When a question is resolved, the forecasts of each participant can be verified as it is clear which response category is the true state of the world (see Atanasov et al., 2017; Tetlock et al., 2014; Tetlock & Gardner, 2016 for more details).[2]

The forecasting tournament featured questions on geopolitical issues from 2011 to 2014. The tournament featured political questions like "Will Italy's Silvio Berlusconi resign, lose re-election/confidence vote, OR otherwise vacate office before 1 January 2012?" Research on

[2] The website https://www.gjopen.com/ illustrates what forecasting tournaments are.

nonmarket strategy (Baron, 1995; Dorobantu et al., 2017) suggests that such political events can change the institutional environment with profound implications for organizational performance. For this reason, geopolitical foresight may be considered important and valuable. For example, organizations that accurately predicted the election of Donald Trump as the 47th president of the United States well in advance may be better prepared for the impact of his policies on their organization or may have networked with relevant political decision makers sooner to shape his policies according to their interests. The forecasts were made by forecasters who participated in the first and second year of the tournament and were assigned to the control group. Focusing on forecasters in the control group eliminated potential bias in the estimates resulting from the inclusion of forecasters who received treatments during the forecasting tournament. It also ensured the plausibility of the assumption that the foresight estimates are stable over time as the interventions in the treatment groups aimed at improving the forecasters' foresight.

These data consist of probabilistic predictions by numerous forecasters over time on a large number and wide range of independent questions about future states of the world varying in their duration (*min* = 3, 1st *quartile* = 40, *median* = 84, *mean* = 122, 3rd *quartile* = 172, *max* = 520). Furthermore, the data were collected from the same forecasters in two consecutive years. Therefore, these data were ideal to quantify the extent to which foresight is luck over one year based on a regression-toward-the-mean approach.

Data Analysis

Estimating the foresight of forecasters. To estimate the foresight of forecasters in years 1 and 2, I took a Bayesian approach (Kruschke et al., 2012) using the R package *brms* (Bürkner, 2017, 2021) to fit the beta IRT model (Eq. 1–5) to the data from year 1 and the data from year 2. *brms* translates standard R syntax into Stan code, which is a Bayesian programming language that

implements Markov chain Monte Carlo (MCMC) sampling via the adaptive Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2011) and the No-U-Turn Sampler (NUTS; Hoffman & Gelman, 2014). HMC and NUTS converge faster than other Bayesian algorithms when fitting high-dimensional models with correlated parameters as in this case (Gelman et al., 2013). More generally, Bayesian data analysis provides a powerful approach to model different likelihoods, explicitly incorporate prior knowledge about model parameters, and draw statistical inferences about model parameters that can be interpreted intuitively (Gelman et al., 2013; McElreath, 2020).

I used the weakly informative default priors by *brms* to specify the beta IRT models (Bürkner, 2017, 2021). The default prior for the fixed mean and variance parameters in the normal and beta IRT models is an improper flat prior over the reals. The intercepts of the variance in the normal and the precision in the beta IRT models represent exceptions because the default for these parameters is a Student-*t* prior with mean $\mu = 0$, scale $\sigma = 2.5$, and degrees of freedom $\nu = 3$. This prior is also applied to all standard deviations[3] of the random effects except for the standard deviation of the foresight parameter, which I fixed to the constant 1 to identify the beta IRT model. The LKJ prior with parameter $\zeta = 1$ (Lewandowski et al., 2009) is the default prior for the correlations between the random effects in the beta IRT model.[4]

I fitted the models using four chains, each with 8,000 iterations consisting of 4,000 warm-up and 4,000 posterior samples resulting in a total of 16,000 posterior samples. I set the initial values for the sampler to 0, the maximum tree depth to 10, and the target average proposal acceptance probability during Stan's adaptation period to 0.85. These settings ensured high-quality posterior samples. I assessed model convergence by inspecting the trace plots (Gabry et al., 2019; Gelman et al., 2013) and the potential scale reduction factor $\hat{R}$ (Brooks & Gelman, 1998; Gelman

[3] The variances τ^2 of the random effects in Eq. 3 and 5 are modeled as standard deviations τ in *brms*.
[4] The covariances $\tau_{X,Y}$ of the parameters in Eq. 3 are modeled as correlations $\rho_{X,Y}$ in brms.

et al., 2013; Gelman & Rubin, 1992; Vehtari et al., 2021) of each parameter in the beta IRT model. The trace plots indicated stationarity, good mixing, and the convergence of the four chains for each parameter. The potential scale reduction factor $\hat{R}$ was 1 with bulk and tail effective sample sizes > 100 per chain for each parameter. This suggests that the algorithms converged successfully to the target posterior distributions of the beta IRT model parameters in both years.

The unidimensionality assumption of the beta IRT model is plausible as all questions in the data were related to geopolitics. Furthermore, the assumption of question independence is plausible as each question did not depend on another question for resolution (e.g., knowing the answer on one question was not required to accurately predict another question). As the unidimensionality assumption and assumption of question independence are plausible, the assumption of local independence is plausible. Therefore, forecasters can be compared on the same foresight dimension even if they forecast different questions. A plot of the question discriminations λ_q suggested substantial variation among them, which lends support to the assumption of question-specific discriminations. While the credible interval of the time effect in the mean parameter model suggested a positive time effect $\hat{\beta}_{\mu 10} = 0.02$ ($EE = 0.00, 95\%\ CI = [0.02; 0.03]$), the credible interval of the time effect in the variance parameter model did not support the notion a time effect $\hat{\beta}_{\phi 10} = 0.00$ ($EE = 0.00, 95\%\ CI = [0.00; 0.00]$). This indicates that while the forecasts became more accurate over time, they did not necessarily become more precise over time. Overall, this suggests that the assumptions of the beta IRT model are plausible.

Quantifying the extent to which foresight is luck. To quantify the extent to which foresight is luck using the regression-toward-the-mean approach, I first extracted the foresight estimates from the beta IRT models of years 1 and 2. Second, I standardized the foresight estimates in each year to ensure that their means and variances are identical (means of 0 and variances of 1).

Furthermore, I plotted the distributions of the standardized foresight estimates in each year. These plots showed that the foresight parameter estimates were normally distributed. A two-sided Kolmogorov-Smirnov test suggested that the distributions were identical ($D = 0.07,\ p = 0.574$). Overall, this suggests that the foresight estimates of year 1 and 2 are identically distributed. Third, I visualized the relationship between the standardized foresight estimates from year 2 on the standardized foresight estimates from year 1 using locally estimated scatterplot smoothing (LOESS) curves with spans of 0.2, 0.4, 0.6, and 0.8. These plots suggested that there may be a potential non-linear relationship between the foresight estimates from years 1 and 2 that may be captured by a reversion-toward-the-mean approach as the foresight estimates also exhibited weak positive dependence: The mean foresight estimates in year 2 of forecasters with above average as well as forecasters with below average foresight in year 1 reverted toward the mean foresight of forecasters (i.e., 0) but not beyond.

Fourth, I implemented the regression-toward-the-mean approach by regressing the standardized foresight estimates from year 2 on the standardized foresight estimates from year 1:

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

where y_i is the standardized foresight estimate of forecaster i in year 2, β_0 denotes the intercept representing the average foresight of the forecasters in year 2, β_1 refers to the effect of the standardized foresight estimate of forecaster i in year 1 x_i on the standardized foresight estimate of forecaster i in year 2 y_i, and ϵ_i is the residual of forecaster i. The regression coefficient β_1 describes the extent to which foresight is skill and the relative regression toward the mean $\rho = 1 - \beta_1$ denotes the extent to which foresight is luck assuming that the relationship between the standardized foresight estimates of years 1 and 2 is linear.

Fifth, to capture the potential non-linear relationship, I fitted a second-order polynomial

regression to model the relationship between the standardized foresight estimate from year 2 and the standardized foresight estimate from year 1, including both linear and quadratic terms:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * x_i^2 + \epsilon_i$$

where β_1 denotes the effect of the standardized foresight estimate of forecaster i in year 1 x_i on the standardized foresight estimate of forecaster i in year 2 y_i, and β_2 refers to the effect of the squared standardized foresight estimate of forecaster i in year 1 x_i^2 on the standardized foresight estimate of forecaster i in year 2 y_i. I also fitted a third-order polynomial regression to model the relationship between the standardized foresight estimate from year 2 and the standardized foresight estimate from year 1, including linear, quadratic, and cubic terms:

$$y_i = \beta_0 + \beta_1 * x_i + \beta_2 * x_i^2 + \beta_3 * x_i^3 + \epsilon_i$$

where β_3 refers to the effect of the cubed standardized foresight estimate of forecaster i in year 1 x_i^3 on the standardized foresight estimate of forecaster i in year 2 y_i. I selected the regression model for the final interpretation based on the Akaike information criterion (AIC), Bayesian information criterion (BIC), and likelihood-ratio tests comparing the goodness of fit of these regression models. All regressions were fitted using ordinary least squares (OLS). All assumptions of OLS were fulfilled.

Note that the coefficients in the second-order and third-order polynomial regressions cannot easily be interpreted as a measure of skill or luck because the relationship between the standardized foresight estimates is nonlinear. For example, forecasters with above or below average foresight in year 1 may regress more strongly toward the mean in year 2 than forecasters with average foresight in year 1. In other words, the size of the regression toward the mean depends on the foresight of forecasters. In this case, the extent to which foresight is luck may be quantified by the reversion toward the mean κ. For the upper portion of the foresight distribution as defined

by cutoff c with $c \geq \mu$, the reversion toward the mean is defined by (Samuels, 1991)

$$\kappa = \frac{E[X_1|X_1 > c] - E[X_2|X_1 > c]}{E[X_1|X_1 > c] - \mu},$$

For the lower portion of the foresight distribution as defined by cutoff c with $c \leq \mu$, the reversion toward the mean is defined by (Samuels, 1991)

$$\kappa = \frac{E[X_1|X_1 < c] - E[X_2|X_1 < c]}{E[X_1|X_1 < c] - \mu},$$

where X_1 and X_2 are identically distributed random variables (e.g., the foresight of forecasters in years 1 and 2), $X_1 > c$ selects all forecasters who possess foresight in year 1 above cutoff c, $X_1 < c$ selects all forecasters who possess foresight in year 1 below cutoff c, and μ is the population mean of X_1 and X_2. The reversion of the upper portion of the foresight distribution toward the mean can be empirically estimated by

$$\hat{\kappa} = \frac{\sum_{x_i > c} x_i - \sum_{x_i > c} y_i}{\sum_{x_i > c} x_i - \sum_{i=1}^{n} x_i}$$

and analogously for the lower portion of the foresight distribution by

$$\hat{\kappa} = \frac{\sum_{x_i < c} x_i - \sum_{x_i < c} y_i}{\sum_{x_i < c} x_i - \sum_{i=1}^{n} x_i}$$

where x_i denotes the foresight of forecaster i at time 0, $x_i > c$ selects all forecasters whose foresight in year 1 is above cutoff c, $x_i < c$ selects all forecasters whose foresight in year 1 is below cutoff c, and y_i refers to the foresight of forecaster i at time 2. Statistical significance may be determined by a bootstrapping approach that estimates 95% percentile confidence intervals of the reversion toward the mean $\hat{\kappa}$ based on 10,000 resamples of the data. In the following, I will report the results of these analyses.

RESULTS

Table 2 shows the results of the first-order, second-order, and third-order polynomial regression

models and their comparison. The regression coefficient showed a strong positive relationship between the standardized foresight estimates from year 1 and the standardized foresight estimates from year 2 $\left(\hat{\beta}_1 = 0.68, SE = 0.05, t = 14.85, p < 0.001\right)$, which implied that foresight regressed by $\hat{\rho} = 1 - 0.68 = 0.32$ toward the mean foresight of forecasters. This suggests that foresight is 68% skill and 32% luck. This means that the foresight of an individual forecaster regresses on average by 32% toward the mean foresight of all forecasters over one year. Figure 1 shows the regression of the foresight estimates toward the mean foresight of all forecasters from year 1 to year 2 based on the first-order linear regression model. Note that this result may somewhat underestimate the extent of skill and overestimate the extent of luck in foresight because any measurement error in relation to the foresight of forecasters attenuates the regression coefficient (Fuller & Hidiroglou, 1978).

Insert Table 2 about here

Insert Figure 1 about here

Regarding the model fit, the results show that the first-order linear regression model had the smallest *BIC* (606.40), whereas the third-order linear regression model had the smallest *AIC* (591.85). The likelihood-ratio tests indicate that the second-order linear regression model did not fit the data better than the first-order linear regression model $(\chi^2(1) = 0.13, p = .608)$. However, the third-order linear regression model fit the data better than the first-order linear regression model $(\chi^2(1) = 4.03, p < 0.01)$. Overall, these results suggest that the third-order linear regression model fit the data best, which implied a non-linear relationship between the standardized foresight estimates. This means that the role of luck in foresight seemed to depend on the foresight of forecasters.

Figure 2 shows the regression of the foresight estimates toward the mean foresight of all forecasters from year 1 to year 2 based on the third-order linear regression model. The third-order linear regression model suggests that the regression toward the mean followed an S-shaped curve, where luck appeared to play a stronger role in high and low foresight but a weaker role in average foresight. To quantify the extent to which foresight is luck depending on the foresight of forecasters, I computed the reversion toward the mean for forecasters with foresight above and below multiple cutoffs. The results displayed in Table 3 suggest that the reversion toward the mean was stronger for forecasters with extreme foresight ($x_i > 3;\ 38\%, 95\%\ CI = [0.11, 0.70]$) and exceptionally good foresight ($x_i > 2;\ 35\%, 95\%\ CI = [0.20, 0.73]$). Furthermore, the reversion toward the mean was stronger for forecasters with below average foresight ($x_i < -1;\ 47\%, 95\%\ CI = [0.26, 0.73]$) and exceptionally poor foresight ($x_i < -2;\ 54\%, 95\%\ CI = [0.31, 0.82]$). In contrast, the reversion toward the mean was weaker for the average forecaster in the upper ($x_i > 0;\ 23\%, 95\%\ CI = [0.06, 0.47]$) and lower portion of the foresight distribution ($x_i < 0;\ 23\%, 95\%\ CI = [0.07, 0.41]$).[5] Figure 3 illustrates the reversion of the foresight estimates toward the mean foresight of all forecasters from year 1 to year 2 for different portions of the foresight distribution as defined by cutoff c.

Insert Figure 2 about here

Insert Figure 3 about here
---

[5] Note that the reversion toward the mean approach only allows for the calculation of effects in upper and lower portions of the foresight distribution as defined by one cutoff. Therefore, it is currently not possible to estimate the reversion toward the mean of average forecasters (i.e., $-1 < x_i < 1$) as this would require two cutoffs for which the reversion toward the mean is not defined (Samuels, 1991). For this reason, estimating the average reversion toward the mean for forecasters in the upper and lower portion of the foresight distribution represents the best estimate of the reversion toward the mean of the average forecaster, albeit the true reversion toward the mean is likely smaller given that forecasters with extreme and exceptionally good foresight exhibit a disproportionately strong reversion toward the mean foresight of forecasters.

Insert Table 3 about here

Furthermore, Figures 2 and 3 suggest that foresight followed a weak pattern of luck: Forecasters who accurately (inaccurately) predicted future states of the world in year 1 predicted future states of the world less accurately (more accurately) in year 2. However, these forecasters did not predict future states of the world worse (better) than forecasters with above (below) average foresight in year 1. This becomes evident by comparing the mean foresight of forecasters in year 2 who showed extreme foresight ($x_i > 3$) and to the mean foresight of exceptionally good forecasters ($2 < x_i < 3$), where the mean difference δ of their foresight is nonsignificant based on 95% percentile confidence intervals of 10,000 bootstraps samples $\left(\hat{\delta} = 0.53, 95\%\, CI = [-0.63, 1.81]\right)$. This is corroborated by comparing the mean foresight of forecasters in year 2 who showed exceptionally poor foresight ($x_i < -2$) to the mean foresight of below average forecasters ($-2 < x_i < -1$), where forecasters with exceptionally poor foresight exhibited worse foresight in year 2 than forecasters with below average foresight in year 2 $\left(\hat{\delta} = -0.52, 95\%\, CI = [-0.98, -0.07]\right)$. Overall, these results suggest that even though luck plays a stronger role in exceptionally good (bad) foresight, such foresight may still signal good (bad) judgment. More generally, these results suggest that even though foresight is to some extent luck, foresight appears to be mainly skill.

DISCUSSION

Motivated by the goal of assessing the relative plausibility of cognitive, evolutionary, and chance theories of foresight, this paper quantified the extent to which foresight is luck using a regression-toward-the-mean approach. The results suggest that foresight is 32% luck, albeit luck plays a greater role among top forecasters (35%) and forecasters with below average foresight (47%) but

a lesser role among average forecasters (23%). The analyses also uncovered that foresight follows a weak pattern of luck. In the following, I will discuss the theoretical and practical implications, acknowledge the limitations of this study, and outline opportunities for future research.

Theoretical Implications

The finding that foresight is 32% luck has important implications for cognitive, evolutionary, and chance explanations of foresight: Cognitive theories suggest that strategists possess "boundedly rational foresight" (Gavetti et al., 2012, p. 9) that is grounded in their mental representations of the world rendering foresight mainly a skill (Csaszar, 2018; Csaszar & Laureiro-Martínez, 2018; Gavetti & Levinthal, 2000). This explanation is consistent with the finding that foresight appears to be mainly a skill but also somewhat luck highlighting that foresight is boundedly rational. It is also consistent with a growing literature highlighting the importance of mental representations and cognitive processes for foresight (Csaszar & Laureiro-Martínez, 2018; Csaszar & Ostler, 2020; Davis et al., 2009, 2009; Eisenhardt & Sull, 2001; Gavetti & Levinthal, 2000; Kapoor & Wilde, 2023; Lee & Csaszar, 2020; Satopää et al., 2021; Sull & Eisenhardt, 2015). Although skill may also refer to any other individual factors than cognition like personality characteristics and motivation, research suggests that such factors only play a minor role in the explanation of foresight (Schuler et al., 2024). Thus, the result that foresight is 32% luck may be viewed as evidence for cognitive explanations of foresight.

Evolutionary theories acknowledge that strategists may exhibit myopic foresight grounded in their cognitive preadaptation that enables them to discover opportunities serendipitously (Denrell et al., 2003; Gavetti & Lecuona Torras, 2021; Gavetti & Menon, 2016). This explanation is partly inconsistent with the finding that foresight appears to be mainly a skill as foresight deriving from the serendipitous discovery of opportunities in the environment should be largely

luck (Busch, 2020, 2024; Denrell et al., 2003). Although forecasters might have been cognitively preadapted to geopolitical questions, the number, breadth (i.e., diversity of knowledge domains), and time horizon of the geopolitical questions covered in the forecasting tournament make this explanation less plausible compared to cognitive explanations that allow for strategists to anticipate a wide variety of distant future states of the world by learning and adapting mental representations. That being said, cognitive preadaptation and serendipity likely still play a role in foresight to some extent, where cognitive preadaptation contributes to skill.

Chance theories question the existence of managerial foresight by emphasizing the role of random processes in accurately predicting future states of the world (Denrell et al., 2015; Denrell & Fang, 2010; Liu, 2020; Liu & de Rond, 2016). However, this explanation is inconsistent with the finding that foresight is on average only 32% luck highlighting the role of cognition and cognitive preadaptation in foresight. Furthermore, foresight followed a weak pattern of luck (Denrell et al., 2019; Liu, 2020): The foresight of forecasters who accurately (inaccurately) predicted future states of the world in year 1 predicted future states of the world less accurately (more accurately) in year 2, and this pattern was stronger for forecasters with exceptionally good and bad foresight. Yet, these forecasters still exhibited on average equal (worse) foresight compared to forecasters with moderately accurate (inaccurate) foresight. Thus, this study refines prior research suggesting that exceptional foresight may be a signal of poor judgment (Denrell & Fang, 2010; Denrell & Liu, 2012) by highlighting that extremely good (bad) foresight may still signal good (bad) judgment. Although chance explanations may be less plausible compared to cognitive and evolutionary explanations, they still legitimately emphasize that foresight involves luck. Taken together, even though cognitive theories appear to be more consistent with the finding that foresight is 32% luck than alternative explanations, these results highlight that cognitive,

evolutionary, and chance explanations of foresight may be best viewed as complements to each other as each theory emphasizes an important factor in the strategist's foresight.

Practical Implications

The findings in relation to cognitive, evolutionary, and chance explanations of foresight also have important practical implications: Cognitive theories suggest that strategists who displayed exceptionally good foresight may be entrusted with the deliberate design of strategies (Mintzberg, 1978; Mintzberg & Waters, 1985) based on their mental representations of the world to increase organizational performance. For example, organizations may select strategists based on their foresight in forecasting tournaments. Strategists may also seek to learn better representations of the world that can guide their strategic decisions (Csaszar et al., 2024; Csaszar & Ostler, 2020; Heshmati & Csaszar, 2024). Furthermore, strategists may still deepen and broaden their knowledge to serendipitously exploit an increasing range and number of opportunities they see that others cannot see when the environment changes (Busch, 2020, 2024; Denrell et al., 2003). At the same time, strategists should rely on evolutionary principles (Levinthal, 2021; Murmann, 2003; Nelson & Winter, 1982) to guide emergent strategies (Mintzberg, 1978; Mintzberg & Waters, 1985). For instance, strategists may guide the organization by communicating a strategic vision and devise internal selection criteria for projects in the organization (Levinthal, 2017, 2021) that guide the generation and promotion of strategic initiatives (Burgelman, 1983a, 1983b). Furthermore, strategists may design organizational structures to support these activities (Csaszar, 2013). More generally, strategists should remain humble as foresight is always to some extent luck (Denrell et al., 2015; Liu, 2020).

Limitations and Future Research Opportunities

This study is limited in two respects, albeit these limitations represent valuable

opportunities for future research. First, this study uses data from a geopolitical forecasting tournament to evaluate the relative merit of cognitive, evolutionary, and chance explanations of foresight. A large and diverse number of questions about the future in this domain provides a good way of assessing the relative plausibility of cognitive and evolutionary theories of foresight. As the number and diversity of the questions increase, cognitive theories become more plausible compared to evolutionary theories as it becomes more difficult to display foresight in a large number of diverse questions simply by cognitive preadaptation. However, future research may subject the relative plausibility of cognitive and evolutionary theories to an even stronger test by quantifying the extent to which foresight is luck using data from multiple domains. Exhibiting foresight across a wide range of knowledge domains (e.g., technology, politics, economics) would render it even harder to display foresight simply by cognitive preadaptation. Studying the extent of foresight in other domains could also provide insights into how the extent of luck and skill in foresight varies across different domains.

Second, this study used data from two years of a geopolitical forecasting tournament. Although this provides important insights about the regression of foresight toward the mean over the time of one year, it remains unclear to what extent foresight regresses toward the mean over longer time horizons (e.g., three to five years). Examining the role of luck in foresight over such long time horizons would require exceptional commitment by forecasters as they would have to participate in two consecutive forecasting tournaments consisting of questions lasting for three to five years. However, examining the role of luck in such distant foresight may provide valuable additional insights for the relative plausibility of cognitive, evolutionary, and chance theories of foresight. Therefore, future research should examine the regression of foresight toward the mean over longer time horizons and evaluate the implications for different explanations of foresight as

the relative plausibility of these theories may change with the time horizon of foresight. For example, luck may play a greater role in long-term foresight rendering evolutionary and chance theories more plausible than cognitive theories for extremely distant foresight.

Conclusion

Cognitive, evolutionary, and chance explanations of foresight each provide unique insights into why strategists may or may not display foresight. While foresight is partly luck, the findings of this paper add to the mounting evidence that foresight is a skill rooted in accurate mental representations of the world. As foresight can be a source of competitive advantage for organizations, understanding why some managers have foresight and how managers may improve their foresight holds great promise for strategy. For this reason, I hope that this study inspires other researchers to rigorously advance our understanding of foresight.

REFERENCES

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P. E., Ungar, L., & Mellers, B. A. (2017). Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atanasov, P., Witkowski, J., Ungar, L., Mellers, B. A., & Tetlock, P. E. (2020). Small steps to accuracy: Incremental belief updaters are better forecasters. *Organizational Behavior and Human Decision Processes*, *160*, 19–35. https://doi.org/10.1016/j.obhdp.2020.02.001

Baron, D. P. (1995). Integrated Strategy: Market and Nonmarket Components. *California Management Review*, *37*(2), 47–65. https://doi.org/10.2307/41165788

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–472). Addison-Wesley Publishing.

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

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Royal Statistical Society. Journal. Series A: General*, *143*(4), 383–404. https://doi.org/10.2307/2982063

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Burgelman, R. A. (1983a). A Process Model of Internal Corporate Venturing in the Diversified Major Firm. *Administrative Science Quarterly*, *28*(2), 223. https://doi.org/10.2307/2392619

Burgelman, R. A. (1983b). Corporate Entrepreneurship and Strategic Management: Insights from a Process Study. *Management Science*, *29*(12), 1349–1364. https://doi.org/10.1287/mnsc.29.12.1349

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*, 1–54. https://doi.org/10.18637/jss.v100.i05

Busch, C. (2020). *The Serendipity Mindset: The Art and Science of Creating Good Luck*. Penguin.

Busch, C. (2024). Towards a Theory of Serendipity: A Systematic Review and Conceptualization. *Journal of Management Studies*, *61*(3), 1110–1151. https://doi.org/10.1111/joms.12890

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item Response Theory. *Annual Review of Statistics and Its Application*, *3*(1), 297–321. https://doi.org/10.1146/annurev-statistics-041715-033702

Carroll, R. J., Primo, D. M., & Richter, B. K. (2016). Using item response theory to improve measurement in strategic management research: An application to corporate social responsibility. *Strategic Management Journal*, *37*(1), 66–85. https://doi.org/10.1002/smj.2463

Csaszar, F. A. (2013). An Efficient Frontier in Organization Design: Organizational Structure as a Determinant of Exploration and Exploitation. *Organization Science*, *24*(4), 1083–1101. https://doi.org/10.1287/orsc.1120.0784

Csaszar, F. A. (2018). What Makes a Decision Strategic? Strategic Representations. *Strategy Science*, *3*(4), 606–619. https://doi.org/10.1287/stsc.2018.0067

Csaszar, F. A., Hinrichs, N., & Heshmati, M. (2024). External representations in strategic decision-making: Understanding strategy's reliance on visuals. *Strategic Management Journal*, *45*(11), 2191–2226. https://doi.org/10.1002/smj.3613

Csaszar, F. A., & Laureiro-Martínez, D. (2018). Individual and Organizational Antecedents of Strategic Foresight: A Representational Approach. *Strategy Science*, *3*(3), 513–532. https://doi.org/10.1287/stsc.2018.0063

Csaszar, F. A., & Levinthal, D. A. (2016). Mental representation and the discovery of new strategies. *Strategic Management Journal*, *37*(10), 2031–2049. https://doi.org/10.1002/smj.2440

Csaszar, F. A., & Ostler, J. (2020). A Contingency Theory of Representational Complexity in Organizations. *Organization Science*, *31*(5), 1198–1219. https://doi.org/10.1287/orsc.2019.1346

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2009). Optimal Structure, Market Dynamism, and the Strategy of Simple Rules. *Administrative Science Quarterly*, *54*(3), 413–452. https://doi.org/10.2189/asqu.2009.54.3.413

Denrell, J. (2004). Random Walks and Sustained Competitive Advantage. *Management Science*, *50*(7), 922–934. https://doi.org/10.1287/mnsc.1030.0143

Denrell, J., & Fang, C. (2010). Predicting the Next Big Thing: Success as a Signal of Poor Judgment. *Management Science*, *56*(10), 1653–1667. https://doi.org/10.1287/mnsc.1100.1220

Denrell, J., Fang, C., & Liu, C. (2015). Perspective—Chance Explanations in the Management

Sciences. *Organization Science*, *26*(3), 923–940. https://doi.org/10.1287/orsc.2014.0946

Denrell, J., Fang, C., & Liu, C. (2019). In Search of Behavioral Opportunities From Misattributions of Luck. *Academy of Management Review*, *44*(4), 896–915. https://doi.org/10.5465/amr.2017.0239

Denrell, J., Fang, C., & Winter, S. G. (2003). The economics of strategic opportunity. *Strategic Management Journal*, *24*(10), 977–990. https://doi.org/10.1002/smj.341

Denrell, J., Fang, C., & Zhao, Z. (2013). Inferring superior capabilities from sustained superior performance: A Bayesian analysis. *Strategic Management Journal*, *34*(2), 182–196. https://doi.org/10.1002/smj.2001

Denrell, J., & Liu, C. (2012). Top performers are not the most impressive when extreme performance indicates unreliability. *Proceedings of the National Academy of Sciences*, *109*(24), 9331–9336. https://doi.org/10.1073/pnas.1116048109

Dorobantu, S., Kaul, A., & Zelner, B. (2017). Nonmarket strategy research through the lens of new institutional economics: An integrative review and future directions. *Strategic Management Journal*, *38*(1), 114–140. https://doi.org/10.1002/smj.2590

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. https://doi.org/10.1016/0370-2693(87)91197-X

Eisenhardt, K. M., & Sull, D. N. (2001). Strategy as simple rules. *Harvard Business Review*, *79*(1), 106–116.

Ferrando, P. J. (2001). A nonlinear congeneric model for continuous item responses. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 293–313. https://doi.org/10.1348/000711001159573

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of Item Response Theory Practices in

Organizational Research: Lessons Learned and Paths Forward. *Organizational Research Methods*, *20*(3), 465–486. https://doi.org/10.1177/1094428116689708

Fuller, W. A., & Hidiroglou, M. A. (1978). Regression Estimation after Correcting for Attenuation. *Journal of the American Statistical Association*, *73*(361), 99–104. https://doi.org/10.1080/01621459.1978.10480011

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

Gavetti, G. (2012). Toward a Behavioral Theory of Strategy. *Organization Science*, *23*(1), 267–285. https://doi.org/10.1287/orsc.1110.0644

Gavetti, G., Greve, H. R., Levinthal, D. A., & Ocasio, W. (2012). The Behavioral Theory of the Firm: Assessment and Prospects. *Academy of Management Annals*, *6*, 1–40. https://doi.org/10.1080/19416520.2012.656841

Gavetti, G., & Lecuona Torras, J. R. (2021). A Neo-Carnegie Approach to the Agency Question: Bridging the Evolutionary and Cognitive Views of Strategy. *Strategy Science*, *6*(4), 353–359. https://doi.org/10.1287/stsc.2021.0149

Gavetti, G., & Levinthal, D. A. (2000). Looking Forward and Looking Backward: Cognitive and Experiential Search. *Administrative Science Quarterly*, *45*(1), 113–137. https://doi.org/10.2307/2666981

Gavetti, G., Levinthal, D. A., & Rivkin, J. W. (2005). Strategy making in novel and complex worlds: The power of analogy. *Strategic Management Journal*, *26*(8), 691–712. https://doi.org/10.1002/smj.475

Gavetti, G., & Menon, A. (2016). Evolution Cum Agency: Toward a Model of Strategic Foresight.

Strategy Science, *1*(3), 207–233. https://doi.org/10.1287/stsc.2016.0018

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*(3), 445–449. https://doi.org/10.1214/08-BA318

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. *arXiv:2011.01808 [Stat]*. http://arxiv.org/abs/2011.01808

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.1984.4767596

Gupta, A. K., & Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications*. CRC Press.

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2), 139–150. https://doi.org/10.2307/2086306

Hambleton, R. K., & Swaminathan, H. (2013). *Item Response Theory: Principles and Applications*. Springer Science & Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response*

Theory. SAGE.

Heshmati, M., & Csaszar, F. A. (2024). Learning Strategic Representations: Exploring the Effects of Taking a Strategy Course. *Organization Science*, *35*(2), 453–473. https://doi.org/10.1287/orsc.2023.1676

Himmelstein, M., Atanasov, P., & Budescu, D. V. (2021). Forecasting forecaster accuracy: Contributions of past performance and individual differences. *Judgment and Decision Making*, *16*(2), 323–362. https://doi.org/10.1017/S1930297500008597

Himmelstein, M., Budescu, D. V., & Han, Y. (2022). The Wisdom of Timely Crowds. In *Judgment in Predictive Analytics*. Springer.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hox, J., Moerbeek, M., & Schoot, R. van de. (2017). *Multilevel Analysis: Techniques and Applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Kapoor, R., & Wilde, D. (2023). Peering into a crystal ball: Forecasting behavior and industry foresight. *Strategic Management Journal*, *44*(3), 704–736. https://doi.org/10.1002/smj.3450

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, *15*(4), 722–752. https://doi.org/10.1177/1094428112457829

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin.

Lee, S., & Csaszar, F. A. (2020). Cognitive and Structural Antecedents of Innovation: A Large-Sample Study. *Strategy Science*, *5*(2), 71–97. https://doi.org/10.1287/stsc.2020.0107

Levinthal, D. A. (2017). Mendel in the C-Suite: Design and the Evolution of Strategies. *Strategy*

Science, *2*(4), 282–287. https://doi.org/10.1287/stsc.2017.0047

Levinthal, D. A. (2021). *Evolutionary Processes and Organizational Adaptation: A Mendelian Perspective on Strategic Management*. Oxford University Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Liu, C. (2020). *Luck: A key idea for business and society*. Routledge.

Liu, C., & de Rond, M. (2016). Good Night, and Good Luck: Perspectives on Luck in Management Scholarship. *Academy of Management Annals*, *10*(1), 409–451. https://doi.org/10.5465/19416520.2016.1120971

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*, *7*, x, 84–x, 84.

March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McGahan, A. M., & Porter, M. E. (1997). How Much Does Industry Matter, Really? *Strategic Management Journal*, *18*(S1), 15–30. https://doi.org/10.1002/(SICI)1097-0266(199707)18:1+<15::AID-SMJ916>3.0.CO;2-1

Mellenbergh, G. J. (1994). A Unidimensional Latent Trait Model for Continuous Item Responses. *Multivariate Behavioral Research*, *29*(3), 223–236. https://doi.org/10.1207/s15327906mbr2903_2

Mellers, B. A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. E. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on*

Psychological Science, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Merkle, E. C., Steyvers, M., Mellers, B. A., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, *3*(1), 1–19. https://doi.org/10.1037/dec0000032

Merkle, E. C., Steyvers, M., Mellers, B. A., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832. https://doi.org/10.1016/j.ijforecast.2017.04.002

Mintzberg, H. (1978). Patterns in Strategy Formation. *Management Science*, *24*(9), 934–948. https://doi.org/10.1287/mnsc.24.9.934

Mintzberg, H., & Waters, J. (1985). Of Strategies, Deliberate and Emergent. *Strategic Management Journal*, *6*(3), 257–272. https://doi.org/10.1002/smj.4250060306

Müller, H. (1987). A rasch model for continuous ratings. *Psychometrika*, *52*(2), 165–181. https://doi.org/10.1007/BF02294232

Murmann, J. P. (2003). *Knowledge and Competitive Advantage: The Coevolution of Firms, Technology, and National Institutions*. Cambridge University Press.

Neal, R. M. (2011). MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change* (digitally reprinted). The Belknap Press of Harvard Univ. Press.

Noel, Y., & Dauvier, B. (2007). A Beta Item Response Model for Continuous Bounded Responses. *Applied Psychological Measurement*, *31*(1), 47–73. https://doi.org/10.1177/0146621605287691

Nott, D. J., Drovandi, C., & Frazier, D. T. (2024). Bayesian Inference for Misspecified Generative

Models. *Annual Review of Statistics and Its Application*, *11*(1), null. https://doi.org/10.1146/annurev-statistics-040522-015915

Packard, M. D., & Clark, B. B. (2020). Mitigating versus Managing Epistemic and Aleatory Uncertainty. *Academy of Management Review*, *45*(4), 872–876. https://doi.org/10.5465/amr.2020.0266

Porter, M. E. (1979). How Competitive Forces Shape Strategy. *Harvard Business Review*.

Porter, M. E. (2008). The Five Competitive Forces That Shape Strategy. *Harvard Business Review*, *86*(1), 78–93.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Rumelt, R. P. (1991). How much does industry matter? *Strategic Management Journal*, *12*(3), 167–185. https://doi.org/10.1002/smj.4250120302

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*, *37*(1, Pt. 2), 68–68.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203–219. https://doi.org/10.1007/BF02291114

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, *39*(1), 111–121. https://doi.org/10.1007/BF02291580

Samuels, M. L. (1991). Statistical Reversion Toward the Mean: More Universal Than Regression Toward the Mean. *The American Statistician*, *45*(4), 344–346.

https://doi.org/10.2307/2684474

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. A. (2021). Bias, Information, Noise: The BIN Model of Forecasting. *Management Science*, *67*(12), 7599–7618. https://doi.org/10.1287/mnsc.2020.3882

Schmittlein, D. C. (1989). Surprising Inferences from Unsurprising Observations: Do Conditional Expectations Really Regress to the Mean? *The American Statistician*, *43*(3), 176–183. https://doi.org/10.2307/2685070

Schuler, B. A., Murmann, J. P., & Beisemann, M. (2022). Integrating the Accuracy and Time Dimension of Foresight. *Academy of Management Proceedings*, *2022*(1), 13788. https://doi.org/10.5465/AMBPP.2022.13788abstract

Schuler, B. A., Murmann, J. P., Beisemann, M., & Satopää, V. (2024). *Individual Foresight: Concept, Operationalization, and Correlates*. https://www.alexandria.unisg.ch/handle/20.500.14171/121555

Simon, H. A. (1947). *Administrative Behavior*. Macmillan.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sull, D. N., & Eisenhardt, K. M. (2015). *Simple rules: How to thrive in a complex world*. Houghton Mifflin Harcourt.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction*. Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting Tournaments: Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in*

Psychological Science, *23*(4), 290–295. https://doi.org/10.1177/0963721414534257

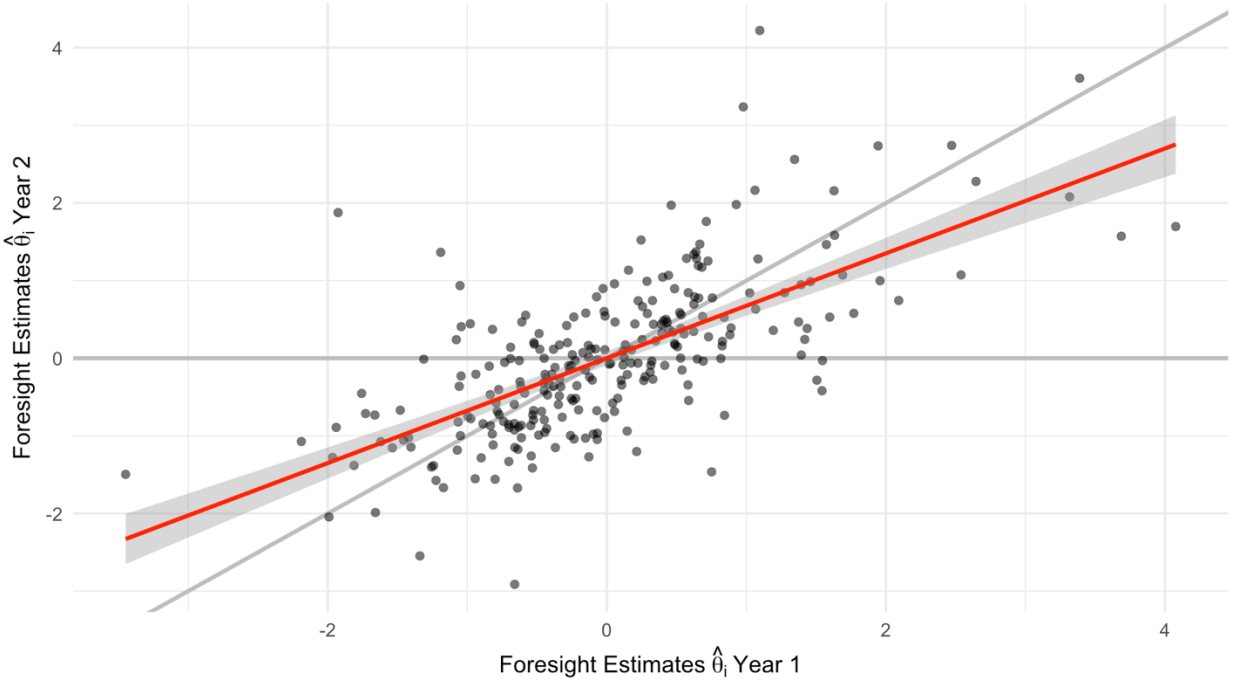
Van der Linden, W. J., & Hambleton, R. (1997). *Handbook of item response theory* (Vol. 1).

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis*, *16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Verkuilen, J., & Smithson, M. (2012). Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution. *Journal of Educational and Behavioral Statistics*, *37*(1), 82–113. https://doi.org/10.3102/1076998610396895

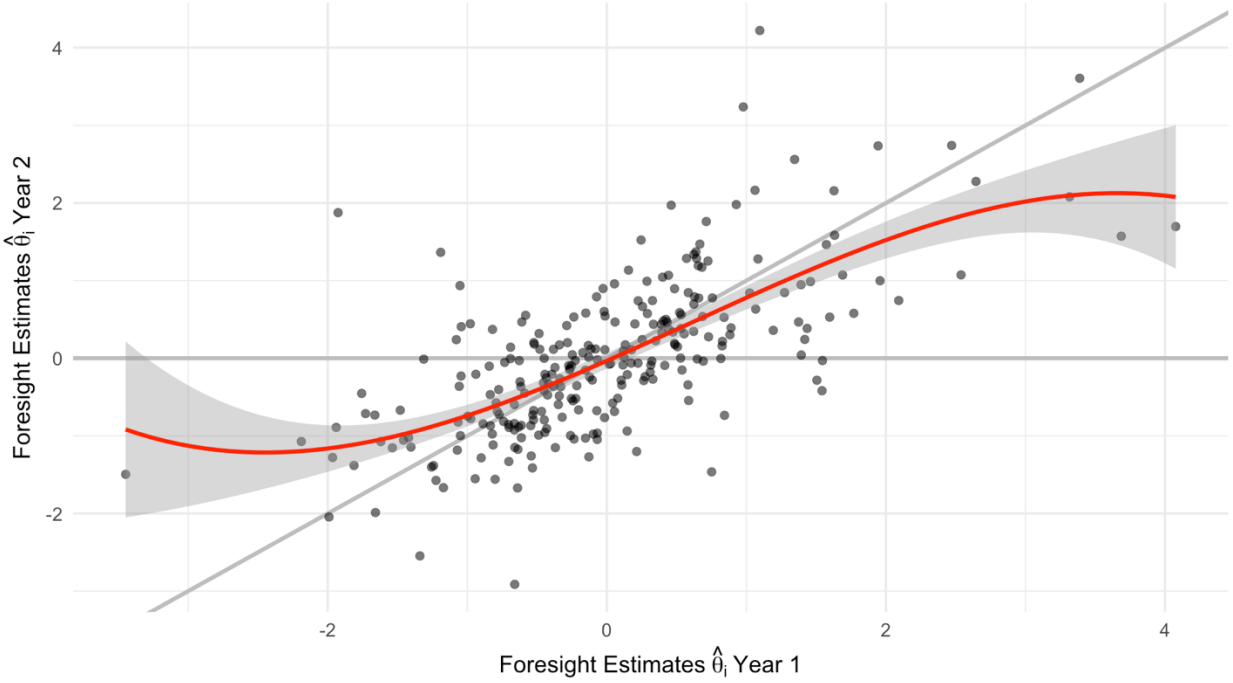
Wang, M. Z. (2023). Changes in industry and corporate effects in the United States, 1978–2019. *Strategic Management Journal*, *44*(2), 477–490. https://doi.org/10.1002/smj.3445

Figure 1. Regression of the Foresight Estimates Toward the Mean Foresight of all Forecasters From Year 1 to Year 2 Based on the First-Order Linear Regression Model



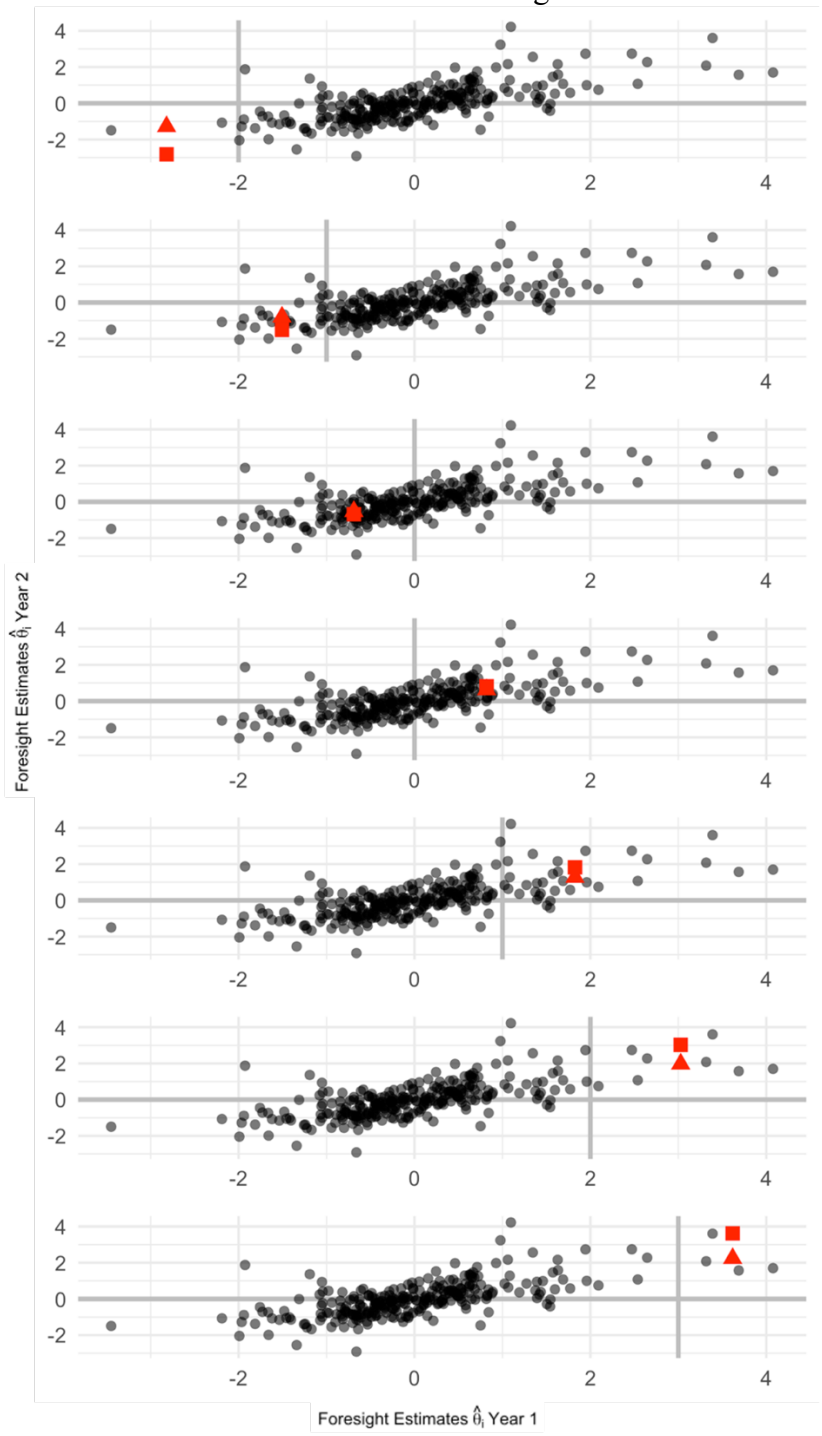
Note. The foresight estimates from the beta IRT model in year 1 and year 2 are standardized. The horizontal grey line indicates the mean foresight of the forecasters μ. The grey 45 degree line $(y = x)$ shows the expected relationship between the foresight estimates of year 1 and 2 if foresight is 100% skill. The red line is the regression line from the first-order linear regression model including a grey 95% confidence band. As the regression line lies between the horizontal and 45 degree line, the foresight estimates regress toward the mean foresight of all forecasters from year 1 to year 2.

Figure 2. Regression of the Foresight Estimates Toward the Mean Foresight of all Forecasters From Year 1 to Year 2 Based on the Third-Order Linear Regression Model.



Note. The foresight estimates from the beta IRT model in year 1 and year 2 are standardized. The horizontal grey line indicates the mean foresight of the forecasters μ. The grey 45 degree line $(y = x)$ shows the expected relationship between the foresight estimates of year 1 and 2 if foresight is 100% skill. The red curve is the regression curve from the third-order linear regression model including a grey 95% confidence band. As the regression line lies between the horizontal and 45 degree line, the foresight estimates regress toward the mean foresight of all forecasters from year 1 to year 2, albeit extreme foresight estimates in year 1 appear to regress more strongly toward the mean.

Figure 3. Reversion of the Foresight Estimates Toward the Mean Foresight of All Forecasters From Year 1 to Year 2 for Different Portions of the Foresight Distribution



Note. The horizontal grey line indicates the mean foresight of the forecasters. The vertical grey line represents the cutoff *c* splitting the distribution of the foresight estimates in year 1 into an upper and lower portion. The red square denotes the mean foresight of the forecasters in year 1 and the red triangle indicates the mean foresight of the forecasters in year 2 in the corresponding portion of the foresight distribution in year 1. The difference between the square and triangle illustrates the absolute reversion toward the mean.

Table 1. Summary of Cognitive, Evolutionary, and Chance Theories of Foresight

	Cognitive Theories	Evolutionary Theories	Chance Theories
Nature of Foresight	Boundedly rational foresight of near, intermediate, and distant future states of the world	Myopic foresight of unique opportunities that can be serendipitously discovered	Foresight may not exist, accurate predictions of future states of the world are due to chance
Explanation for Foresight	Accurate mental representations of the world that enable strategists to predict future states of the world	Cognitive preadaptation (deep domain knowledge) to a future environment that enables strategists to see unique opportunities that other strategists cannot see once the future environment realizes	Random processes in the world determine future states of the world that are accurately or inaccurately predicted by strategists due to chance
Extent of Luck in Foresight	Somewhat luck but mainly skill as strategists can learn and adapt mental representations, albeit their bounded rationality and the inherent randomness of the world preclude perfect foresight	Mainly luck but also somewhat skill as strategists may deepen and broaden their domain knowledge, attend to changes in the environment, and exert effort to seize opportunities in the environment, albeit the strategist mostly cannot control the environment determining what deep domain knowledge becomes valuable	Mostly luck as strategists cannot control random processes and consequently, not predict future states of the world accurately over time

Table 2. Results of the First-Order, Second-Order, and Third-Order Polynomial Linear Regression Models and Their Comparison

	Model 1: Linear Effect				Model 2: Quadratic Effect				Model 3: Cubic Effect			
	Model Estimates											
Predictor	$\hat{\beta}$	*SE*	*t*	*p*	$\hat{\beta}$	*SE*	*t*	*p*	$\hat{\beta}$	*SE*	*t*	*p*
Intercept	0.00	0.05	0.00	1.000	-0.01	0.05	-0.24	0.814	-0.03	0.05	-0.65	0.517
Foresight Year 1	0.68	0.05	14.85	***	0.67	0.05	13.82	***	0.79	0.07	12.13	***
Foresight Year 1 Squared					0.01	0.02	0.51	0.612	0.05	0.03	1.90	†
Foresight Year 1 Cubed									-0.03	0.01	-2.75	**
Statistic	*Model Fit*											
R^2	0.46				0.46				0.47			
F	*F*(1, 263) = 220.50***				F(2, 262) = 110.10***				F(3, 261) = 77.74***			
AIC	595.66				597.40				591.85			
BIC	606.40				611.72				609.75			
LRT					$\chi^2(1) = 0.13, p = .608$				$\chi^2(1) = 4.03$**			

Note. Each regression model predicted the standardized foresight estimates from year 2 from the standardized foresight estimates from year 1 based on $n = 265$ observations. $\hat{\beta}$ = estimated regression coefficient; *SE* = standard error; *t* = *t*-statistic; *p* = *p*-value; R^2 = coefficient of determination quantifying the variation in the dependent variable that is explained by the independent variable(s); *F* = *F*-test indicating if the regression explains significant variation in the dependent variable; *AIC* = Akaike information criterion; *BIC* = Bayesian information criterion; *LRT* = likelihood-ratio test compares the goodness of fit of two nested models by testing the null hypothesis that the two models fit the data equally well. A significant likelihood-ratio test suggests that the simpler model describes the data significantly worse and therefore, the more complex model should be selected for interpretation. In this case, the likelihood-ratio tests compared Model 2 to Model 1 and Model 3 to Model 2. Robustness checks using robust standard errors and removing one outlier with high leverage in the data did not affect the interpretation of the results.

† $p < .10$
** $p < .01$
*** $p < .001$

Table 3. Reversion Toward the Mean Foresight of Forecasters for Different Portions of the Foresight Distribution

Portion	n	$\hat{\kappa}$	*LL*	*UL*	Sig
$x_i > 3$	4	0.38	0.11	0.70	*
$x_i > 2$	8	0.35	0.20	0.73	*
$x_i > 1$	33	0.30	0.10	0.48	*
$x_i > 0$	121	0.23	0.06	0.47	*
$x_i < 0$	144	0.23	0.07	0.41	*
$x_i < -1$	32	0.47	0.26	0.73	*
$x_i < -2$	2	0.54	0.31	0.82	*
$x_i < -3$	1	0.56			

Note. Portion = portion of the foresight distribution of the standardized foresight estimates in year 1 x_i for which the reversion toward the mean was calculated; n = number of forecasters in the portion of the distribution; $\hat{\kappa}$ = reversion toward the mean; LL = lower level of the 95% percentile bootstrap confidence interval based on 10,000 resamples; UL = upper level of the 95% percentile bootstrap confidence interval based on 10,000 resamples; Sig = statistical significance based on the 95% percentile bootstrap confidence interval, where * denotes statistical significance at $\alpha = .05$ if the 95% confidence interval does not include 0. The reversion toward the mean quantifies the extent to which foresight is luck in the corresponding portion of the foresight distribution.

APPENDIX A. INTRODUCTION TO ITEM RESPONSE THEORY MODELS

In the following, I will provide a general introduction to item response theory (IRT) models by describing two basic IRT models that illustrate the general structure of IRT models: The 1 and 2 Parameter Logistic models. In general, IRT models assume that the responses of persons to items can be described as a function of their latent abilities, the item characteristics, and a random factor. The 1 Parameter Logistic (1PL) and 2 Parameter Logistic (2PL) models represent basic IRT models for Bernoulli distributed responses (e.g., whether an item was correctly answered or not). Both models describe the responses of persons to a set of items as a probabilistic function of the latent abilities of persons, the difficulties of the items, and the discrimination of the items. Although other IRT models for such dichotomous data exist (Guttman, 1944; Lazarsfeld & Henry, 1968; Lord, 1952), logistic models have become the standard for this data type. More formally, the 1PL (or Rasch) model can be defined as follows (Bond & Fox, 2013; Rasch, 1980):

$$\begin{aligned} Y_{iq} &\sim Bernoulli\left(\pi_{iq}\right) \\ \pi_{iq} &= \frac{1}{1 + e^{-\lambda(\theta_i - \delta_q)}}, \end{aligned} \tag{A1}$$

where the response Y_{iq} of person $i = 1, \dots, N$ ($N \in \mathbb{Z}^+$ the number of persons) on item $q = 1, \dots, Q$ ($Q \in \mathbb{Z}^+$ the number of items) is either correct (1) or incorrect (0) and thus, follows a Bernoulli distribution with success probability π_{iq}; θ_i represents the latent ability of person i; δ_q is the difficulty of item q describing how difficult it is to answer an item correctly; and λ is the discrimination indicating how well the question differentiates between different degrees of the latent ability. By modeling the difficulty and discrimination of items, IRT models control for the influence of item characteristics on the responses of persons. The model ensures that the values lie within the range of the success probability $\pi_{iq} \in [0,1]$, which is the probability that person i solves item q correctly.

It is important to note that the 1PL model assumes the discrimination λ to be constant across all items. In other words, the model assumes all items to differentiate equally well between different degrees of the latent ability. However, items often differ in their discrimination, which can be modeled by the 2PL (or Birnbaum) model that can be formally defined as (Birnbaum, 1968):

$$\begin{aligned} Y_{iq} &\sim Bernoulli(\pi_{iq}) \\ \pi_{iq} &= \frac{1}{1 + e^{-\lambda_q(\theta_i - \delta_q)}}, \end{aligned} \tag{A2}$$

where the only difference to the 1PL model is that the discrimination λ_q varies across items q. Note that these definitions describe IRT models in the difficulty-discrimination parameterization.

The 2PL IRT model (and analogously the 1PL IRT model) can be rewritten as a multilevel model (Hox et al., 2017; Raudenbush & Bryk, 2002) with two levels, where the level-1 model can be formally defined as

$$\begin{aligned} Y_{iq} &\sim Bernoulli(\pi_{iq}) \\ logit(\pi_{iq}) &= \beta_{0q} + \lambda_q \theta_i, \end{aligned} \tag{3}$$

the level-2 model for the parameters varying across questions can be formally written as

$$\begin{aligned} \beta_{0q} &= \beta_{00} + u_{0q} \\ \lambda_q &= \lambda_0 + v_q, \end{aligned} \tag{4}$$

where the random effects pertaining to the level-2 model of the questions are assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{0q} \\ v_q \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{u0}^2 & \\ 0 & \tau_v^2 \end{pmatrix} \right), \tag{5}$$

and the level-2 model for the latent ability parameter varying across persons can be formally described with

$$\theta_i = \theta_0 + z_i, \tag{6}$$

where the random effect pertaining to the level-2 model of the latent ability parameter is assumed to be normally distributed with

$$z_i \sim \mathcal{N}(0, \tau_z^2). \tag{7}$$

In this multilevel intercept-slope parameterization, the probability of success π_{iq} is linked to the linear equation $\beta_{0q} + \lambda_q\theta_i$ by the logit link. The logit link ensures that the number space of $\pi_{iq} \in [0,1]$ matches the number space of $(\beta_{0q} + \lambda_q\theta_i) \in \mathbb{R}$. β_{0q} represents the easiness of item q, which is the logical equivalent of δ_q describing the difficulty of item q. Larger β_{0q} indicate a higher probability to answer item q correctly on average. The intercept-slope parameterization suggests that the item discrimination λ_q can be reinterpreted as the effect size of the latent ability θ_i on the probability of success π_{iq}. In other words, the item discrimination λ_q determines how useful the latent ability θ_i is for answering item q correctly. β_{00}, λ_0, and θ_0 refer to the means of the corresponding parameters across all items Q or all persons N, respectively. u_{0q}, v_q, and z_i represent the random effects of the corresponding model parameters describing how each item q or person i differs from its corresponding mean parameter. τ_{u0}^2, τ_v^2, and τ_z^2 describe the variances of the corresponding random effects.

1PL and 2PL IRT models typically assume the latent abilities θ to be unidimensional, which means that any systematic differences in the persons' responses to the items result from the same underlying ability (Hambleton & Swaminathan, 2013). If the latent abilities θ are unidimensional and the item responses are independent from each other conditional on the latent ability (i.e., the residuals of the items in the model with a unidimensional foresight parameter θ_i are uncorrelated), the items are assumed to be locally independent (Hambleton & Swaminathan, 2013). If the assumption of local independence is fulfilled, persons can be compared regarding

their latent ability even if they answered different item subsets as each item exclusively measures the latent ability. Furthermore, the parameters are typically assumed to be uncorrelated even though correlated parameters can improve the parameter estimates (Bürkner, 2021). Although IRT models have been most popular to model dichotomous responses, IRT models can be constructed for various response distributions, such as probabilistic forecasts (Ferrando, 2001; Mellenbergh, 1994; Müller, 1987; Samejima, 1972, 1973, 1974).

APPENDIX B. EVALUATION OF THE BETA ITEM RESPONSE THEORY MODEL OF PROBABILISTIC FORECASTS

In the following, I will evaluate the beta item response theory (IRT) model by comparing it to an alternative IRT model synthesized from prior research (Himmelstein et al., 2021, 2022; Merkle et al., 2016, 2017): the normal IRT model. Bayesian analyses based on 22,616 forecasts by 428 forecasters on 102 questions in a forecasting tournament suggest that the measurement of foresight by the beta IRT model is more objective, equally reliable, and more valid compared to the normal IRT model. The beta IRT model is more objective than the normal IRT model because it uses minimally informative prior distributions for the model parameters, whereas prior research has used informative priors for the parameters of the normal IRT model.[6] Post hoc analyses show that the largest improvement in validity results from modeling all forecasts made by forecasters. Overall, these results suggest that the beta IRT model of probabilistic forecasts estimates the foresight of persons more accurately than alternative IRT models.

The Normal Item Response Theory Model of Probabilistic Forecasts

In the following, I will describe the normal IRT model of probabilistic forecasts that represents a synthesis of the IRT models proposed by prior research (cf. Table B1).

Insert Table B1 about here

The normal IRT model describes the forecasters' probit-transformed probabilistic forecasts on the true response category as a probabilistic function of the forecasters' foresight, the difficulties and discriminations of questions, and the timing of the probabilistic forecasts.[7] The

[6] It is important to note that subjective priors may be desirable if they represent valid prior knowledge.

[7] The probit transformation ensures that the number space of the probabilistic forecasts $f \in [0,1]$ matches the number space of the normal distribution $\mathcal{N}(\mu, \sigma) \in \mathbb{R}$. Extreme probabilistic forecasts of 1 and 0 are transformed to 0.999 and 0.001 before the probit-transformation is applied.

normal IRT model can be written as a multilevel model with two levels, where the level-1 model can be formally defined as:

$$\begin{aligned} Y_{iq}^{*} &\sim Normal(\mu_{iq}, \sigma_q) \\ \mu_{iq} &= \beta_{\mu 0q} + \beta_{\mu 1q} t_{iq} + \lambda_q \theta_i \\ \sigma_q &= exp(\beta_{\sigma 0q}), \end{aligned} \tag{B1}$$

where the response Y_{iq}^{*} is the probit-transformed probabilistic forecast of forecaster $i = 1, \ldots, N$ ($N \in \mathbb{Z}^{+}$ the number of forecasters) on question $q = 1, \ldots, Q$ ($Q \in \mathbb{Z}^{+}$ the number of questions) that follows a Normal distribution with mean μ_{iq} and standard deviation σ_q; $\beta_{\mu 0q}$ refers to the easiness of question q; $\beta_{\mu 1q}$ represents a time effect capturing linear changes in the easiness of question q over time; t_{iq} captures the timing of the probit-transformed probabilistic forecast Y_{iq}^{*} of forecaster i on question q (i.e., when forecaster i made their forecast on question q) with $t_{iq} = 0$ indicating the first day question q was forecastable; and $\beta_{\sigma 0q}$ is the standard deviation of question q, where the natural exponential function ensures that the standard deviation σ_q is positive.

The level-2 model for the parameters varying across questions can be formally written as

$$\begin{aligned} \beta_{\mu 0q} &= \beta_{\mu 00} + u_{\mu 0q} \\ \beta_{\mu 1q} &= \beta_{\mu 10} + u_{\mu 1q} \\ \lambda_q &= \lambda_0 + v_q \\ \beta_{\sigma 0q} &= \beta_{\sigma 00} + u_{\sigma 0q}, \end{aligned} \tag{B2}$$

where the random effects pertaining to the level-2 model of the questions are assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{\mu 0q} \\ u_{\mu 1q} \\ v_q \\ u_{\sigma 0q} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{u\mu 0} & & & \\ 0 & \tau^2_{u\mu 1} & & \\ 0 & 0 & \tau^2_{v} & \\ 0 & 0 & 0 & \tau^2_{u\sigma 0} \end{pmatrix} \right) \tag{B3}$$

and the level-2 model for the latent ability parameter varying across persons can be formally described with

$$\theta_i = \theta_0 + z_i, \tag{B4}$$

where the random effect pertaining to the level-2 model of the latent ability parameter is assumed to be normally distributed with

$$z_i \sim \mathcal{N}(0, \tau^2_z), \tag{B5}$$

where τ^2_z is fixed to 1 to identify the model.

Although the normal IRT model of probabilistic forecasts represents an important advancement in the measurement of foresight, it is limited in four ways: First, it assumes the probabilistic forecasts to be normally distributed, which can result in misleading inferences (Gelman et al., 2013; Nott et al., 2024). Specifically, the normal IRT model may underestimate or overestimate the foresight of forecasters when they provide perfectly inaccurate forecasts (i.e., forecasts of 0 on the true category) or perfectly accurate forecasts (i.e., forecasts of 1 on the true category) as such extreme forecasts influence the foresight parameter estimates more strongly than other forecasts. This problem may be exacerbated by the fact that such extreme forecasts occur frequently in forecasting tournaments. For example, 11.85% of the 22,616 forecasts in the geopolitical data that I use to compare the normal to the beta IRT model are extreme forecasts. For this reason, research suggests that continuous response variables bounded in the interval $[0; 1]$ should be modeled as beta distributed as it naturally reflects the boundedness of the response variable and the parameters of the beta distribution have intuitive interpretations the data (Gupta

& Nadarajah, 2004; Smithson & Verkuilen, 2006; Verkuilen & Smithson, 2012).

Second, the normal IRT model assumes the parameters to be uncorrelated. Although this may be a reasonable assumption in some contexts, the model parameters estimated based on data from forecasting tournaments are likely correlated. For example, the easiness of the questions (i.e., the intercept of the mean model) should be negatively related to the time effect capturing changes in the easiness over time (i.e., the slope of the mean model) as the time effect should be smaller for questions that are easy from the very beginning. More generally, modeling correlations between the parameters may be preferable, because this often improves the model fit and usually results in more accurate parameter estimates by IRT models (Bürkner, 2021).

Third, the normal IRT model includes a time effect in the mean model but not in the variance model. Although mean model time effects are essential for capturing changes in the average forecasts over time, variance model time effects may convey additional insight into changes in the variation of the forecasts over time, which could improve the model fit and increase the accuracy of IRT model parameter estimates.

Fourth, the normal IRT model only considers the initial forecast on each question for each forecaster. While this makes the normal IRT model computationally cheaper, it also omits potentially valuable information on the foresight of forecasters contained in their updates over time (Atanasov et al., 2020; Kapoor & Wilde, 2023). This is especially problematic as forecasters frequently update their forecasts in forecasting tournaments. For example, the geopolitical data I use to compare the normal to the beta IRT model contain 7,276 forecast updates.

The beta IRT model of probabilistic forecasts makes better use of the available information in forecasting tournaments by modeling (1) the probabilistic forecasts as beta distributed, (2) correlations between the parameters, (3) time effects in the variance model, and (4) all forecasts

made by forecasters. One explanation for why prior research has used the simpler normal rather than the beta IRT model may be that the models were fitted using Bayesian Markov chain Monte Carlo (MCMC) Gibbs sampling (Geman & Geman, 1984) implemented in JAGS, which can struggle with complex models (Bürkner, 2017). For example, one comparatively complex IRT model proposed by prior research did not converge using this approach (Bo et al., 2017). To remedy this issue, I fitted the models using the Bayesian adaptive Hamiltonian Monte Carlo (HMC; Duane et al., 1987; Neal, 2011) and the No-U-Turn Sampler (NUTS; Hoffman & Gelman, 2014) algorithms implemented in *Stan* and *brms* (Bürkner, 2017, 2021). Next, I will compare the normal and beta IRT model in terms of objectivity, reliability, and validity based on data from a geopolitical forecasting tournament.

Data

The data from the geopolitical forecasting tournament consist of 22,616 probabilistic forecasts (15,340 initial forecasts and 7,276 updates) by 428 forecasters on 102 questions. The forecasts were made by forecasters in the control group in the first year of the tournament. Focusing on forecasters in the control group eliminates potential bias in the estimates resulting from treatments conducted during the forecasting tournament. Fitting the models to data from the first year of the tournament enables me to conduct predictive validity analyses with data from subsequent years of the geopolitical forecasting tournament. All questions were related to geopolitically relevant events. I refer the reader to prior research for more details on the forecasting tournament (see Atanasov et al., 2017; Mellers et al., 2015; Tetlock et al., 2014).

The mean and variance parameter estimates of the normal and beta IRT models are on different scales because they are linked to their corresponding distributional parameters by different link functions. While the mean parameter estimates of the normal IRT model are on the

identity scale, the mean parameter estimates of the beta IRT model are on the logit scale. Although the variance parameter estimates of the normal and beta IRT models are both on the log scale, the normal IRT model estimates the variance, whereas the beta IRT model estimates the precision of their corresponding response distributions. To ensure the comparability of the models despite these differences, I transformed the posterior samples of the model parameter estimates in the variance model of the normal IRT model by applying the $expit$ function to the corresponding posterior draws so that the scales of the normal IRT model parameters match the scales of the beta IRT model parameters. All mean parameter estimates are on the logit scale and all variance parameter estimates are precision estimates on the log scale.

Data Analysis

I fitted the normal and beta IRT models using the same procedure as outlined in the paper. To evaluate whether the beta IRT model represents an improvement over the normal IRT model, I compared the normal and beta IRT models on six dimensions. First, I plotted and correlated the foresight estimates of the normal and beta IRT models to evaluate the similarity of these parameter estimates between the models. Second, I conducted posterior-predictive checks (Box, 1980; Gelman et al., 1996, 2013, 2020) to assess the fit of the models to the data. Third, although I used the same approach to fit the normal and beta IRT models, I discuss the model fitting approach by previous research to contrast the objectivity of the normal and beta IRT models. Fourth, I compared the errors and width of the 95% credible intervals of the foresight parameter estimates by the normal and beta IRT models to evaluate differences in the measurement reliability of the models. Fifth, I correlated the foresight estimates by the normal and beta IRT models with the linear time-weighted Brier scores (TWBS; Schuler et al., 2022) to evaluate the criterion validity of the parameter estimates. Finally, I correlated the foresight estimates from the first year with the

foresight estimates from the second year of the geopolitical forecasting tournament to assess the predictive validity of the normal and beta IRT models. I used the same data analysis approach outlined above to estimate the foresight of 265 forecasters in the control group who also participated in the second year of the tournament and made 24,465 forecasts on 110 questions. As researchers and practitioners are generally interested in the relative accuracy of forecasters, correlations are appropriate to compare the beta to the normal IRT model. In the following, I will first briefly explain the model results before I compare the normal and beta IRT models.

Results

Mean and variance parameter estimates. Tables B2 and B3 show the fixed and random mean and variance parameter estimates of the normal and beta IRT models. For the sake of brevity, I focus on important aspects of the model results. The results suggest that the fixed mean and variance model parameter estimates of the normal and beta IRT models are similar. The intercept estimate of the precision $\hat{\beta}_{\phi 00}$ in the fixed variance parameter model is an exception, because it appears to be smaller for the normal compared to the beta IRT model. The results also suggest that the standard deviation and correlation estimates of the normal and beta IRT models are similar. The standard deviation estimate of the discrimination $\hat{\tau}_v$ is an exception, because it appears to be larger compared to the beta IRT model. While the normal IRT model assumes the parameters to be uncorrelated, four out of ten correlation estimates of the beta IRT model are statistically significant. This suggests that modeling correlations may be reasonable. Overall, the fixed and random mean and variance model parameter estimates of the normal and beta IRT models are generally similar.

Insert Table B2 about here

Insert Table B3 about here

Foresight parameter estimates. Figure B1 shows the posterior means and 95% credible intervals of the foresight parameter estimates by the normal and beta IRT models (sorted by size). Figure B2 shows the scatterplot of the foresight parameters estimates by the beta and normal IRT models. The figures suggest that the foresight parameter estimates by the normal and beta IRT models are similar. This is corroborated by the correlation between the foresight estimates $r = 0.94$ ($95\%\ CI = [0.93; 0.95], t = 57.79, p < 0.001$), which suggests that the rank order between the estimates is similar but not identical as the confidence interval of the correlation estimate does not contain 0 or 1. That is, the normal and beta IRT models appear to arrive at similar conclusions about the relative foresight of forecasters.

Insert Figure B1 about here

Insert Figure B2 about here

Model fit. Figure B3 depicts the graphical posterior predictive checks of the normal and beta IRT models. Graphical posterior predictive checks compare the observed forecasts y against 10 simulated datasets y_{rep} from the posterior predictive distribution to assess the model fit to the data (Box, 1980; Gelman et al., 1996, 2013, 2020). The posterior predictive checks suggest that the normal IRT model fit is rather bad, whereas the beta IRT model fit is good. Although the normal IRT model describes the probit-transformed forecasts in the interval $[-2.5, 2.5]$ relatively well, it inaccurately predicts probit-transformed forecasts outside this interval. This is potentially problematic as such extreme forecasts are particularly informative about the foresight of forecasters. In contrast, the beta IRT model accurately predicts the forecasts across the entire interval $(0,1)$. This suggests that the beta likelihood is a better choice than the normal likelihood.

Insert Figure B3 about here

Objectivity. Objectivity is an important quality criterion of research designs and statistical methods as it provides the foundation for the reliability, validity, and replicability of research findings. While Bayesian statistics is a powerful approach to fit IRT models, Bayesian models make assumptions about the model parameters in the form of prior probability distributions that can influence the model parameter estimates (Gelman et al., 2013). As such priors are inherently subjective (Gelman, 2008), Bayesian model parameter estimates may be considered subjective. Although setting priors should be considered a strength of Bayesian model estimation as this enables researchers and practitioners to incorporate prior knowledge explicitly (Gelman et al., 2013; McElreath, 2020), setting weakly informative priors for the normal and beta IRT models renders the model parameter estimates virtually objective. However, prior research has set informative priors for the intercept and time effect in the mean parameter model of the normal IRT model (e.g., $\mathrm{N}(0, 0.5)$; Merkle et al., 2016). Therefore, the foresight parameter estimates of the beta IRT model can be considered more objective than those of prior normal IRT models.

Reliability. The errors and 95% credible intervals of the foresight parameter posterior distributions estimated by the normal and beta IRT models represent measures of reliability that quantify the uncertainty pertaining to each individual foresight parameter estimate (see Figure B1).[8] Larger errors and wider credible intervals indicate a lower reliability of (i.e., higher

[8] IRT models relying on frequentist parameter estimation use the concept of item and test information to compute the conditional standard error of measurement for the foresight estimates (Hambleton & Swaminathan, 2013). The conditional standard error of measurement describes reliability as a function of the foresight parameter. This enables researchers and practitioners to evaluate how reliably forecasting tournaments measure different degrees of foresight. For example, forecasting tournaments may measure average foresight accurately but above-average foresight inaccurately. However, the conditional standard error of measurement assumes that every person forecasts every question, which is usually not the case in forecasting tournaments as forecasters usually select a subset of questions to forecast (Merkle et al., 2017). In contrast, IRT models relying on Bayesian parameter estimation do not make this assumption because they provide individual reliability estimates for each forecaster's foresight parameter estimate. This can be considered an advantage of Bayesian over frequentist model parameter estimation.

uncertainty about) the foresight parameter estimates. To evaluate the reliability of the foresight parameter estimates, I compared the average errors and the average width of the 95% credible intervals of the normal and beta IRT models. The average error was 0.50 for the normal and 0.49 for the beta IRT model. The average width of the 95% credible intervals was 1.97 for the normal and 1.92 for the beta IRT model. Unequal variances *t*-tests (linearity, independence, and normality assumptions were met) indicated that the average errors ($b = 0.01, SE = 0.01, t = 1.06, p = 0.290$) and the average width of the 95% credible intervals ($b = 0.05, SE = 0.05, t = 1.02, p = 0.307$) did not differ significantly between the normal and beta IRT model. This suggests that the normal and beta IRT models measure foresight equally reliably.

Validity. I compared the validity of the normal and beta IRT models based on the predictive validity of the normal and beta IRT models using correlations between the foresight parameter estimates in year 1 and 2 of the geopolitical forecasting tournament. Using data from year 1 and 2 maximized the sample size of forecasters in the control group who participated in two years of the forecasting tournament. The data from year 2 consisted of 265 forecasters who made 24,465 forecasts (15,797 initial forecasts and 8,668 updates) on 110 questions. The foresight parameter estimates from year 1 and 2 were positively correlated for the normal ($r = 0.59, 95\%\, CI = [0.51;\ 0.67], t(263) = 11.94, p < 0.001$) and the beta IRT model ($r = 0.68, 95\%\, CI = [0.60;\ 0.74], t(263) = 14.85, p < 0.001$), see also Figure B4. The predictive validity of the beta is greater than the predictive validity of the normal IRT model $\left(\hat{\delta} = 0.14, 95\%\ \text{BCa}\ CI = [0.03, 0.24]\right)$. These results suggest that the beta IRT model has a higher predictive validity than the normal IRT model.

Insert Figure B4 about here

Post Hoc Analyses

To understand why the beta IRT model outperforms the normal IRT model in terms of predictive validity, I compared the predictive validities of different normal and beta IRT model specifications. Remember that the beta extends the normal IRT model by modeling (1) the probabilistic forecasts as beta distributed, (2) correlations between the parameters, (3) time effects in the variance model, and (4) all forecasts made by the participants in forecasting tournaments. Therefore, I compared different specifications of the normal and beta IRT models in terms of their predictive validity to assess what extensions improve the predictive validity. All models were fitted in *brms* with the same settings as described above. Table B4 shows the results of the post hoc analyses. The results suggest that modeling all forecasts made by the participants in forecasting tournaments or modeling correlations and time effects jointly improves the predictive validity of the normal IRT model so that it does not differ significantly from the beta IRT model anymore. Strikingly, the misspecification of the likelihood (i.e., modeling the probabilistic forecasts as normal instead of beta distributed) does not seem to impair the predictive validity. However, using the beta instead of the normal likelihood offers the advantage that the probabilistic forecasts are modeled and predicted on the probability scale. Overall, these results suggest that the beta outperforms the normal IRT model mainly because it considers all forecasts, albeit modeling correlations and time effects may somewhat contribute to this performance.

Insert Table B4 about here

Conclusion

This appendix compared the beta IRT model to the beta IRT model of probability forecasts can be considered more objective than the normal IRT model, which has been used by prior research to measure foresight (Merkle et al., 2016, 2017). Prior research has specified informative

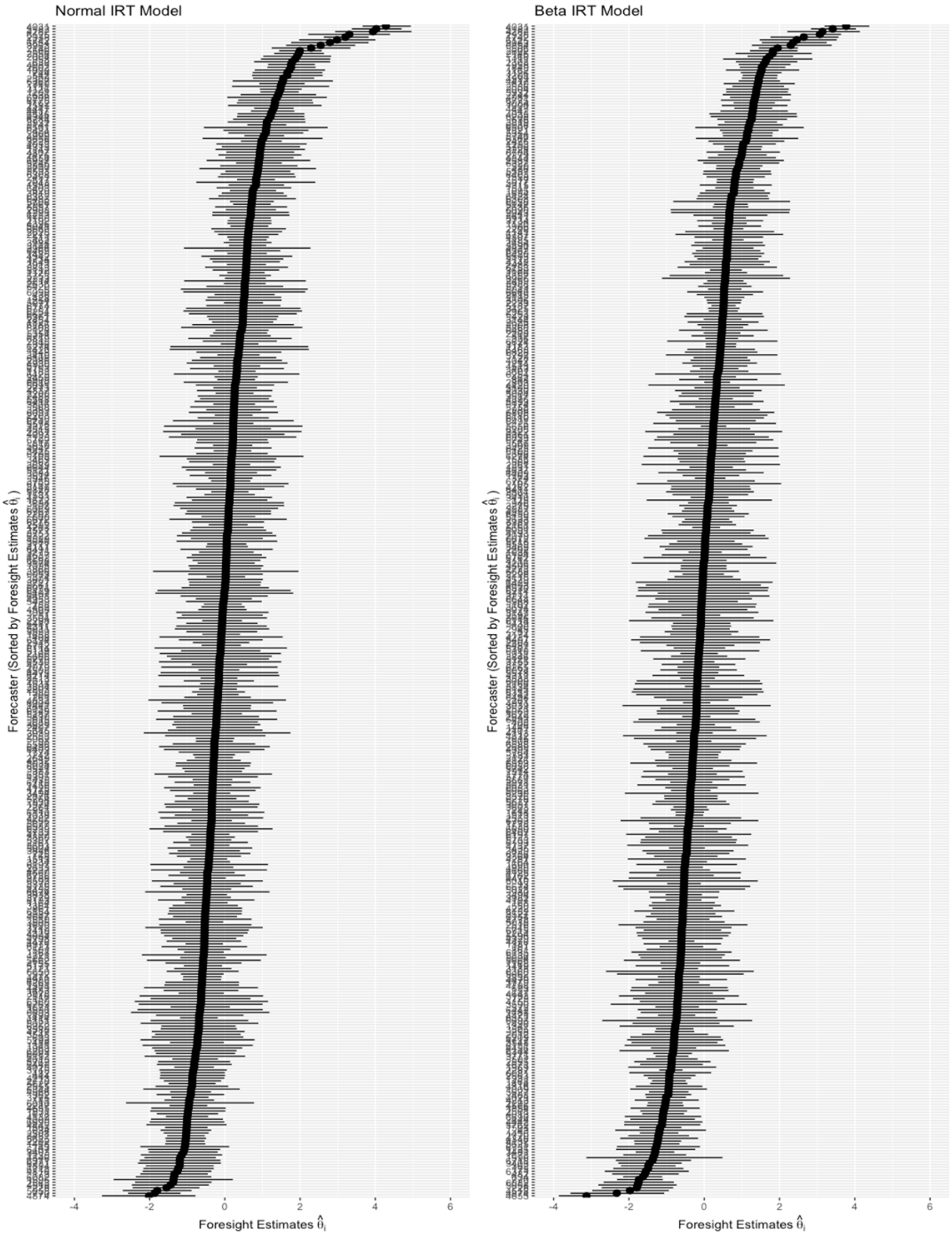
priors for some normal IRT model parameters (Merkle et al., 2016, 2017), whereas I specified uninformative priors for all normal and beta IRT model parameters. One explanation for why prior research set informative priors for some normal IRT model parameters may be that the normal IRT model was fitted using MCMC Gibbs sampling (Geman & Geman, 1984), which can struggle with complex models like the normal IRT model. In contrast, I fitted the normal and beta IRT models using HMC NUTS sampling (Duane et al., 1987; Hoffman & Gelman, 2014; Neal, 2011), which works well with complex models (Bürkner, 2017, 2021). Although the normal and beta IRT models can in principle both be fitted using uninformative priors as demonstrated in this paper, the beta IRT model and the Bayesian model fitting procedure can still be considered an improvement in objectivity compared to previous research.

The normal and beta IRT models are equally reliable as their average estimated errors and average 95% credible intervals are not significantly different. This suggests that the model specification may not affect the reliability of the foresight parameter estimates. One explanation for this finding may be that the reliability of the foresight parameter estimates depends on (1) how many forecasts each forecaster made, (2) the number of questions each forecaster forecast, and (3) what questions each forecaster forecast. The foresight parameter estimates are likely accurate if forecasters make many forecasts on many questions. If forecasters only forecast a subset of questions, the foresight parameter estimates are likely accurate if forecasters forecast questions with high discriminations and difficulties matching their latent ability. More generally, this highlights that researchers and practitioners should incentivize forecasters to regularly forecast all questions in a forecasting tournament to ensure the reliable measurement of foresight.

The beta IRT model outperforms the normal IRT model in terms of predictive validity. The predictive validity can be considered the most important validity facet because IRT models of

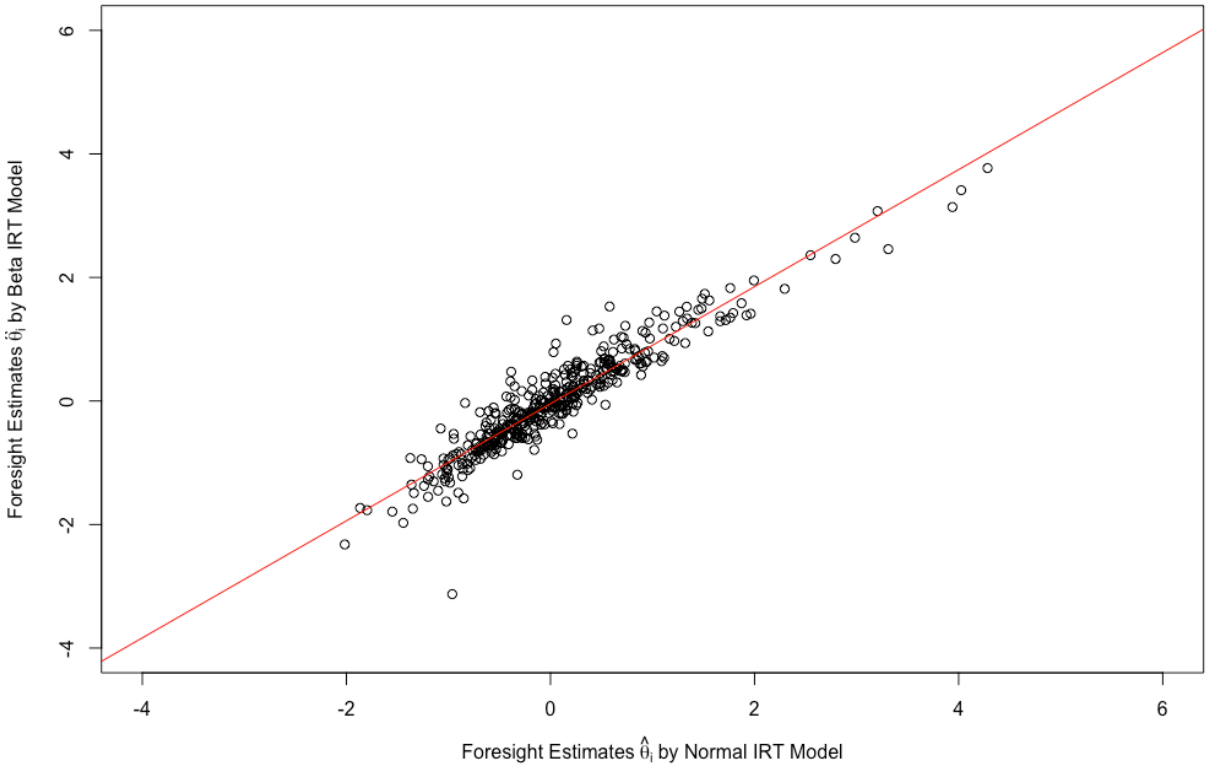
probability judgments are designed to identify forecasters with superior foresight (Mellers et al., 2015; Tetlock & Gardner, 2016). Conceptualizing foresight as an ability (Schuler et al., 2024) implies that any measurement of foresight should be stable over time. The higher predictive validity of the beta IRT model represents an improvement in validity over previous IRT models. Taken together, the results suggest that the beta IRT model of probabilistic forecasts estimates the foresight of persons more accurately than the normal IRT model.

Figure B1. Posterior Means and 95% Credible Intervals of the Foresight Parameter Estimates by the Normal and Beta IRT Models (Sorted by Size)



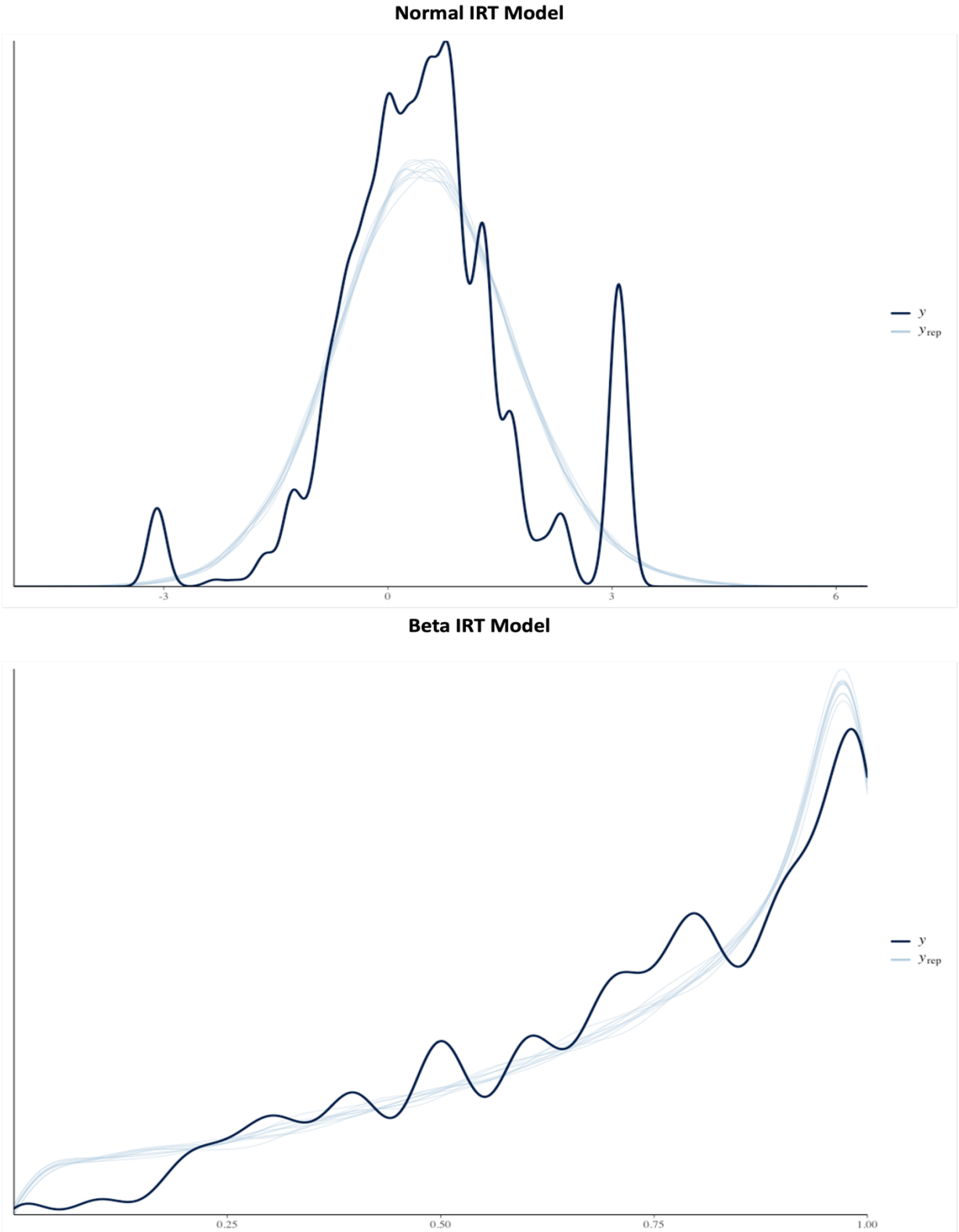
Note. The black dots are the posterior means. The grey bars are the 95% credible intervals. Wider credible intervals reflect more uncertainty about the foresight parameter estimates.

Figure B2. Scatterplot of the Foresight Parameter Estimates by the Beta and Normal Item Response Theory Models



Note. The red regression line is based on the simple linear regression of $\hat{\theta}_i^{beta}$ on $\hat{\theta}_i^{normal}$.

Figure B3. Graphical Posterior Predictive Checks of The Normal and Beta IRT Models



Note. y = observed forecasts; y_{rep} = 10 simulated datasets from the posterior predictive distributions.

Figure B4. Scatterplots of the Foresight Parameters from Years 1 and 2 as Estimated by the Beta and Normal Item Response Theory Models

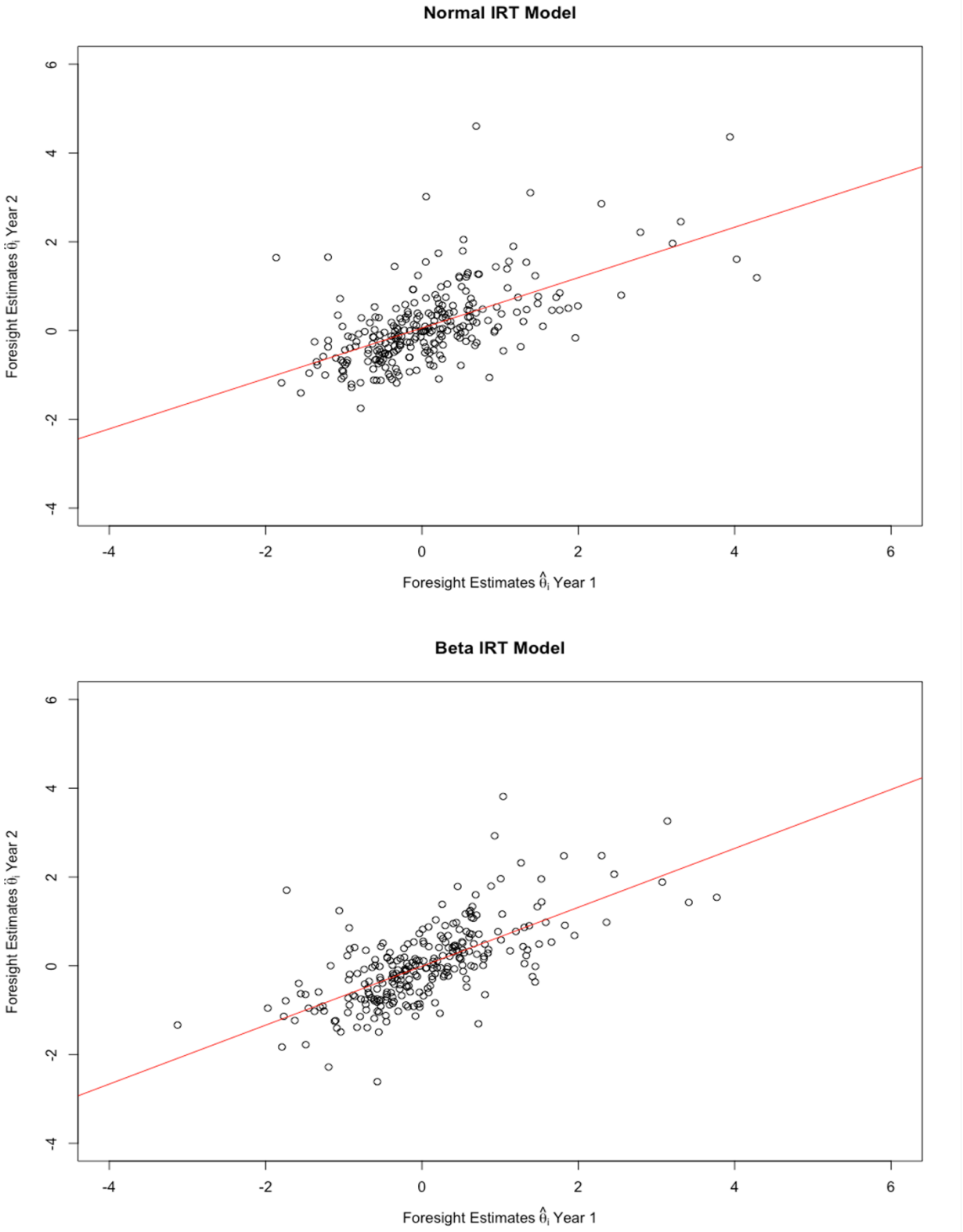


Table B1. IRT Models of Probabilistic Forecasts Proposed by Prior Research

	Citation	Likelihood	Correlations	Time Effects	Data	Priors
1	Noel & Dauvier (2007)	Beta	No	Mean: No Variance: No	Initial forecasts	—
2	Bo, Budescu, Lewis, Tetlock, & Mellers (2017)	Multinomial	No	Mean: Yes Variance: No	Scores	Less informative: $\mathcal{N}(0, 25)$
3	Merkle, Steyvers, Mellers, & Tetlock (2016)	Normal	No	Mean: Yes Variance: No	Initial forecasts	Informative: $\mathcal{N}(0, 0.5)$
4	Merkle, Steyvers, Mellers, & Tetlock (2017)	Normal	No	Mean: Yes Variance: No	Initial forecasts	Informative: $\mathcal{N}(0, 0.5)$
5	Himmelstein, Atanasov, & Budescu (2021)	Normal	No	Mean: Yes Variance: No	Initial forecasts	Not reported
6	Himmelstein, Budescu, & Han (2022)	Normal	No	Mean: Yes Variance: No	Initial forecasts	Not reported
7	Normal IRT Model	Normal	No	Mean: Yes Variance: No	Initial forecasts	Minimally informative
8	Beta IRT Model	Beta	Yes	Mean: Yes Variance: Yes	All Forecasts	Minimally informative

Note. Model 1 is a frequentist IRT model that uses a joint maximum likelihood procedure to estimate the parameters; model 2 did not converge as indicated by the potential scale reduction factor $\hat{R}$, which ranged between 1.1 and 1.5 (Bo, Budescu, Lewis, Tetlock, & Mellers, 2017: 95); models 5 and 6 essentially follow the structure of models 3 and 4 but do not report the specification of the variance model, the priors, or convergence diagnostics; model 7 is defined in equations B1-B5 in this appendix, and model 8 is described in equations 1–5 in the paper; models 2, 3, and 4 were fitted using MCMC Gibbs sampling; models 5, 6, 7, and 8 were fitted using HMC NUTS sampling. The column Time Effects indicates whether the IRT model specified time effects in the mean or variance model.

Table B2. Fixed Mean and Variance Parameter Estimates of the Normal and Beta IRT Models

	Normal IRT Model				Beta IRT Model			
	Est	Error	LL	UL	Est	Error	LL	UL
Mean Parameters								
$\beta_{\mu 00}$	0.38	0.08	0.23	0.53	0.23	0.06	0.12	0.34
$\beta_{\mu 10}$	0.02	0.00	0.02	0.03	0.02	0.00	0.02	0.03
λ_0	0.34	0.05	0.25	0.43	0.24	0.03	0.19	0.29
Precision Parameters								
$\beta_{\phi 00}$	0.37	0.01	0.35	0.39	0.92	0.04	0.85	1.00
$\beta_{\phi 10}$					0.00	0.00	0.00	0.00

Note. Est = estimate; Error = estimated error based on the standard deviation of the posterior distribution; LL = lower limit of the 95% credibility interval based on the 2.5% quantile of the posterior distribution; UL = upper limit of the 95% credibility interval based on the 97.5% quantile of the posterior distribution. The mean parameter estimates are on the logit scale. The precision parameter estimates are on the log scale. The intercept of the variance model $\beta_{\phi 00}$ of the normal IRT model is based on transformed posterior draws of $\beta_{\sigma 00}$ using the $expit$ function.

Table B3. Standard Deviation and Correlation Estimates of the Normal and Beta IRT Models

	$u_{\mu 0q}$	$u_{\mu 1q}$	v_q	$u_{\phi 0q}$	$u_{\phi 1q}$
$u_{\mu 0q}$	0.69 (0.05) [0.59; 0.80] 0.52 (0.04) [0.45; 0.60]	0	0	0	—
$u_{\mu 1q}$	0.16 (0.11) [-0.06; 0.36]	0.03 (0.00) [0.02; 0.03] 0.02 (0.00) [0.02; 0.03]	0	0	—
v_q	0.57 (0.08) [0.39; 0.71]	0.14 (0.12) [-0.09; 0.37]	6.53 (1.31) [4.70; 9.77] 2.52 (0.26) [2.09; 3.11]	0	—
$u_{\phi 0q}$	0.20 (0.10) [-0.01; 0.39]	0.02 (0.12) [-0.21; 0.24]	0.24 (0.11) [0.02; 0.44]	0.26 (0.01) [0.25; 0.27] 0.35 (0.03) [0.30; 0.41]	—
$u_{\phi 1q}$	0.35 (0.11) [0.12; 0.55]	0.12 (0.18) [-0.24; 0.46]	0.40 (0.12) [0.16; 0.61]	-0.15 (0.13) [-0.40; 0.11]	— 0.01 (0.00) [0.01; 0.01]

Note. The upper / lower part of the diagonal shows the standard deviation estimates $\hat{\tau}$ of the normal / beta IRT model. The standard deviation of the foresight parameter τ_z was fixed to 1 to identify the models. The upper triangular matrix contains zeros, because the normal IRT model assumes the parameters to be uncorrelated. The lower triangular matrix contains the correlation estimates of the beta IRT model. Estimates in parentheses represent the estimated errors based on the standard deviation of the posterior distributions. Estimates in brackets show the 95% credible intervals based on the 2.5% and 97.5% quantiles of the posterior distributions. The estimates of the mean model parameters are on the logit scale and the estimates of the variance model parameters are on the log scale. The standard deviation estimate of the variance model intercept in the normal IRT model $\hat{\tau}_{u\phi 0}$ is based on the transformed posterior draws of $\hat{\tau}_{u\sigma 0}$ using the $expit$ function.

Table B4. Post Hoc Analyses Comparing the Predictive Validity of the Normal and Beta IRT Models Based on Different Model Specifications of the Normal IRT Model

Normal Model Specification			Predictive Validity				Difference to Beta Model		
Cor	Time	Updates	r	LL	UL	$t(263)$	$\hat{\delta}$	LL	UL
No	No	No	0.59	0.51	0.67	11.94	0.14*	0.03	0.24
Yes	No	No	0.60	0.52	0.68	12.31	0.12*	0.02	0.22
No	Yes	No	0.61	0.52	0.68	12.35	0.12*	0.01	0.22
No	No	Yes	0.67	0.60	0.74	14.82	0.00	-0.03	0.08
Yes	Yes	No	0.62	0.54	0.69	12.75	0.10	-0.01	0.20
Yes	No	Yes	0.68	0.61	0.74	15.06	-0.01	-0.04	0.08
No	Yes	Yes	0.68	0.61	0.74	15.22	-0.02	-0.04	0.07
Yes	Yes	Yes	0.70	0.63	0.76	15.82	-0.04	-0.07	0.01

Note. Cor = correlation between model parameters estimated; Time = fixed and random time effects in the variance parameter model specified; Updates = models fitted to initial forecasts and updates (Yes) or only to initial forecasts (No); LL = lower limit of the 95% confidence interval; UL = upper limit of the 95% confidence interval; $\hat{\delta}$ = estimated difference between beta and normal IRT model. * indicates statistical significance at $\alpha = .05$ if the 95% confidence interval does not include 0. LL and UL of $\hat{\delta}$ represent BCa bootstrap confidence intervals based on 10,000 replications. The model specification in the first row refers to the baseline normal IRT model.