3 Individual Foresight: Concept, Operationalization, and Correlates

Benedikt Alexander Schuler[a]*
Johann Peter Murmann[b]
Marie Beisemann[c]
Ville Satopää[d]

[a] University of St.Gallen, Dufourstrasse 40a, 9000 St.Gallen, Switzerland, benedikt.schuler@unisg.ch
[b] University of St.Gallen, Dufourstrasse 40a, 9000 St.Gallen, Switzerland, j.peter.murmann@unisg.ch
[c] TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, Germany, beisemann@statistik.tu-dortmund.de
[d] INSEAD, Boulevard de Constance, 77305 Fontainebleau Cedex, France, ville.satopaa@insead.edu
* Corresponding author

Abstract

Judgmental forecasting research on superforecasters has demonstrated that individuals differ in their foresight. However, the concept underlying this work focuses on accuracy and does not fully incorporate the time dimension of foresight. We reconceptualize foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. To operationalize foresight in forecasting tournaments, we propose various strictly proper scoring rules and compare them with existing scoring rules using a simulation study and real-world forecasting data consisting of 414,168 scores for 9,694 forecasters on 498 questions from a four-year geopolitical forecasting tournament. The results suggest that the linear time-weighted Brier score should be the default operationalization of foresight and that probability training and teaming interventions as proposed by prior research may not improve foresight as we conceptualize it. We contribute to judgmental forecasting research by clarifying the concept, operationalization, and correlates of foresight.

Keywords: foresight, time-weighted Brier score, forecasting framework, forecasting tournaments, judgmental forecasting, superforecasting

3.1 Introduction

Forecasting is key to successful managerial decision-making as the utility of a decision alternative usually depends on the future state of the world. Managers can draw on statistical and judgmental forecasting techniques to forecast future states of the world (Armstrong, 2001; Bunn & Wright, 1991; Goodwin & Wright, 1993, 2014; Petropoulos et al., 2022): Statistical forecasting techniques employ statistical models, whereas judgmental forecasting techniques rely on human judgment to predict the future. Initial research suggested that statistical forecasting is superior to judgmental forecasting (e.g., Hogarth and Makridakis 1981, Makridakis 1986) because human judgment can be biased and noisy (Kahneman, 2011; Kahneman et al., 2021; Kahneman & Lovallo, 1993; Tversky & Kahneman, 1974). But later research using forecasting tournaments (Atanasov et al., 2017; Tetlock et al., 2014) demonstrated that some forecasters are superforecasters who forecast the future more accurately than others (Mellers et al., 2014; Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2016). This research has demonstrated that individuals differ in their foresight—their ability to forecast future states of the world.[9]

However, the concept of foresight underlying this research stream focuses on accuracy and does not fully incorporate the time dimension of foresight. Prior research in this domain has solely examined 'forecasting accuracy' as the focal phenomenon (see Table 3.1). Accordingly, prior research has operationalized foresight using scoring rules that mainly evaluate the accuracy of forecasters' predictions in forecasting tournaments. The seminal research, for example, used Brier scores (BS; Brier, 1950) to operationalize foresight and classify the top 2% of forecasters as superforecasters (Mellers et al., 2014; Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2016). The BS considers the accuracy but omits the timing of forecasts by computing the simple average of the forecast errors over the days a question was forecastable starting from the initial predictions made by the forecasters. This enables forecasters to score better on the BS by delaying their predictions until more and higher quality signals become available, which makes forecasting easier over time.

[9] Building on prior research (Lawrence et al., 2006; Petropoulos et al., 2022), we consider judgmental forecasting to be a method for predicting future states of the world, where forecasters make predictions based on their causal mental models of the world and the information they possess. In contrast, we construe foresight as a concept describing an individual ability necessary for successful judgmental forecasting. Although crowds can also possess foresight (Atanasov et al., 2017, 2024; Satopää et al., 2023; Surowiecki, 2005), we always mean the foresight of individuals when we refer to foresight in this paper. Note that our conceptualization of foresight is in principle generalizable to the foresight of crowds.

To remedy this issue, three studies used the mean daily Brier score (MDBS) to operationalize foresight (Atanasov et al., 2017, 2020; Mellers, Stone, Atanasov, et al., 2015). The MDBS considers the accuracy and timing of forecasts by imputing the daily average group forecast if forecasters delayed their initial predictions and computing the simple average of the forecast errors over all days a question was forecastable. But forecasters can still benefit from delaying their predictions as the group average usually improves when more and higher quality signals become available over time. Furthermore, both the BS and the MDBS weight forecast errors equally over time even though forecast errors should be weighted progressively heavier: If it becomes easier to forecast future states of the world over time, later errors under easier conditions indicate worse foresight compared to early errors under more difficult conditions. As prior research suggests that the availability and quality of signals increase over time (Moore et al., 2017; Satopää et al., 2021), the current concept and operationalization do not fully reflect the time dimension of foresight. This limits our ability to observe and explain interindividual differences in forecasters' foresight. For these reasons, the goal in this paper is to develop a concept and operationalization that fully integrate the accuracy and time dimension of foresight.

To provide the theoretical basis of this integration, we propose a forecasting framework suggesting that predicting future states of the world accurately becomes easier over time as the availability and quality of signals generally increase over time. A signal represents information with predictive value that enables forecasters to predict future states of the world more accurately. Based on this forecasting framework, we conceptualize foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. Using this concept, we propose various strictly proper scoring rules to operationalize foresight in forecasting tournaments (Atanasov et al., 2017; Tetlock et al., 2014). After proving analytically that the proposed scoring rules are strictly proper (Gneiting & Raftery, 2007), we evaluate the scoring rules under different theoretical signal trajectories in a simulation study. The results show that scoring rules considering the accuracy and timing of forecasts accurately measure the true foresight of forecasters irrespective of the true signal trajectory.

We also evaluate the scoring rules empirically using real-world forecasting data of 414,168 scores for 9,694 forecasters on 498 questions from a four-year geopolitical forecasting tournament collected in the context of government intelligence analysis. Measures of linear association suggest that scoring rules considering the accuracy and timing of forecasts lead to different conclusions about the relative foresight of forecasters and exhibit different correlational patterns with 29 variables compared to the BS. These results indicate that the different classes of scoring rules measure different concepts of foresight. As the linear time-weighted Brier score (TWBS) assigns better scores to superforecasters than other scoring rules, the results suggest that the linear TWBS should be the default operationalization of foresight. The results also show that probability training and teaming interventions as proposed by prior research may not improve foresight as we conceptualize it.

We make three contributions to the literature. First, we conceptualize foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. In contrast to prior work, our conceptualization fully integrates the accuracy and time dimension of foresight. Second, we introduce and evaluate a set of scoring rules considering the accuracy and timing of forecasts. Our results suggest that the linear TWBS should be the default operationalization of foresight, which enables researchers and practitioners to better identify individuals who possess superior foresight. Third, we establish the correlational patterns for the different concepts and operationalizations of foresight. This provides the foundation for future research on foresight.

The remainder of this article is structured as follows: First, we briefly review previous work on foresight. Second, we delineate the forecasting framework and define foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. Third, we define the scoring rules to operationalize our concept of foresight in forecasting tournaments. Fourth, we demonstrate in a simulation study that scoring rules considering the accuracy and timing of forecasts measure the true foresight of forecasters better than the BS under all theoretical signal trajectories. Fifth, we apply the scoring rules to real-world geopolitical forecasting data and show that they lead to different conclusions about the foresight of forecasters, exhibit different correlational patterns with variables previously examined in relation to forecasting ability, and that the linear TWBS is the preferable scoring rule. Finally, we discuss the theoretical and practical implications of our work, consider the limitations of our study, and outline opportunities for future research.

Table 3.1 Overview of the Concept, Measurement, and Operationalization of Foresight in International Peer-Reviewed Journal and Conference Publications

Study	**Concept**	**Measurement**	**Operationalization**
Atanasov et al. (2017)	Accuracy	Accuracy and Timing	Mean Daily Brier Score
Atanasov et al. (2020)	Accuracy	Accuracy and Timing	Mean Daily Brier Score
Baron et al. (2014)	Accuracy	Accuracy	Brier Score
Chang et al. (2016)	Accuracy	Accuracy	Brier Score
Dana et al. (2019)	Accuracy	Accuracy	Brier Score
Friedman et al. (2018)	Accuracy	Accuracy	Brier Score
Himmelstein et al. (2021)	Accuracy	Accuracy	Brier Score
Horowitz et al. (2019)	Accuracy	Accuracy	Brier Score
Karvetski et al. (2022)	Accuracy	Accuracy	Brier Score
Mellers & Tetlock (2019)	Accuracy	Accuracy	Brier Score
Mellers et al. (2014)	Accuracy	Accuracy	Brier Score
Mellers et al. (2017)	Accuracy	Accuracy	Brier Score
Mellers et al. (2019)	Accuracy	Accuracy	Brier Score
Mellers et al. (2023)	Accuracy	Accuracy	Brier Score
Mellers, Stone, Atanasov, et al. (2015)	Accuracy	Accuracy and Timing	Mean Daily Brier Score
Mellers, Stone, Murray, et al. (2015)	Accuracy	Accuracy	Brier Score
Moore et al. (2017)	Accuracy	Accuracy	Brier Score
Satopää et al. (2014)	Accuracy	Accuracy	Brier Score
Satopää et al. (2021)	Accuracy	Accuracy	Brier Score
Tetlock & Mellers (2014)	Accuracy	Accuracy	Brier Score
Tetlock et al. (2014)	Accuracy	Accuracy	Brier Score
Ungar et al. (2012)	Accuracy	Accuracy	Brier Score

Note. Although some studies operationalized foresight using the Mean Daily Brier Score, which de facto partially measures the timing of forecasts, these publications conceptualize foresight only in terms of accuracy.

3.2 Foresight

Judgmental forecasting research on superforecasting has demonstrated that forecasters differ in their foresight—their ability to forecast future states of the world accurately (Mellers et al., 2014; Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2016). In forecasting tournaments (Atanasov et al., 2017; Tetlock et al., 2014), the top 2% of forecasters—called superforecasters—were able to forecast future states of the world more accurately than the other forecasters across a variety of geopolitical forecasting questions three years in a row (Mellers, Stone, Murray, et al., 2015). Forecasting tournaments are prediction polls in which participants make probability forecasts on questions about the future state of the world at a specific point in time (Atanasov et al., 2017; Tetlock et al., 2014). As forecasters can update their probability forecasts over time until the question is resolved, forecasters can integrate new information in their forecasts. Although accurately forecasting future states of the world may partly be luck (Barney, 1986; Mauboussin, 2012), the superforecasters' predictive accuracy across multiple questions suggests that foresight is also an ability in which forecasters display stable interindividual differences.

Subsequent research has examined the aggregate profile of superforecasters to understand why they forecast future states of the world more accurately than others. Various personality traits, situational variables, and behaviors are related to predictive accuracy (Atanasov et al., 2020; Atanasov & Himmelstein, 2022; Karvetski et al., 2022; Mellers et al., 2014; Mellers, Stone, Atanasov, et al., 2015; Mellers, Stone, Murray, et al., 2015): Forecasters with above average fluid and crystallized intelligence, need for cognition, open-minded thinking, cognitive control, dialectical complexity, and competitiveness predicted the future more accurately than others. Furthermore, forecasters who participated in cognitive debiasing training, became members of teams, or were tracked into elite teams predicted the future more accurately than others. Moreover, forecasters who cultivated a scientific worldview, embraced a growth mindset, developed a nuanced understanding of uncertainty, collected more information on the questions to be forecast, and updated their beliefs and forecasts regularly in small steps predicted the future more accurately than others. Generally, good calibration (Moore et al., 2017) as well as little bias and noise in the interpretation of information (Satopää et al., 2021) are elements of foresight.

Previous research has implicitly acknowledged that the availability of signals may increase over time. For example, research has plotted the information revelation over time (Arrow et al., 2008; Himmelstein et al., 2021, 2022) and suggested that an increase in available signals results in a higher predictive accuracy of forecasters (Moore et al., 2017). Furthermore, research building on the partial information framework (Satopää et al., 2016) contended that some signals about the future state of the world may become available later, which can confine forecasters' predictive accuracy in the beginning (Satopää et al., 2021). Although this work overall acknowledges that the accuracy of forecasts is bounded by the available signals at a specific point in time, neither of these studies fully conceptualizes foresight as an ability that is constituted by the accuracy and timing of forecasts.

3.3 Reconceptualizing Foresight

3.3.1 Forecasting Framework

The extent to which future states of the world can be forecast accurately depends on whether the world is assumed to be a deterministic system—a system in which the initial states and causal mechanisms determine all future states of the system—or an indeterministic system—a system in which future states of the system are at least partly random (Lorenz, 1969). We take an indeterministic view of the world by assuming future states of the world to be partly determined by causal mechanisms operating in the world and partly created by stochastic processes. Although stochastic processes preclude a perfectly predictable world, forecasters can imperfectly forecast future states of the world by observing the present state of the world and inferring the causal mechanisms operating in the world that allow them to project how the present state of the world could evolve in the future (Lorenz, 1969; Simon, 1996). Forecasters observe the present state of the world by gathering information, for example, from their personal experiences, other individuals, or media. Forecasters infer the causal mechanisms operating in the world by observing changes in states of the world over time and interpreting these changes through the conception of causal mechanisms that can explain these changes (Cheng, 1997). In short, forecasters forecast future states of the world based on causal mental models of the world and information about the present state of the world.

The entirety of information about the present state of the world and the causal processes operating in the world that is available at a specific point in time constitutes the information universe (Satopää et al., 2016). Forecasters predict future states of the world by sampling information from this information universe at a specific point in time (Satopää et al., 2016). The information universe dynamically expands over time due to the continuous interaction of the present state of the world with the causal mechanisms and stochastic processes operating in the world, which produces more and higher quality information over time. This reduces epistemic uncertainty—imperfect knowledge (Tannenbaum et al., 2017)—over time.

The increasing quantity and quality of signals make it easier to forecast future states of the world accurately over time for two reasons. First, forecasters can observe the present state of the world more accurately by updating their initial observation based on the higher quantity and quality of signals created by causal mechanisms and stochastic processes over time (Atanasov et al., 2020; Kapoor & Wilde, 2023). Second, forecasters can infer the relevant causal mechanisms operating in the world more accurately by evaluating the conceived causal mechanisms in light of the higher quantity and quality of signals in a Bayesian manner (Harrison & Stevens, 1976; West & Harrison, 2006). As forecasters' predictive accuracy depends partly on the accurate observation of the present state of the world and partly on their accurate knowledge of the relevant causal mechanisms operating on the present state of the world, it becomes easier to forecast future states of the world accurately over time.

It is important to recognize that it may become more difficult to forecast future states of the world accurately over time in some instances at least for two reasons (Van den Broeke et al., 2019): First, higher quantities of information may hurt forecasting accuracy over time due to information overload (Edmunds & Morris, 2000; Gross, 1964). Second, forecasters may pay attention to lower instead of higher quality information (i.e., noise instead of signals) (Gigerenzer & Todd, 1999; Kahneman et al., 2021; Satopää et al., 2021). While these factors may hurt forecasting accuracy over time in some instances, empirical evidence shows that forecasters usually predict future states of the world more accurately over time as the quantity and quality of information generally increases over time (Himmelstein et al., 2022; Moore et al., 2017).

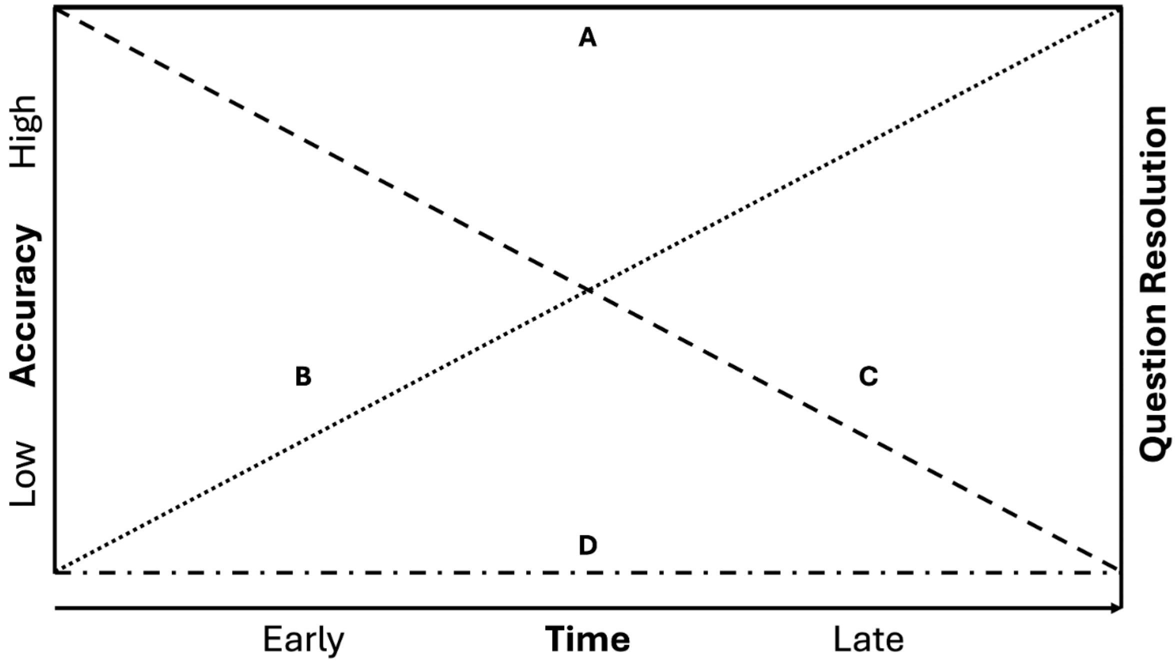
3.3.2 Reconceptualizing Foresight

Based on our forecasting framework, we reconceptualize foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. According to this definition, forecasters possess perfect foresight if they predict future states of the world with perfect accuracy over time, for example, from the beginning to the end of a question in a forecasting tournament. Forecasters may deviate from perfect foresight in three ways: First, forecasters may provide inaccurate forecasts over time. For instance, one forecaster may consistently provide perfectly accurate forecasts at each point in time, whereas another forecaster may consistently make inaccurate forecasts at each point in time. Second, forecasters may start forecasting at different points in time. For example, one forecaster may consistently forecast a question perfectly accurate from the beginning to the end of a question, whereas another forecaster may also perfectly forecast the question but only starts forecasting one day before the question ends. Third, forecasters may provide inaccurate forecasts at different points in time. For example, one forecaster may be inaccurate at the beginning but accurate at the end of a question, whereas another forecaster may be accurate at the beginning but inaccurate at the end of a question. In all these cases, the former forecasters possess superior foresight compared to the latter forecasters as it becomes easier to forecast future states of the world accurately over time (see also Figure 3.1).

It may seem counterintuitive that a forecaster who is inaccurate at the beginning but accurate at the end of a question possesses superior foresight compared to another forecaster who is accurate at the beginning but inaccurate at the end of the question. This appears to contradict our forecasting framework positing that it is easier to accurately forecast future states of the world later rather than earlier. But if it becomes easier to forecast future states of the world over time, forecasting future states of the world accurately at the beginning but inaccurately at the end of a question suggests that the initial accurate forecasts were due to luck rather than true foresight. Forecasters who possess true foresight either update their initial inaccurate forecasts so that their forecasts become more accurate over time or do not revise their initial perfectly accurate forecast.[10] Taken together, we consider foresight to be a bidimensional construct consisting of an accuracy and time dimension that is revealed in the accuracy of forecasts over time. In the following, we will propose a corresponding operationalization of foresight.

[10] One may argue that an accurate early forecast is worth more than an accurate late forecast as decisions may be based on earlier rather than later forecasts in practice. However, only accurate early forecasts that are rooted in true foresight will enable managers to consistently make better decisions. In the long term, lucky accurate early forecasts will ensue unlucky inaccurate early forecasts resulting in poor decisions. Furthermore, decision makers often adjust and reverse their decisions based on later, more inaccurate forecasts, which often results in negative outcomes (Fildes et al., 2009; Van den Broeke et al., 2019). Thus, it is theoretically and practically useful to define foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time.

Figure 3.1 Prototypical Forecasters Predicting Future States of the World With Varying Accuracy Over Time



Note. Solid line = forecaster A who makes perfectly accurate forecasts over time (i.e., from the beginning to the resolution of the question); dotted line = forecaster B who makes early inaccurate but late accurate forecasts; dashed line = forecaster C who makes early accurate but inaccurate late forecasts; dot-dashed line = forecaster D who makes perfectly inaccurate forecasts over time (i.e., from the beginning to the resolution of the question). According to our forecasting framework suggesting that it becomes easier to predict future states of the world accurately over time, the foresight of the prototypical forecasters can be ranked as follows: A > B > C > D.

3.4 Operationalizing Foresight

First, we will briefly explain what forecasting tournaments are, what strictly proper scoring rules are, and why forecasting tournaments in conjunction with strictly proper scoring rules are ideal to examine foresight. Second, we will describe the Brier score (Brier, 1950), the mean daily Brier score (MDBS), and ignorance-prior Brier score (IPBS). Third, we will propose the time-weighted Brier score (TWBS) operationalizing foresight according to our reconceptualization.

3.4.1 Forecasting Tournaments and Strictly Proper Scoring Rules

Forecasting tournaments are prediction polls in which participants make probabilistic forecasts (Budescu & Du, 2007) on questions that each relate to a future state of the world at a specific point in time (Atanasov et al., 2017; Tetlock et al., 2014). For each question, the participants allocate probabilities to a finite number of categories representing an exhaustive set of answers to the question. When the participants have made a probability forecast, this probability forecast is retained until they update their probability forecast, which they can do at any time until the question is resolved. The probability forecasts of each participant on a question are recorded daily. A question is resolved when the true state of the world at the resolution date is known. When a question is resolved, the forecasts of each participant can be evaluated, as it is clear which category is the true state of the world (see Atanasov et al. 2017, Mellers, Stone, Murray, et al. 2015, and Tetlock et al. 2014).

The optimal elicitation and evaluation of probability forecasts in forecasting tournaments requires strictly proper scoring rules (Gneiting & Raftery, 2007; Winkler, 1994). Strictly proper scoring rules incentivize forecasters to make careful and honest forecasts that correspond to their true beliefs about the future states of the world (Wallsten & Budescu, 1983), as they can only maximize their expected reward when they forecast their true beliefs (Gneiting & Raftery, 2007; Winkler, 1969). Furthermore, strictly proper scoring rules quantify the quality of forecasts depending on the realized future state of the world by rewarding forecasters for providing calibrated and sharp[11] probability forecasts (Gneiting & Raftery, 2007). Probability forecasts are calibrated if the empirical frequencies of the event outcomes align with the forecaster's probability forecasts (Gneiting et al., 2007). Specifically, if the forecaster is calibrated, then 25% of all those events, for which they forecast 0.25, occur, and this alignment does not hold only for 0.25 but for any probability in the unit interval. Probability forecasts are sharp if they have high confidence (Gneiting & Katzfuss, 2014; Gneiting & Raftery, 2007). For instance, if forecasters allocate a probability of 1 to a category, they provide a maximally sharp forecast. Under a strictly proper scoring rule, forecasters receive the maximum reward if they are both calibrated and maximally sharp, leading to perfect predictions of the future state of the world (see Gneiting and Raftery 2007 for a formal definition). For example, if the true probability that it rains on a given day is on average 50% and a forecaster allocates a probability of 1 to rain on days that it is actually raining and a probability of 1 to no rain when it is actually not raining, the forecaster is perfectly calibrated and maximally sharp, which corresponds to perfect predictions of future states of the world.

Forecasting tournaments are the ideal research design to examine foresight, as both the accuracy and timing of forecasts can be measured. Accuracy can be measured by computing the deviations of the probability forecasts from the true state of the world.

---

[11] Mellers, Stone, Murray, et al. (2015) and Satopää et al. (2021) refer to this as the forecasters' resolution.

Timing can be measured by recording when the forecasters made their probability forecasts, which is done by default in forecasting tournaments. Based on this information, it can be measured how accurately a forecaster predicted the future state of the world relative to the point in time when the question was resolved. However, existing strictly proper scoring rules mainly focus on the accuracy of probability forecasts and thus, do not fully consider the timing of probability forecasts as an additional quality criterion. In the following, we will generalize the BS, which has previously been used to measure foresight in forecasting tournaments, to the TWBS, which fully incorporates the timing of probability forecasts as an additional quality criterion.

3.4.2 Brier Scores

The current conceptualization of foresight is usually operationalized by the Brier score (Brier, 1950), which is a strictly proper scoring rule.[12] The Brier score (BS) is formally defined as

$$BS_{iq} = \frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2, \tag{3.1}$$

where BS_{iq} is the BS of forecaster $i = 1, \ldots, N$ ($N \in \mathbb{N}$ the number of forecasters) on question $q = 1, \ldots, Q$ ($Q \in \mathbb{N}$ the number of questions); T_{iq} is the number of time-units (usually days) $t = 1, \ldots, T_{iq}$ forecaster i had an active forecast on question q; C_q is the number of possible disjunct classes $c = 1, \ldots, C_q$ of question q the event can fall into; f_{iqtc} is the probability forecast of forecaster i on question q at time t for class c; and o_{qc} indicates whether the class c is the true state of the world at the time of the resolution of question q ($o_{qc} = 1$ if yes, $o_{qc} = 0$ if not).

The BS is a mean squared error. It measures the deviation of the forecasts f_{iqt} of forecaster i on question q at time t from the true state of the world when question q is resolved. As the BS provides the smallest reward for a probability forecast of $1/C_q$ (e.g., a score of 0.5 for $C_q = 2$) and provides the largest reward for a probability forecast of 1 on the category representing the true future state of the world to be forecast, the BS is a symmetric scoring rule (Winkler, 1994). Furthermore, researchers who use the BS typically do not penalize forecasters for providing no forecast at a specific day. In other

[12] Although other scoring rules exist (cf. Gneiting and Katzfuss 2014), research on forecasting tournaments has mainly used the Brier and related scores. For this reason, we focus on and extend the Brier score.

words, they exclude missing values, which may contain potentially valuable information about the foresight of forecasters. Furthermore, the forecast errors are equally weighted. In essence, the BS rewards forecasters for calibrated and sharp forecasts but does not consider the timing of forecasts. Consequently, forecasters can delay their initial forecasts until more information becomes available, which makes it easier to forecast future states of the world accurately over time. That is, the BS measures foresight conceptualized as the ability to consistently forecast future states of the world accurately.

Partly recognizing the neglect of the time dimension, three studies used the mean daily Brier score (MDBS) to operationalize foresight (Atanasov et al., 2017, 2020; Mellers, Stone, Atanasov, et al., 2015). The MDBS is formally defined as

$$MDBS_{iq} = \frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2 \tag{3.2}$$

$$\text{with } \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2 = \frac{1}{|\boldsymbol{N}_{qt}|} \sum_{n \in \boldsymbol{N}_{qt}} \sum_{c=1}^{C_q} (f_{nqtc} - o_{qc})^2$$

$$\text{if forecaster } i \text{ made no forecast on question } q \text{ at time } t,$$

where $MDBS_{iq}$ is the MDBS of forecaster i on question q; $\boldsymbol{N}_{qt}$ is the set of indices of forecasters who had an active forecast on question q at time t; $|\cdot|$ is the cardinality of the set (i.e., the number of elements in the set); and f_{nqtc} is the probability forecast of forecaster n on question q at time t for class c.

The MDBS extends the BS by assigning forecasters who delayed their initial prediction on the question with the mean daily Brier score of all forecasters who have provided a forecast as of that specific day. Thereby, the MDBS incentivizes forecasters to forecast future states of the world when they believe that they know more than the average forecaster. While this partly considers the timing of forecasts, this still incentivizes forecasters to delay forecasting until they think that they possess more information than the average forecaster. In the meantime, they benefit from other forecasters who make more accurate forecasts over time as it becomes easier to forecast future states of the world over time. This likely distorts the measurement of interindividual differences in foresight.

To remedy this issue, one can conceive the ignorance prior Brier score (IPBS), which assigns a score of $1 - 1/C_q$ (e.g., a score of 0.5 for $C_q = 2$) to forecasters who provided no forecast at a day. The score of $1 - 1/C_q$ reflects the ignorance prior as this score results from the assignment of equal probabilities to the potential outcomes.

Thereby, the IPBS incentivizes forecasters to forecast future states of the world when they believe that they know something more than chance. While this enables the undistorted measurement of interindividual differences in foresight, it still does not fully reflect the time dimension of foresight according to our reconceptualization and forecasting framework. The forecasts are equally weighted so that a forecaster who makes accurate early but inaccurate late forecasts is considered to have equal foresight compared to a forecaster who makes inaccurate early but accurate late forecasts. In the following, we propose the time-weighted Brier score to solve this issue.

3.4.3 Time-Weighted Brier Score

To operationalize our reconceptualization of foresight, we propose the Time-Weighted Brier score (TWBS), which is formally defined as

$$TWBS_{iq} = \frac{\sum_{t=1}^{T_q} w_t \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} w_t} \tag{3.3}$$

$$\text{with} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2$$

$$= 1 - 1/C_q \text{ if forecaster } i \text{ made no forecast on question } q \text{ at time } t,$$

where $TWBS_{iq}$ is the TWBS of forecaster i on question q; T_q is the number of time-units (usually days) $t = 1, \dots, T_q$ question q was forecastable; and $w_t > 0$ is any weight at time t that is strictly monotonically increasing over time $t = 1, \dots, T_q$.

Extending the BS from a mean squared error to a weighted mean squared error, the TWBS weights the forecast error $\sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2$ depending on the timing of forecaster i's forecast (i.e., when forecaster i made their forecast) by multiplying the forecast error with the value of w_t at time t when forecaster i made their forecast. Thereby, the TWBS weights the forecast error $\sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2$ of forecaster i on question q at time t heavier, the closer forecaster i made their forecast relative to the realization of the future state of the world. Errors are punished more severely over time as it becomes easier to forecast future states of the world over time. This incentivizes forecasters to update their forecasts regularly. Like the IPBS, the TWBS penalizes forecasters for providing no forecast at a day with a score of $1 - 1/C_q$. Thus, the TWBS incentivizes forecasters to forecast future states of the world when they believe that they know something more than chance. The TWBS operationalizes foresight conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time (see Appendix 3.A for a detailed explanation and example).

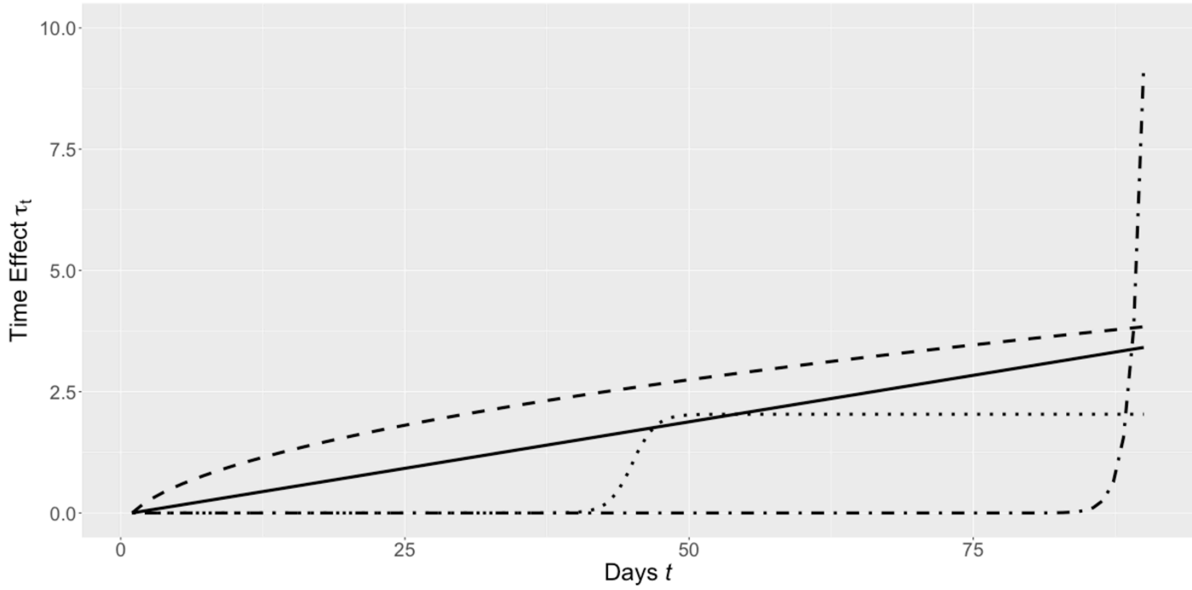
The TWBS is a strictly proper scoring rule that represents a mathematical generalization of the BS (see Appendix 3.B for the proof). The TWBS can be easily used in practice, as forecasting tournaments contain all information necessary for the computation of the TWBS by default (i.e., the forecasts f_{iqt} of forecaster i on question q at time t). Thus, the TWBS can always be calculated when the Brier Score can be calculated. The TWBS can be used in conjunction with mechanisms ensuring incentive-compatible forecasting competitions as the TWBS is bounded (Witkowski et al., 2022). Furthermore, the logic of the time-weighted average can also be used in conjunction with other (e.g., ordered) scoring rules (Gneiting & Katzfuss, 2014; Jose et al., 2008). More generally, the TWBS can be used to measure the performance of a forecasting system producing time-series probability forecasts (Regnier, 2018).

As the TWBS allows practitioners to determine the time-dependent weights w_t themselves, they can model various theoretical signal trajectories describing the quantity and quality of signals that become available at a specific point in time based on their theoretical expectations.[13] For example, practitioners may assume a linear true signal trajectory, which means that the quantity and quality of signals increase at a constant rate over time $(w_t = t)$. They may also assume a square root signal trajectory, which implies that signals continuously become available but the rate and quality at which the signals become available diminishes towards the resolution date $\left(w_t = \sqrt{t}\right)$. Practitioners may also assume that the availability and quality of signals increases rapidly at a specific point in time, which corresponds to a logistic signal trajectory $\left(w_t = \frac{1}{1+e^{-1(t-0.5T)}}\right)$. T is the number of days the event is forecastable and $0.5T$ means that the rapid increase takes place after 50% of the days. Lastly, practitioners may assume that the highest quality and quantity of signals become rapidly available briefly before the event realizes, which corresponds to an exponential signal trajectory $(w_t = e^t)$. Figure 3.2 illustrates these signal trajectories and provides real-world

[13] There are two general ways to determine the time-dependent weights of the TWBS. First, researchers and practitioners may assess the quantity and quality of signals a priori based on their theoretical expectations. For example, when forecasting stock market prices, researchers and practitioners may assume a linear signal trajectory and use the linear TWBS. Second, researchers and practitioners may assess the quantity and quality of signals a posteriori by estimating the weights of the TWBS based on empirical proxies like the standard deviation of the forecasts on a question at each point in time (as a measure of uncertainty), the number of newspaper reports, or Google trends. Although both approaches are feasible, we recommend the a priori approach as the purpose of strictly proper scoring rules is not only the evaluation but also the elicitation of forecasts, which requires researchers and practitioners to determine the incentive structure of the score a priori (Gneiting & Raftery, 2007).

examples. Taken together, the TWBS is a strictly proper scoring rule that quantifies how accurately a forecaster predicted the future state of the world over time.

Figure 3.2 Illustration of the Theoretical Signal Trajectories



Note. Solid line = linear signal trajectory (e.g., forecasting stock market prices at the end of the day, where the prices are updated every second); dashed line = square root signal trajectory (e.g., forecasting the results of an election, where earlier vote updates are more indicative of the final election results than later vote updates); dotted line = logistic signal trajectory (e.g., forecasting a business decision, such as whether Elon Musk will acquire Twitter, where the decision is made and announced before the scheduled end of the question); dot-dashed line = exponential signal trajectory (e.g., forecasting a sports game, where the score is 0-0 until close to the end of the game when a team scores last minute). It is evident that the theoretical signal trajectories differ greatly in their assumption of how many signals become available on each day t and thus, how much easier it becomes to forecast future states of the world accurately, which is reflected in the time effect τ_t.

3.5 Simulation Study

We conducted a simulation study to provide insights into three questions: First, do scores considering the timing of forecasts offer an advantage over the BS in terms of measuring foresight in a setting where it becomes easier to forecast future states of the world over time? Second, what is a reasonable default weight w_t of the TWBS to measure the true foresight of forecasters under various theoretical signal trajectories? This question is important to answer because ex ante our forecasting framework and the TWBS suggest that matching the score to an anticipated signal trajectory is important. Third, what is the best score to measure foresight conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time? A simulation study is the ideal design to answer these questions because the data-generating population model is known to researchers. Therefore, we can know the true foresight of forecasters and the true signal trajectory influencing the difficulty of forecasting future states of the world. This allows us to evaluate how accurately different scoring rules measure the true foresight of forecasters under different true theoretical signal trajectories. To this end, we correlate each score with the true foresight of forecasters.

3.5.1 Design

The simulation study consisted of a 2x2x4 design yielding 16 conditions with 1000 simulation trials each. In each simulation trial $s = 1, 2, \ldots, 1000$, we generated a forecasting tournament where $N = 200$ forecasters predict $Q = 10$ questions that consisted of $c = 1, 2$ categories and was forecastable on $t = 1, 2, \ldots, T$ days with $T = 90$. The number of forecasters, the number of questions, and the number of days the question was forecastable are typical of forecasting tournaments. However, the number of questions predicted by each forecaster is rather at the lower boundary. Although questions in forecasting tournaments can have more than two categories, we chose this value for simplicity.

We modelled the forecast f_{iqt} of forecaster i on question q at time t as beta distributed with a mean of

$$E(F_{iqt}) = \frac{1}{1 + e^{-\lambda_q(\theta_i - (\delta_q - \tau_t))}} \tag{3.4}$$

where θ_i is the true foresight of forecaster i; λ_q is the discrimination of question q, which determines the impact of θ_i on their forecasts on question q and thus captures natural differences in how useful forecasters' true foresight is to forecast a question; δ_q is the difficulty of question q, which represents how difficult it is to forecast question q irrespective of θ_i; and τ_t is the time effect at time t, which is subtracted from δ_q as it represents a proxy for the amount and quality of signals that become available at time t. This reflects the assumption that questions become easier over time. The modelling approach follows the logic of a two-parameter item response theory (IRT) model (see for the binomial case Birnbaum 1968), which is a statistical framework for modelling how person abilities and item characteristics influence responses to items. Although IRT models are primarily used to measure interindividual differences in abilities, they also describe the process how forecasters make predictions: Forecasters have a stable ability that determines the forecasts on the true category in interaction with the difficulty of the question. This stable ability can be interpreted as the result of underlying cognitive processes, such as selecting and interpreting information from the information universe (cf. Satopää et al., 2021).

We characterized forecasters by five parameters: their (1) foresight θ_i describing their ability to forecast future states of the world early and consistently accurate over time; (2) forecast precision π_i indicating how noisy the forecasts of forecaster i are; (3) strategic forecast delay σ_i defining to what extent forecaster i delays their initial forecast until more information becomes available; (4) miscellaneous forecast delay μ_i representing to what extent forecaster i delays their initial forecast for other reasons (e.g., because they did not have time to make a forecast initially); and (5) updating behavior υ_i specifying the number of updates forecaster i makes.

We sampled the person parameters from a multivariate normal distribution,

$$\begin{pmatrix}\theta_i\\ \pi_i\\ \sigma_i\\ \mu_i\\ \upsilon_i\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\ 15\\ -1\\ -2\\ 3\end{pmatrix}, \begin{pmatrix}1 & & & & \\ 0.5 & 2.5 & & & \\ -1.25 & -1.25 & 1 & & \\ -0.5 & -0.5 & 1.25 & 1 & \\ 0.5 & 0.5 & -1.25 & -0.5 & 1\end{pmatrix}\right),$$

with means $\sigma_i = 0$ (high strategic forecast delay) and $\mu_i = -1$ (high miscellaneous forecast delay) depending on the simulation condition. The covariances between the parameters equal medium correlations ± 0.5 and reflect empirical findings that foresight θ_i should be positively related to the precision of forecasts π_i and the number of updates υ_i (Tetlock & Gardner, 2016). Furthermore, our forecasting framework suggests that foresight θ_i should be negatively related to the missing value rate μ_i and delaying behavior σ_i as forecasters who possess foresight should be able to accurately forecast future states of the world early. For the question parameters, we sampled the difficulties δ_q and the discriminations λ_q from a multivariate normal distribution, $\begin{pmatrix}\delta_q\\ \lambda_q\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\ 1\end{pmatrix}, \begin{pmatrix}1 & \\ 0.3 & 1\end{pmatrix}\right)$. The assumed correlation between the difficulty δ_q and discrimination λ_q of question q is typical of 2 PL IRT models (Birnbaum, 1968). We truncated the lower bound of precision π_i to 2, the lower bound of updating behavior υ_i to 0, the difficulty so that $\delta_q \in [-1, 3]$, and the discrimination so that $\lambda_q \in [0.3, 1.5]$ via rejection sampling. This avoided trials with extreme and unrealistic parameters, which ensures stable simulation trials.

We modelled four theoretical signal trajectories $\tau \in \{\tau_{lin}, \tau_{sqrt}, \tau_{logit}, \tau_{exp}\}$ that determined τ_t: a signal trajectory following a linear function $\tau_{lin}(t) = t$; a signal trajectory following a square root function $\tau_{sqrt}(t) = \sqrt{t}$; a signal trajectory following a logistic function $\tau_{logit}(t) = \frac{1}{1+e^{-1(t-0.5T)}}$, where T is the number of days the event is forecastable; and a signal trajectory following an exponential function $\tau_{exp}(t) = e^{t}$. τ_t was standardized so that it matched the scale of the other model parameters. The minimum of the standardized time effect τ_t was set to 0 by subtracting the minimum of the standardized time effect τ_t from the standardized time effect τ_t to ensure that the standardized time effect is positive and thus, reduces the difficulty δ_q over time. Corresponding to our forecasting framework, all theoretical signal trajectories share the assumption that the availability and quality of signals regarding the specific future state of the world to be forecast increases over time. This ensures that forecasting future states of the world becomes easier over time. However, the theoretical signal trajectories differ in their assumption of how many high-quality signals become available on each day t.

Based on these forecaster and question parameters, we simulated the manifest forecasts f_{iqtc} in two steps: First, we modelled the potential forecasts f^{*}_{iqt1} of forecaster i on day t on question q and category $c = 1$ by drawing a random value rounded to two digits from a beta distribution with a mean dependent on the item parameters as shown in Eq. (3.4) and with forecaster i's precision π_i, so that shape parameters are $\alpha = E\left(F_{iqt}\right) * \pi_i$ and $\beta = \left(1 - E\left(F_{iqt}\right)\right) * \pi_i$. π_i influences how precise (i.e., variable) the potential forecasts f^{*}_{iqt1} are, where smaller π_i create noisy and larger π_i create precise forecasts. Second, we modeled the initial forecasts of forecasters based on their strategic forecast delay σ_i and miscellaneous forecast delay μ_i. We calculated the strategic forecast delay σ_i and miscellaneous forecast delay μ_i by setting negative values to 0 and positive values to 1. If the miscellaneous forecast delay μ_i was 1, forecaster i made their initial forecast at day $T * \zeta_i$ rounded to integers, where ζ_i is a number randomly drawn from a uniform distribution with bounds $[0.01, 0.50]$. If the strategic forecast delay σ_i was 1, forecaster i made their initial forecast at day $T * \zeta_i$ rounded to integers, where ζ_i is a number randomly drawn from a uniform distribution with bounds $[0.51, 1]$. This implies that forecasters who engaged in miscellaneous forecast delay made their initial forecast in the first 50% of days, whereas forecasters who engage in strategic forecast delay made their initial forecast in the last 50% of days. Having simulated the manifest forecasts on category 1 f_{iqt1}, we set $f_{iqt2} = 1 - f_{iqt1}$ for the days forecaster i made a forecast. Generally, we chose all values because they result in realistic but conservative forecasts (Figures 3.C1–3.C8 in Appendix 3.C show exemplary simulation data for various parameter values). For example, the simulated initial forecasts occur on average slightly earlier compared to the real-world forecasts. This tends to help scores mainly rewarding accurate forecasts. For each condition and trial, we saved the manifest forecasts f_{iqtc} to compute each score, average the scores across questions, and correlate the average scores with the true foresight θ_i.

3.5.2 Analysis

For each condition and trial, we computed the BS, MDBS, IPBS, and various versions of the TWBS for each forecaster i on question q using the manifest forecasts f_{iqtc}:

$$linear\ TWBS_{iq} = \frac{\sum_{t=1}^{T_q} t \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} t} \tag{3.5}$$

$$square\ root\ TWBS_{iq} = \frac{\sum_{t=1}^{T_q} \sqrt{t} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} \sqrt{t}} \tag{3.6}$$

$$logistic\ TWBS_{iq} = \frac{\sum_{t=1}^{T_q} \frac{1}{1 + e^{-1(t-0.5T)}} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} \frac{1}{1 + e^{-1(t-0.5T)}}} \tag{3.7}$$

$$exponential\ TWBS_{iq} = \frac{\sum_{t=1}^{T_q} e^t \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} e^t} \tag{3.8}$$

$$\text{with} \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2 = 1 - \frac{1}{C_q}$$

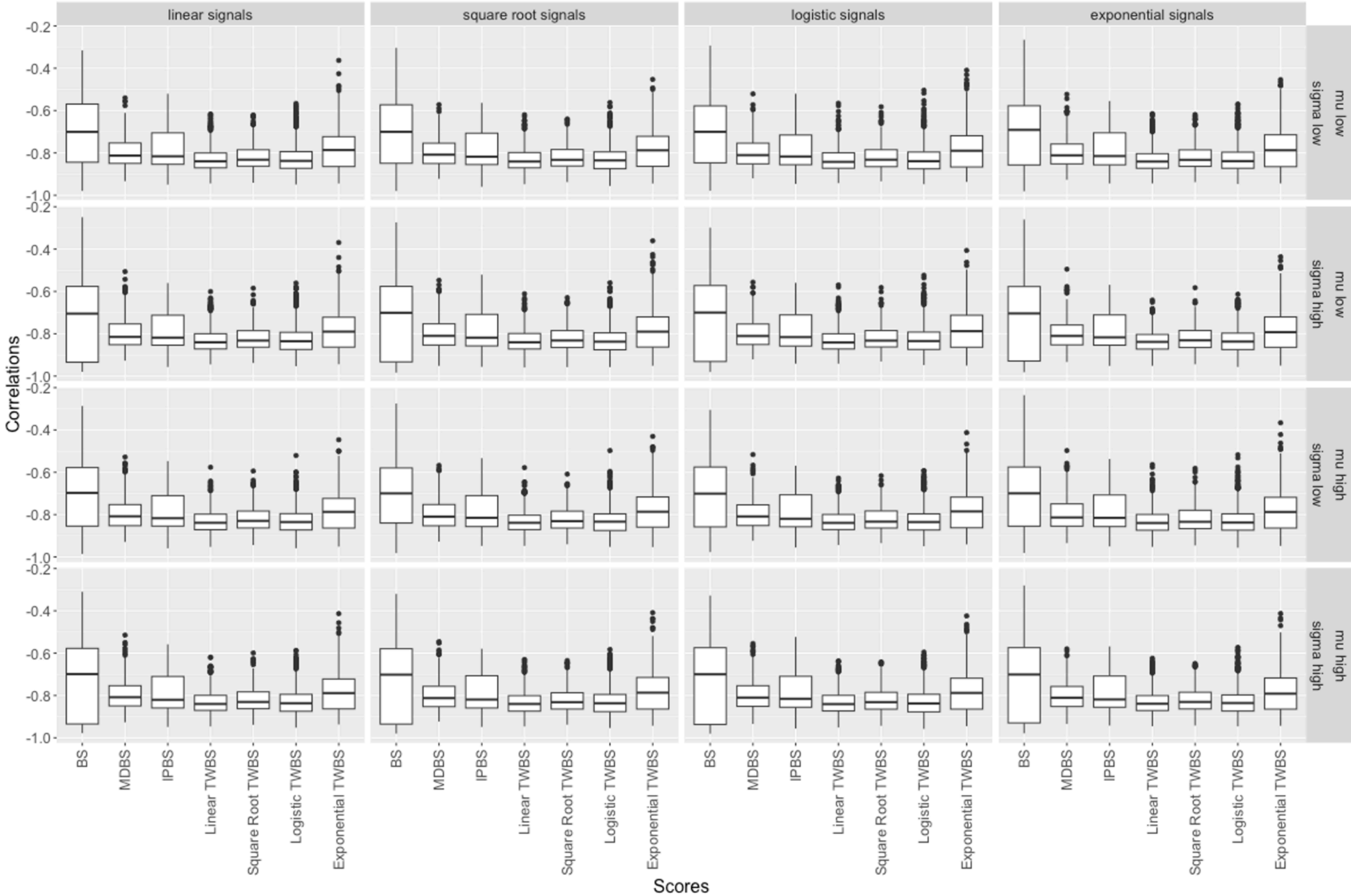
$$\text{if forecaster i made no forecast on question } q \text{ at time } t.$$

Then, we averaged the scores across questions and correlated the average scores with the true foresight of forecasters θ_i. Thus, we obtained 1000 correlations per condition and per scoring rule. These correlations represent the dependent variable of our simulation study. We analyzed our results graphically by drawing boxplots of the correlations for each scoring rule across conditions.

3.5.3 Results

Figure 3.3 displays boxplots of the correlations between the true foresight of forecasters θ_i and each scoring rule across all 1000 trials by simulation condition. As larger values of θ_i represent a better true foresight of forecasters and lower values on the scoring rules indicate better forecasts, more negative correlations indicate a more accurate measurement of the true foresight of forecasters by the corresponding scoring rule.

Figure 3.3 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition



Note. The linear, square root, and logistic TWBS measured the true foresight of forecasters θ_i best in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

The results displayed in Figure 3.3 indicate that the MDBS, IPBS, linear TWBS, square root TWBS, logistic TWBS, and exponential TWBS measure the true foresight θ_i on average more accurately than the BS across all conditions as indicated by the more negative correlations. Although there are only small differences in the average of the better performing scores, there are large differences in the performance stability of the scoring rules: The ranges and interquartile ranges of the linear, square root, and logistic TWBS are considerably smaller than those of the MDBS, IPBS, and exponential TWBS across the simulation conditions. This indicates that the linear, square root, and logistic TWBS measure the true foresight of forecasters more consistently than the remaining scoring rules. It is striking that the performance of all scores is consistent across all simulation conditions. Irrespective of the true underlying signal trajectory τ, strategic forecast delay σ_i and miscellaneous forecast delay μ_i, the linear, square root, and logistic TWBS robustly measure the true foresight of forecasters accurately. In other words, these scores accurately extract information about the forecasters' true foresight under various true signal trajectories even when forecasters delay their forecasts more or less often for strategic or miscellaneous reasons.

We conducted several robustness checks to corroborate our findings (see Appendix 3.D): First, we presumed the correlations between the person parameters to be 0.3 and 0. Second, we assumed that forecasters forecast one and 20 instead of 10 questions. Third, we assumed that the questions were forecastable 365 instead of 90 days. Fourth, we assumed that for ten percent of the questions the quality of the signals decreases over time (i.e., the signals represent noise instead of information) up until the final day of the forecast. The robustness checks yielded the same qualitative results with two exceptions: First, the linear, square root, and logistic TWBS had no advantage over the BS, MDBS, and IPBS in average accuracy and consistency when the correlations between the person parameters were 0. One explanation for this finding is that the TWBS extracts information about the true foresight of forecasters based on the forecast delays. If the forecast delays are uncorrelated with the true foresight of forecasters, this information represents noise and cannot improve the accuracy and consistency of the TWBS compared to the other scores. Second, the logistic TWBS performed slightly worse in terms of average accuracy and consistency than the linear and square root TWBS when forecasters only forecast 1 question. This suggests that the logistic TWBS is less accurate and consistent than the linear and square root TWBS in the measurement of the true foresight for forecasters who only forecast one question. Therefore, both $w_t = \tau_{linear}$ and $w_t = \tau_{sqrt}$ are reasonable default weights for the TWBS. Generally, the advantage of the TWBS over the other scores increases with the number of questions the forecasters forecast as indicated by the results where forecasters forecast 20 instead of 10 questions. Furthermore, the TWBS outperforms the other scoring rules the longer the questions were forecastable as shown by the results where forecasters forecast 10 questions each lasting 365 days. Remarkably, these results even hold when the quality of signals decreases over time for ten percent of the questions. Taken together, the results suggest that the linear and square root TWBS are the best scores to measure foresight defined as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time.

3.6 Application to Geopolitical Forecasting Data

To evaluate the scoring rules empirically, we conducted two analyses using geopolitical forecasting data gathered in the context of government intelligence analysis consisting of 414,168 scores for 9,694 forecasters on 498 questions over a period of four years: First, we computed measures of linear association between the scoring rules to assess whether the scoring rules come to different conclusions about the true foresight of forecasters. Second, we calculated measures of linear association between the scoring rules and 29 variables to determine whether the scoring rules exhibit different correlational patterns, which would suggest that the scoring rules measure different concepts of foresight. These analyses may also provide empirical insight into whether one of the strictly proper scoring rules is preferable over the others. In the following, we will describe the data, explain our analytical approach, and report the results.

3.6.1 Geopolitical Forecasting Data

We used publicly available data from a geopolitical forecasting tournament that was run over four years from 2011 to 2015 as part of the Good Judgment Project that took part in the Aggregative Contingent Estimation project sponsored by the Intelligence Advanced Research Projects Activity.[14] Forecasters made probabilistic forecasts on questions, such as "Will Bashar al-Assad remain President of Syria through 31 January 2012?" The questions were forecastable on average on 112 days ($SD = 93, median = 83, range = 548$); 382 questions had two, 47 questions had three, 50 questions had four, and 19 questions had five answer categories. All questions were related to politics or politically relevant events. As part of the project, participants were randomly assigned to various treatment groups (Mellers et al., 2014): For example, the

[14] The data is publicly available at https://dataverse.harvard.edu/dataverse/gjp.

training group received a probability training, the teaming group consisted of forecasters who were assigned to teams, and the tracking group consisted of teams of superforecasters.[15] In addition, measures of various psychological constructs were collected as part of the project. Table 3.E1 in Appendix 3.E provides an overview including descriptive statistics of the treatment groups and empirical measures. For the sake of brevity, we refer the reader to the public data for a full description of the treatment groups as well as measures and to previous research for more details on the forecasting tournament (cf. Mellers et al. 2014, Mellers, Stone, Atanasov, et al. 2015, Mellers, Stone, Murray, et al. 2015, Moore et al. 2017, Satopää et al. 2021)

3.6.2 Analytical Approach

We fitted ordinary least squares (OLS) regressions to obtain a measure of linear association between the scoring rules by regressing the scores on each other (e.g., the linear TWBS on the BS). Furthermore, we fitted OLS regressions to obtain a measure of linear association between each score and the empirical measures listed in Table 3.E1 by regressing each score on each empirical measure (e.g., the linear TWBS on training). We computed heteroscedasticity-consistent standard errors clustered by forecasters to account for the nested data structure resulting from the fact that forecasters usually forecast more than one question (Liang & Zeger, 1986). Furthermore, we included question and group fixed effects to account for differences in the scores attributable to forecasters deciding to forecast different questions and their assignment to different experimental groups during the tournament. Note that we did not include year fixed effects, as differences between years are indirectly controlled for by the question fixed effects (each question belonged to a specific year). We computed linear models with one predictor, because we did not have a theoretical justification for nonlinear relationships and the estimated regression weight is conceptually similar to the Pearson correlation coefficient, which is the standard measure of linear association for simple data structures but would yield biased estimates in this case due to the dependencies in the data.

As the normal Q-Q plots of most models indicated that the errors were non-normally distributed, we implemented a nonparametric cluster bootstrap approach to compute 99.92% studentized *t* confidence intervals based on 10,000 bootstrap replications as measures of uncertainty and significance of the estimated parameters

[15] As the superforecaster status depended on prior forecasting performance, this group was not randomly assigned. Furthermore, forecasters in the teaming and tracking group still made their own predictions even though they worked in teams.

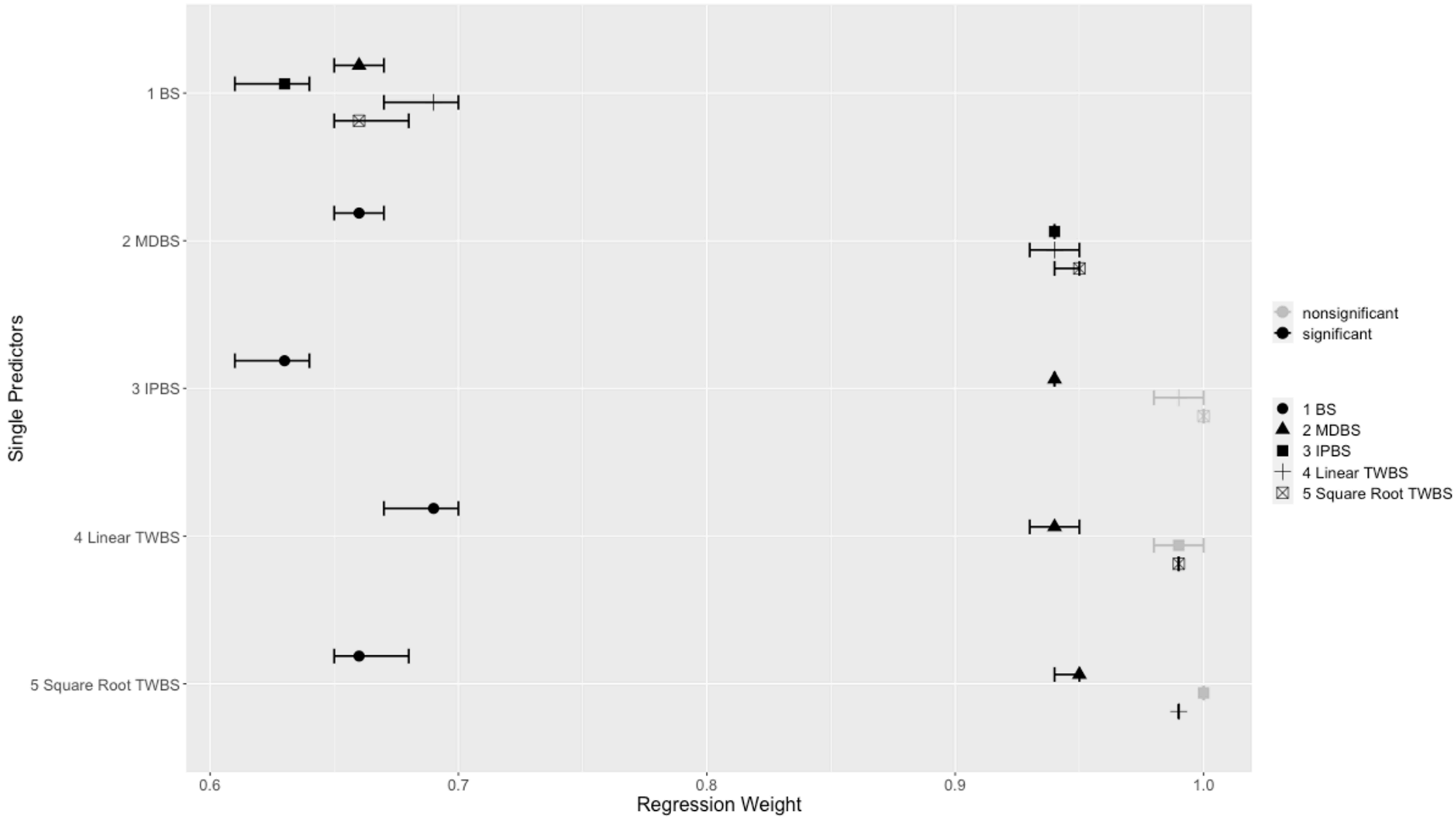
(Efron & Tibshirani, 1994). We computed 99.92% confidence intervals, which correspond to the Bonferroni corrected alpha level $\alpha \approx 0.08\%$ for 29 two-sided tests per score, to prevent an inflation of the type I error rate. All analyses were done in-sample. Our analytical approach is described in detail in Appendix 3.E.

3.6.3 Results

Figures 3.4 and 3.5 show the results of the OLS linear regressions and the 99.92% studentized *t* bootstrap confidence intervals. For the sake of brevity, we only report the precise quantitative results (cf. Appendix 3.F Table 3.F1) in the paper when they add insight to the qualitative description of the results. The results in Figure 3.4 suggest that all scoring rules are positively related to each other. While the BS exhibits medium strong positive linear associations with all other scoring rules, ranging from $b = 0.63$ to $b = 0.69$, the scoring rules considering the accuracy and timing of forecasts exhibit very strong positive linear associations among each other, ranging from $b = 0.94$ to $b = 1.00$. All scoring rules are positively related to each other as all scores reward accurate forecasts. Yet, the BS differs markedly from the other scoring rules, which consider both the accuracy and timing of forecasts.

Strikingly, the linear associations between the IPBS and the linear TWBS ($b = 0.99, 99.92\%\ \text{CI} = [0.99; 1.00]$) as well as the IPBS and the square root TWBS ($b = 1.00, 99.92\%\ \text{CI} = [1.00; 1.00]$) do not differ significantly from 1. This means that the IPBS statistically comes to the same conclusion about the relative foresight of forecasters as the linear and square root TWBS even though the scoring rules differ in their weighting of the forecast errors over time. In contrast, the linear and square root TWBS come to small but significantly different conclusions about the true foresight of forecasters ($b = 0.99, 99.92\%\ \text{CI} = [0.99; 0.99]$). Taken together, our results suggest that the BS comes to different conclusions about the true foresight of forecasters than the other scoring rules but there only seem to be small differences between the MDBS, IPBS, linear TWBS, and square root TWBS.

Figure 3.4 Results of the OLS Linear Regressions and the 99.92% Studentized *t* Bootstrap Confidence Intervals: Scoring Rules



Note. Shapes depict the regression weights *b* and the corresponding 99.92% studentized *t* bootstrap confidence intervals of a model regressing each scoring rule on a single predictor (i.e., one other scoring rule). Black shapes mark regression weights that significantly differ from 1, whereas grey shapes indicate regression weights that do not significantly differ from 1.

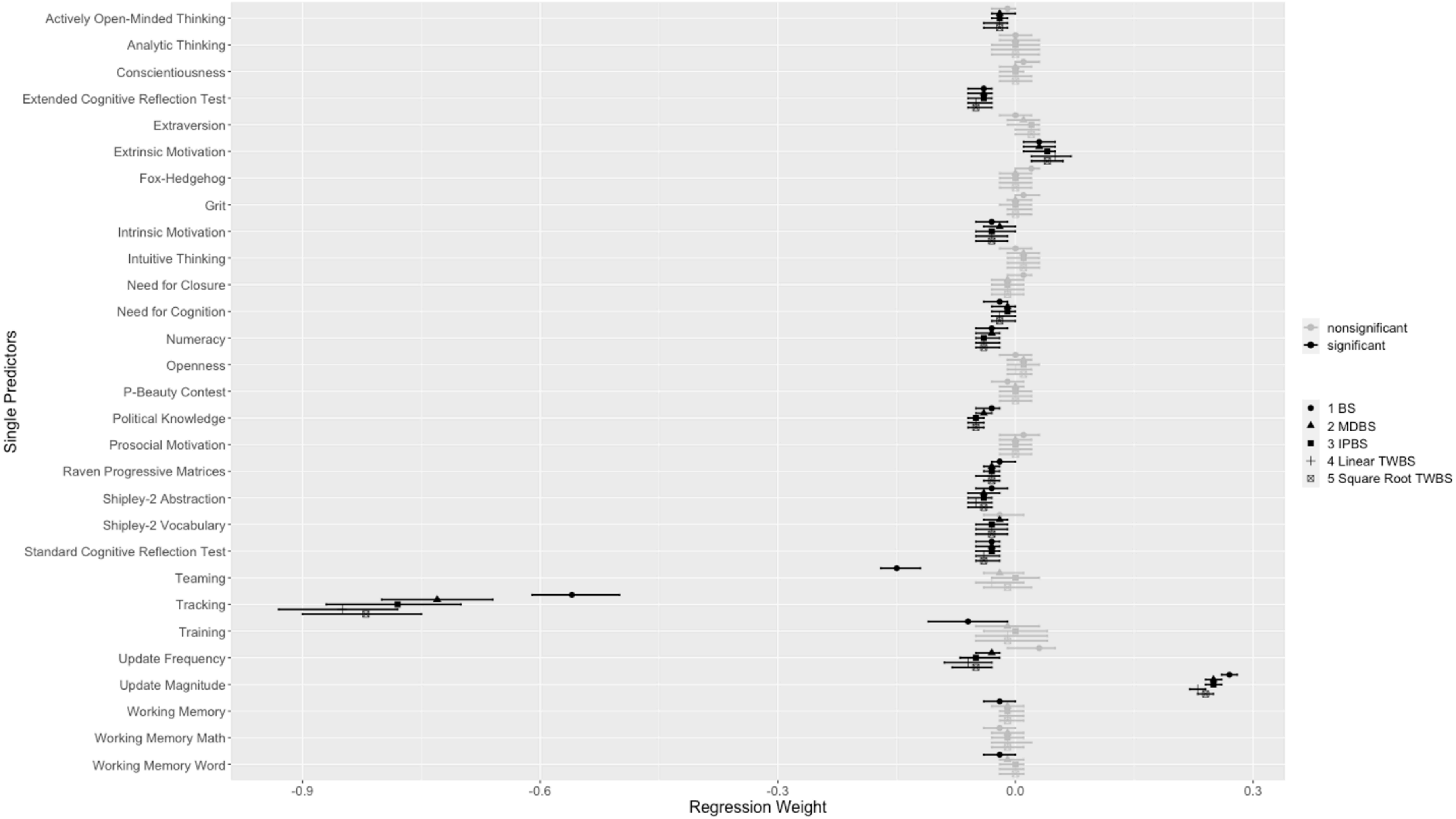
The results in Figure 3.5 suggest that most constructs exhibit only very small positive or no linear associations with the scoring rules. This means that interindividual differences in cognitive styles and personality characteristics generally seem to explain only to a very small degree why forecasters possess foresight. The update magnitude represents an exception because it exhibits small linear associations with the scoring rules, ranging from $b = 0.24$ to $b = 0.28$. This can be explained by the fact that forecasters who revise their forecasts due to new information have to make larger updates if the new information disconfirms their initial predictions. In other words, forecasters only have to update their forecasts if their predictions turn out to be wrong over time. Thus, larger updates indicate worse foresight on all scoring rules.

The training and teaming interventions only exhibit small negative linear associations with the BS but no linear associations with the other scoring rules. This suggests that the training and teaming interventions were successful in improving the accuracy of forecasts but not in continuously improving accuracy over time. More generally, such differential correlational patterns can also be observed with other variables: Actively Open-Minded Thinking, Fox-Hedgehog, Shipley-2 Vocabulary, Update Frequency, Working Memory, and Working Memory Word. These results suggest that the BS measures something different compared to the MDBS, IPBS, linear TWBS, and square root TWBS.

The tracking intervention (i.e., assigning superforecasters to the same team) displays the strongest negative linear associations with the scoring rules compared to all other variables and interventions. Nonparametric 95% bootstrap confidence intervals of the estimated differences between the linear associations of the tracking intervention with the scoring rules (e.g., $\hat{\delta} = b_{linear\ TWBS} - b_{square\ root\ TWBS}$) suggest that the tracking intervention's negative relationship with the linear TWBS ($b = -0.85, 99.92\%\ \text{CI} = [-0.92; -0.78]$) is stronger than its negative associations with the square root TWBS $\left(\hat{\delta} = -0.03, 95\%\ \text{CI} = [-0.03; -0.02]\right)$, IPBS $\left(\hat{\delta} = -0.07, 95\%\ \text{CI} = [-0.08; -0.06]\right)$, MDBS $\left(\hat{\delta} = -0.12, 95\%\ \text{CI} = [-0.13; -0.11]\right)$, and BS $\left(\hat{\delta} = -0.29, 95\%\ \text{CI} = [-0.36; -0.22]\right)$. As the tracking group consists of superforecasters, these results suggest that the linear TWBS assigns better scores to superforecasters and rewards forecasters with superior foresight more than the other scoring rules despite being strongly related to them. This finding provides first empirical evidence that there is an empirical difference between the linear TWBS and the other scoring rules as the tracking group is the only variable (along with the update magnitude), in which differences between the scoring rules are likely to become apparent given that the effect sizes of the other variables were only very small ($b = -0.06$ to $b = 0.05$) or statistically nonsignificant. The sensitivity analyses (see Appendix 3.G) further corroborated these conclusions (Katsagounos et al., 2021).

Taken together, the simulation study and the empirical application suggest that the linear TWBS should be the default operationalization of foresight. The simulation study indicates that the linear TWBS measures the true foresight of forecasters more accurately and consistently than alternative scoring rules under various theoretical signal trajectories. The empirical application shows that the linear TWBS (1) comes to different conclusions about the true foresight of forecasters than the BS, (2) measures something different than the BS, and (3) better identifies forecasters with superior foresight than the BS, MDBS, IPBS, and square root TWBS. Therefore, we recommend the linear TWBS as the standard operationalization of foresight conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time.

Figure 3.5 Results of the OLS Linear Regressions and the 99.92% Studentized *t* Bootstrap Confidence Intervals: Variables



Note. Shapes depict the regression weights *b* and the corresponding 99.92% studentized *t* bootstrap confidence intervals of a model regressing each scoring rule on a single predictor (i.e., one variable). Black shapes mark regression weights that significantly differ from 0, whereas grey shapes indicate regression weights that do not significantly differ from 0.

3.7 Discussion

The goal in this paper was to develop a reconceptualization of foresight that integrates the dimensions of accuracy and time. To provide the theoretical basis for this integration, we proposed a forecasting framework suggesting that forecasting future states of the world accurately becomes easier over time as the quantity and quality of signals increases over time. Based on this forecasting framework, we reconceptualized foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. To operationalize foresight in forecasting tournaments, we proposed various strictly proper scoring rules considering the accuracy and timing of forecasts. Taken together, the simulation study and empirical application suggest that the linear TWBS should be the standard operationalization of foresight. In the following, we will discuss how the forecasting framework, reconceptualization of foresight, and linear TWBS contribute to the emergent literature on foresight before we acknowledge the limitations of our study, outline opportunities for future research, and articulate the practical implications.

3.7.1 Contributions

Our paper contributes to the literature by clarifying the concept, operationalization, and correlates of foresight. We reconceptualized foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. Our concept suggests that foresight is a bidimensional construct consisting of an accuracy and time dimension that is revealed in the accuracy of forecasts over time. In contrast, prior research has either assumed that foresight is completely described by the accuracy of forecasts (Mellers et al., 2014; Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2016) or only partly considered the timing of forecasts (Atanasov et al., 2020; Mellers, Stone, Atanasov, et al., 2015). As our concept of foresight describes the general situation in which one or more forecasters provide one or more forecasts at the same or different points in time, our concept generalizes the prior concept of foresight that is only applicable when forecasters make forecasts at the same point in time. Overall, our reconceptualization contributes to the literature on judgmental forecasting by integrating the accuracy and time dimension of foresight. It also contributes to conversations on foresight in the forecasting and management literature by revising our understanding of what constitutes foresight (Csaszar & Laureiro-Martínez, 2018; Fergnani, 2022; Gavetti & Menon, 2016; Hyndman & Koehler, 2006; Iden et al., 2017; Kapoor & Wilde, 2023; Makridakis, 1993; Marinković et al., 2022; Peterson & Wu, 2021; Rohrbeck et al., 2015).

To provide the theoretical basis for this integration, we proposed a forecasting framework suggesting that forecasting future states of the world accurately becomes easier over time as the quantity and quality of signals increase over time. Prior research mainly relied on the partial information framework (Satopää et al., 2016, 2021), which proposes that forecasters predict future states of the world by sampling and interpreting signals from a static information universe containing all past, present, and future signals. Our forecasting framework extends the partial information framework by (1) suggesting that the signals of the information universe are created by causal mechanisms and stochastic processes and (2) assuming the information universe to be dynamically expanding over time instead of static as the causal mechanisms and stochastic processes create more and higher quality signals over time. Thus, forecasters also differ in the accuracy of their forecasts because they sample signals from the information universe at different times. More generally, our forecasting framework can explain on a more fundamental level why future states of the world are forecastable and why they become easier to forecast accurately over time. In this way, our framework is a step towards a more comprehensive theory of forecasting (Fergnani & Chermack, 2021; Gordon et al., 2020; Lawrence et al., 2006).

From our forecasting framework, we derived the TWBS to operationalize foresight. As the TWBS is grounded in an explicit theoretical rationale that matches our concept of foresight, it represents the theoretically preferred scoring rule to measure foresight conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. In contrast, the BS quantifies foresight conceptualized as the ability to forecast future states of the world accurately (e.g., Mellers et al., 2014), and the MDBS measures foresight conceptualized as the ability to make better forecasts than the average crowd forecast on a day (e.g., Atanasov et al., 2017). While these scoring rules are adequate operationalizations of their corresponding concepts of foresight, we argue that the linear TWBS should be the default operationalization of foresight as we conceptualize it. In general, researchers and practitioners should make the concept of foresight underlying their work explicit and choose a corresponding strictly proper scoring rule.

To evaluate the empirical differences between the BS, MDBS, IPBS, and TWBS, we computed the relationships between the strictly proper scoring rules and various variables based on data gathered in the context of government intelligence analysis consisting of 414,168 scores for 9,694 forecasters on 498 questions over a period of four years. Coinciding with prior research (Mellers, Stone, Atanasov, et al., 2015;

Mellers, Stone, Murray, et al., 2015), we found interindividual differences in thinking styles and personality characteristics to be related to foresight. But these variables explained interindividual differences in foresight only to a small extent in our study. This means that the best way to predict a person's foresight—given our present knowledge—appears to be a person's past foresight, which is consistent with prior research findings (Atanasov & Himmelstein, 2022). As the linear TWBS rewards superforecasters with better scores than the other scoring rules in our study, the linear TWBS can be considered the empirically preferable scoring rule and consequently, should be the default operationalization of foresight. This finding also highlights that while the accuracy of forecasts can provide important information about the foresight of forecasters, the timing of forecasts seems to contain additional valuable information about the foresight of forecasters.

Our analyses also uncovered unexpected correlational patterns. The BS suggests that interventions, such as providing 30 minute online probability trainings and putting regular forecasters in teams, improve foresight (Mellers et al., 2014; Tetlock & Gardner, 2016). In contrast, the linear TWBS suggests that these simple interventions do not improve foresight. Only putting the top 2% forecasters in teams (i.e., the tracking condition) was positively related to foresight here conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time. One explanation for this finding could be that the training and teaming treatments create latency costs (e.g., training and organizing team meetings requires time), which may negatively affect the foresight of forecasters in these conditions as measured by the linear TWBS. Taken together, the linear TWBS seems to be the theoretically and empirically preferred strictly proper scoring rule to measure foresight conceptualized as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time.

3.7.2 Practical Implications, Limitations, and Future Research Opportunities

Our theory and data suggest that foresight should be operationalized by the linear TWBS. The linear TWBS enables researchers and practitioners to better identify forecasters who possess superior foresight (Mellers, Stone, Murray, et al., 2015; Tetlock & Gardner, 2016). Note that the linear TWBS can in principle be used for any probabilistic forecasts that can be updated over time. As the results of our simulation study—like any other simulation study—depend on the generative model, future research should examine alternative generative models to further corroborate our conclusions. Furthermore, while our empirical analyses provided first evidence that the linear TWBS is the preferred scoring rule to measure foresight conceptualized as the ability to forecast future states of the world early and consistently accurate over time, future research should conduct convergent and divergent validity analyses with constructs that were carefully selected based on theory to further establish the construct validity of the linear TWBS relative to other scoring rules that consider the accuracy and timing of forecasts. Such analyses may also reveal variables that are better predictors of the foresight of forecasters than interindividual differences in thinking styles and personality. In light of our surprising findings in relation to the superforecasting literature, future research should replicate and examine potential explanations for these findings. These efforts would also represent an important step towards an integrative theory explaining why some persons possess foresight. Such a theory would also enable us to design better interventions that improve foresight.

To conclude, this study contributes to the emergent literature on foresight by clarifying the concept, operationalization, and correlates of foresight. The rigorous study of foresight using forecasting tournaments as proposed by Phil Tetlock, Barbara Mellers, and other pioneers is truly a ‘game changer’ (Tetlock et al., 2014: 290) for judgmental forecasting research. Based on subjective probability forecasts, forecasting tournaments enable researchers and practitioners to predict events that are important for decisions in public and private organizations but elude more traditional forecasting methods relying on objective frequentist probabilities (Mellers et al., 2023; Tetlock et al., 2023). Although this research has yielded many fascinating empirical findings, it might be time to develop a coherent theory explaining why some forecasters possess superior foresight. Given that judgmental forecasting is so consequential, we hope that our conceptualization of foresight and the linear TWBS equip researchers and practitioners with the conceptual and methodological tools to systematically advance our understanding of foresight and design interventions that improve foresight.

3.8 References

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

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., Levmore, S., Litan, R., Milgrom, P., Nelson, F. D., Neumann, G. R., Ottaviani, M., Schelling, T. C., Shiller, R. J., Smith, V. L., Snowberg, E., Sunstein, C. R., Tetlock, P. C., Tetlock, P. E., … Zitzewitz, E. (2008). The promise of prediction markets. *Science*, *320*(5878), 877–878. https://doi.org/10.1126/science.1157679

Atanasov, P., & Himmelstein, M. (2022). Talent spotting in crowd prediction. In M. Seifert (Ed.), *Judgment in Predictive Analytics*. Springer. https://psyarxiv.com/rm49a/

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., Mellers, B., & Tetlock, P. (2024). Crowd prediction systems: Markets, polls, and elite forecasters. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2023.12.009

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

Barney, J. B. (1986). Strategic factor markets: Expectations, luck, and business strategy. *Management Science*, *32*(10), 1231–1241. https://doi.org/10.1287/mnsc.32.10.1231

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

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.

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

Budescu, D. V., & Du, N. (2007). Coherence and consistency of investors' probability judgments. *Management Science*, *53*(11), 1731–1744. https://doi.org/10.1287/mnsc.1070.0727

Bunn, D., & Wright, G. (1991). Interaction of judgmental and statistical forecasting methods—Issues and analysis. *Management Science*, *37*(5), 501–518. https://doi.org/10.1287/mnsc.37.5.501

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

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367. https://doi.org/10.1037/0033-295X.104.2.367

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

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

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management*, *20*(1), 17–28. https://doi.org/10.1016/S0268-4012(99)00051-1

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall.

Fergnani, A. (2022). Corporate foresight: A new frontier for strategy and management. *Academy of Management Perspectives*, *36*(2), 820–844. https://doi.org/10.5465/amp.2018.0178

Fergnani, A., & Chermack, T. J. (2021). The resistance to scientific theory in futures and foresight, and what to do about it. *Futures & Foresight Science*, *3*(3–4), 1–16. https://doi.org/10.1002/ffo2.61

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018).

The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, *62*(2), 410–422. https://doi.org/10.1093/isq/sqx078

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

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the guidance provided by research. *International Journal of Forecasting*, *9*(2), 147–161. https://doi.org/10.1016/0169-2070(93)90001-4

Goodwin, P., & Wright, G. (2014). *Decision analysis for management judgment*. John Wiley & Sons.

Gordon, A. V., Ramic, M., Rohrbeck, R., & Spaniol, M. J. (2020). 50 Years of corporate and organizational foresight: Looking back and going forward. *Technological Forecasting and Social Change*, *154*, 119966. https://doi.org/10.1016/j.techfore.2020.119966

Gross, B. M. (1964). *The managing of organizations: The administrative struggle*. Free Press of Glencoe.

Harrison, P. J., & Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(3), 205–228. https://doi.org/10.1111/j.2517-6161.1976.tb01586.x

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.

Hogarth, R. M., & Makridakis, S. (1981). Forecasting and planning: An evaluation. *Management Science*, *27*(2), 115–138. https://doi.org/10.1287/mnsc.27.2.115

Horowitz, M., Stewart, B. M., Tingley, D., Bishop, M., Samotin, L. R., Roberts, M., Chang, W., Mellers, B. A., & Tetlock, P. E. (2019). What makes foreign policy teams tick: Explaining variation in group performance at geopolitical forecasting. *Journal of Politics*, *81*(4), 1388–1404. https://doi.org/10.1086/704437

Hossain, T., & Okui, R. (2013). The binarized scoring rule. *The Review of Economic Studies*, *80*(3), 984–1001. https://doi.org/10.1093/restud/rdt006

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Iden, J., Methlie, L. B., & Christensen, G. E. (2017). The nature of strategic foresight research: A systematic literature review. *Technological Forecasting and Social Change*, *116*, 87–97. https://doi.org/10.1016/j.techfore.2016.11.002

Johnstone, D. J., Jose, V. R. R., & Winkler, R. L. (2011). Tailored scoring rules for probabilities. *Decision Analysis*, *8*(4), 256–268. https://doi.org/10.1287/deca.1110.0216

Jose, V. R. (2009). A characterization for the spherical scoring rule. *Theory and Decision*, *66*(3), 263–281. https://doi.org/10.1007/s11238-007-9067-x

Jose, V. R. R. (2017). Percentage and relative error measures in forecast evaluation. *Operations Research*, *65*(1), 200–211. https://doi.org/10.1287/opre.2016.1550

Jose, V. R. R., Nau, R. F., & Winkler, R. L. (2008). Scoring rules, generalized entropy, and utility maximization. *Operations Research*, *56*(5), 1146–1157. https://doi.org/10.1287/opre.1070.0498

Jose, V. R. R., & Winkler, R. L. (2009). Evaluating quantile assessments. *Operations Research*, *57*(5), 1287–1297. https://doi.org/10.1287/opre.1080.0665

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

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science*, *39*(1), 17–31. https://doi.org/10.1287/mnsc.39.1.17

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.

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

Karvetski, C. W., Meinel, C., Maxwell, D. T., Lu, Y., Mellers, B. A., & Tetlock, P. E. (2022). What do forecasting rationales reveal about thinking patterns of top geopolitical forecasters? *International Journal of Forecasting*, *38*(2), 688–704. https://doi.org/10.1016/j.ijforecast.2021.09.003

Katsagounos, I., Thomakos, D. D., Litsiou, K., & Nikolopoulos, K. (2021). Superforecasting reality check: Evidence from a small pool of experts and expedited identification. *European Journal of Operational Research*, *289*(1), 107–117. https://doi.org/10.1016/j.ejor.2020.06.042

Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, *22*(3), 493–518. https://doi.org/10.1016/j.ijforecast.2006.03.007

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lichtendahl, K. C., & Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science*, *53*(11), 1745–1755. https://doi.org/10.1287/mnsc.1070.0729

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289–307.

Makridakis, S. (1986). The art and science of forecasting—An assessment and future-directions. *International Journal of Forecasting*, *2*(1), 15–39. https://doi.org/10.1016/0169-2070(86)90028-2

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3

Marinković, M., Al-Tabbaa, O., Khan, Z., & Wu, J. (2022). Corporate foresight: A systematic literature review and future research trajectories. *Journal of Business Research*, *144*, 289–311. https://doi.org/10.1016/j.jbusres.2022.01.097

Mauboussin, M. J. (2012). *The success equation: Untangling skill and luck in business, sports, and investing*. Harvard Business Review Press.

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study.

Judgment and Decision Making, *12*(4), 369–381.

Mellers, B. A., McCoy, J. P., Lu, L., & Tetlock, P. E. (2023). Human and algorithmic predictions in geopolitical forecasting: Quantifying uncertainty in hard-to-quantify domains. *Perspectives on Psychological Science*, 17456916231185339. https://doi.org/10.1177/17456916231185339

Mellers, B. A., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. E. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology-Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

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

Mellers, B. A., & Tetlock, P. E. (2019). From discipline-centered rivalries to solution-centered science: Producing better probability estimates for policy makers. *American Psychologist*, *74*(3), 290–300. https://doi.org/10.1037/amp0000429

Mellers, B. A., Tetlock, P. E., & Arkes, H. R. (2019). Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition*, *188*, 19–26. https://doi.org/10.1016/j.cognition.2018.10.021

Mellers, B. A., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Moore, D. A., Swift, S. A., Minster, A., Mellers, B. A., Ungar, L., Tetlock, P. E., Yang, H. H. J., & Tenney, E. R. (2017). Confidence calibration in a multiyear geopolitical forecasting competition. *Management Science*, *63*(11), 3552–3565. https://doi.org/10.1287/mnsc.2016.2525

Peterson, A., & Wu, A. (2021). Entrepreneurial learning and strategic foresight. *Strategic Management Journal*, *42*, 2357–2388. https://doi.org/10.1002/smj.3327

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino

Oliveira, F. L., De Baets, S., Dokumentov, A., … Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*(3), 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Regnier, E. (2018). Probability forecasts made at multiple lead times. *Management Science*, *64*(5), 2407–2426. https://doi.org/10.1287/mnsc.2016.2720

Rohrbeck, R., Battistella, C., & Huizingh, E. (2015). Corporate foresight: An emerging field with a rich tradition. *Technological Forecasting and Social Change*, *101*, 1–9. https://doi.org/10.1016/j.techfore.2015.11.002

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting*, *30*(2), 344–356. https://doi.org/10.1016/j.ijforecast.2013.09.009

Satopää, V. A., Pemantle, R., & Ungar, L. H. (2016). Modeling probability forecasts via information diversity. *Journal of the American Statistical Association*, *111*(516), 1623–1633. https://doi.org/10.1080/01621459.2015.1100621

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

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. A. (2023). Decomposing the effects of crowd-wisdom aggregators: The bias–information–noise (BIN) model. *International Journal of Forecasting*, *39*(1), 470–485. https://doi.org/10.1016/j.ijforecast.2021.12.010

Simon, H. A. (1996). *The sciences of the artificial* (Third edition). The MIT Press.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tannenbaum, D., Fox, C. R., & Ulkumen, G. (2017). Judgment extremity and accuracy under epistemic vs. Aleatory uncertainty. *Management Science*, *63*(2), 497–518. https://doi.org/10.1287/mnsc.2015.2344

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

Tetlock, P. E., Lu, Y., & Mellers, B. A. (2023). False dichotomy alert: Improving subjective-probability estimates vs. raising awareness of systemic risk. *International Journal of Forecasting*, *39*(2), 1021–1025. https://doi.org/10.1016/j.ijforecast.2022.02.008

Tetlock, P. E., & Mellers, B. A. (2014). Judging political judgment. *Proceedings of the National Academy of Sciences*, *111*(32), 11574–11575. https://doi.org/10.1073/pnas.1412524111

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Ungar, L., Mellers, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J., & Swift, S. (2012). The good judgment project: A large scale test of different methods of combining expert predictions. *AAAI Fall Symposium Series*.

Van den Broeke, M., De Baets, S., Vereecke, A., Baecke, P., & Vanderheyden, K. (2019). Judgmental forecast adjustments over different time horizons. *Omega*, *87*, 34–45. https://doi.org/10.1016/j.omega.2018.09.008

Wallsten, T. S., & Budescu, D. V. (1983). State of the art—encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*(2), 151–173. https://doi.org/10.1287/mnsc.29.2.151

West, M., & Harrison, J. (2006). *Bayesian forecasting and dynamic models*. Springer Science & Business Media.

Winkler, R. L. (1969). Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, *64*(327), 1073–1078. https://doi.org/10.1080/01621459.1969.10501037

Winkler, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science*, *40*(11), 1395–1405. https://doi.org/10.1287/mnsc.40.11.1395

Witkowski, J., Freeman, R., Vaughan, J. W., Pennock, D. M., & Krause, A. (2022). Incentive-compatible forecasting competitions. *Management Science*. https://doi.org/10.1287/mnsc.2022.4410

3.9 Appendices

3.9.1 Appendix 3.A Correspondence Between Our Conceptualization and Operationalization of Foresight

In the following, we will explain how our conceptualization of foresight as the ability to predict future states of the world accurately, where accuracy becomes continuously more important over time corresponds to the time-weighted Brier score (TWBS) using the prototypical forecasters in Figure 3.1 and Table 3.A1 as example:

Table 3.A1 Exemplary Ranking of Prototypical Forecasters in Figure 3.1

Forecaster	Day 1	Day 2	Day 3	Linear TWBS	Rank
A	1	1	1	0	1
B	0	0.5	1	0.5	2
C	1	0.5	0	1.17	3
D	0	0	0	2	4

Note. Values in columns 'Day X' indicate forecasts on the true category on the corresponding day X for a question with two outcome categories. The rank of the forecasters is based on the linear time-weighted Brier score (TWBS).

Forecaster A predicts the future state of the world with perfect accuracy over time (i.e., from the beginning to the resolution of the question). According to our definition of foresight, forecaster A possesses perfect foresight and receives the best TWBS possible.

Forecaster B predicts the future state of the world with perfect inaccuracy at the beginning of the question but improves that accuracy over time until forecaster B achieves perfect accuracy at the resolution of the question. Based on our definition of foresight, forecaster B possesses worse foresight than forecaster A as forecaster B provided inaccurate early predictions. Although the TWBS highly rewards forecaster B for making increasingly accurate predictions over time as accuracy becomes continuously more important over time, the TWBS still punishes forecaster B to some extent for the inaccurate early predictions. Therefore, forecaster B receives a worse TWBS than forecaster A.

Forecaster C predicts the future state of the world with perfect accuracy at the beginning of the question, but their accuracy deteriorates over time until forecaster C reaches perfect inaccuracy at the resolution of the question. According to our definition of foresight, forecaster C possesses worse foresight than forecaster B as forecaster C provided inaccurate late predictions, and accuracy becomes continuously more important over time. In other words, our reconceptualization allows us to rank forecasters B and C. Although the TWBS rewards forecaster C to some extent for making accurate early predictions, the TWBS heavily punishes forecaster C for the inaccurate late predictions. Therefore, forecaster C receives a worse TWBS than forecaster B.

Forecaster D predicts the future state of the world with perfect inaccuracy over time (i.e., from the beginning to the resolution of the question). According to our definition of foresight, forecaster D possesses the worst foresight and receives the worst possible TWBS.

3.9.2 Appendix 3.B Proof that the Time-Weighted Brier Score Is a Strictly Proper Scoring Rule

We define the Time-Weighted Brier Score (TWBS) as

$$TWBS_{iq} = \frac{\sum_{t=1}^{T_q} w_t \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} w_t}$$

$$\text{with } \sum_{c=1}^{C_q} (f_{iqtc} - o_{qc})^2 = 1 - \frac{1}{C_q}$$

$$\text{if forecaster i made no forecast on question } q \text{ at time } t,$$

and weights $w_t > 0$ that are (strictly) monotonically increasing over time. It is immediately obvious that the TWBS is a scoring rule (cf. Gneiting and Raftery 2007). It remains to be shown that it is furthermore strictly proper.

For a scoring rule to be strictly proper, its expectation must be uniquely minimized by the "true probability." For instance, the Brier score is known to be strictly proper. This means that, if $\boldsymbol{f}_{iqt} = (f_{iqt1}, \ldots, f_{iqtC})$ is a vector of probabilities representing forecaster *i*'s true belief about the event $\boldsymbol{o}_q = (o_1, \ldots, o_C)$ at time *t*, then the expected Brier score of a candidate forecast $\boldsymbol{f}'_{iqt} = (f'_{iqt1}, \ldots, f'_{iqtC})$ under the forecasters' true belief $\boldsymbol{f}_{iqt}$ is $E_{\boldsymbol{f}_{iqt}}[\sum_{c=1}^{C_q} (f'_{iqtc} - o_c)^2]$ and this expectation is minimized when $\boldsymbol{f}_{iqt} = \boldsymbol{f}'_{iqt}$; that is, when the forecaster reports their true belief. Therefore, so long as the forecaster seeks to minimize the Brier score, they will report their true beliefs.

To extend this discussion to our $TWBS_{iq}$, collect all forecaster *i*'s predictions for question q into a vector $\boldsymbol{f}_{iq} = (\boldsymbol{f}_{iq1}, \ldots, \boldsymbol{f}_{iqT_q})$. The Time-Weighted Brier Score $TWBS_{iq}$ is a weighted-average of T_q individual Brier scores. Given that the weights are non-random, the expectation of $TWBS_{iq}$ for some candidate prediction $\boldsymbol{f}'_{iq} =$

$(\boldsymbol{f}'_{iq1},\ldots,\boldsymbol{f}'_{iqT_q})$ under the forecaster's true belief $\boldsymbol{f}_{iq}$ is a weighted average of the expectations of the individual Brier scores:

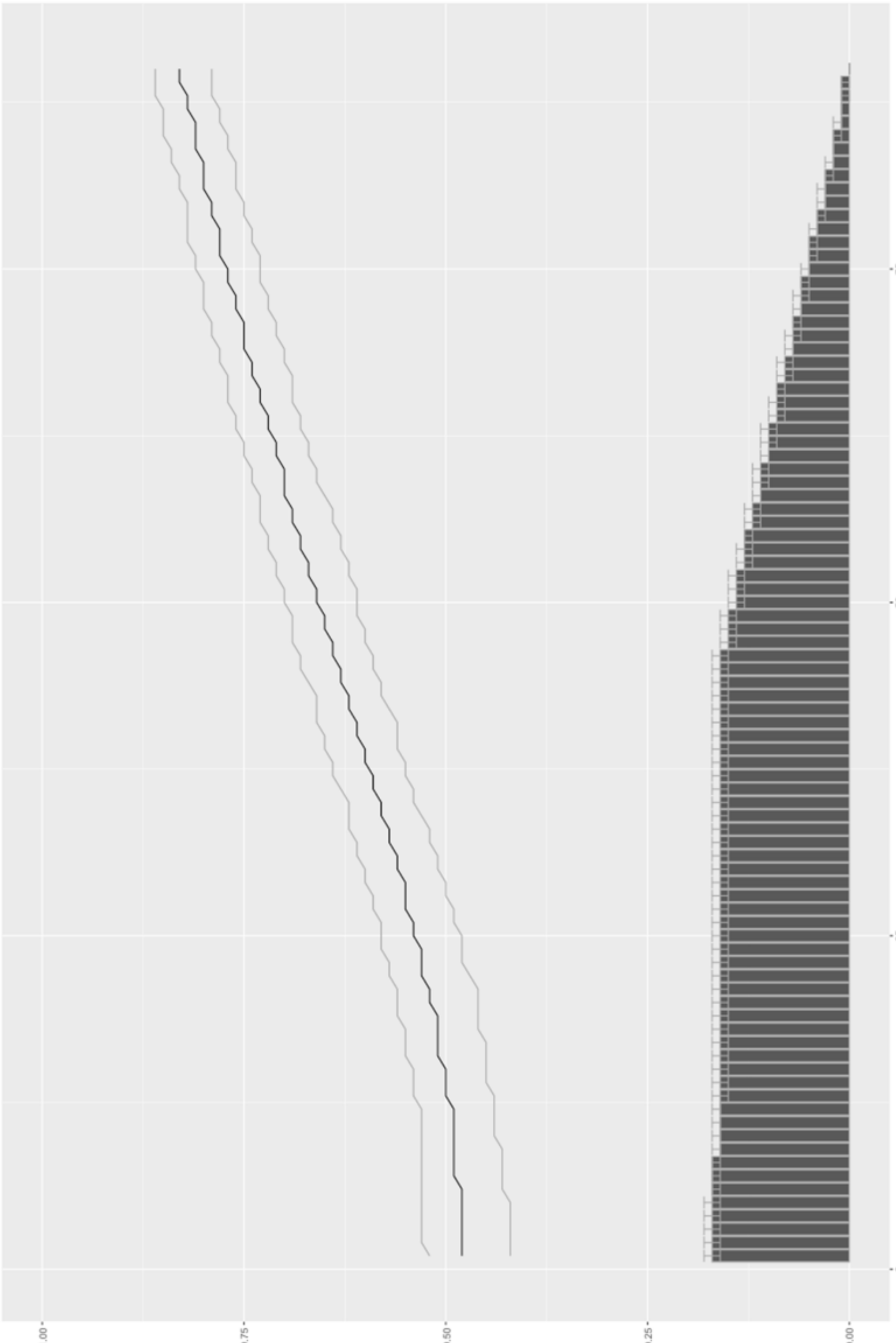
$$E_{\boldsymbol{f}_{iq}}[TWBS_{iq}] = \frac{\sum_{t=1}^{T_q} w_t E_{\boldsymbol{f}_{iqt}}[\sum_{c=1}^{C_q}(f'_{iqtc} - o_{qc})^2]}{\sum_{t=1}^{T_q} w_t}.$$

By our earlier argument, the individual Brier score at time *t* is minimized when $\boldsymbol{f}_{iqt} = \boldsymbol{f}'_{iqt}$ for all $t = 1,\ldots,T_q$. Given that a weighted average is uniquely minimized when all the individual components take on their smallest values, $E_{\boldsymbol{f}_{iq}}[TWBS_{iq}]$ is minimized uniquely when $\boldsymbol{f}_{iq} = \boldsymbol{f}'_{iq}$, showing that our TWBS is a strictly proper scoring rule.

Remark 1: The above argument is not unique to the Brier score, but in fact holds for any strictly proper scoring rule such as the log or spherical scoring rules. **Remark 2:** By our procedure, the forecaster receives the score $1 - 1/C_q$ at a time point if they do not report a forecast. The forecaster may fail to report a forecast for various reasons. For instance, they may be absent from the forecasting platform, had forgotten that forecasting for a particular question had already begun, etc. If the forecaster did not actively choose to abstain from forecasting, then there was no choice involved, and it does not make sense to discuss incentives or properness of scoring rule. If, however, the forecaster actively chose to abstain, the properness of the Brier score shows that their true belief coincided with the uniform forecast; else they would have not abstained and would have instead reported a forecast with a lower expected score than that of the uniform forecast. Therefore, our TWBS score ensures that whenever the forecaster does make a choice, they have the proper incentives to report truthfully.

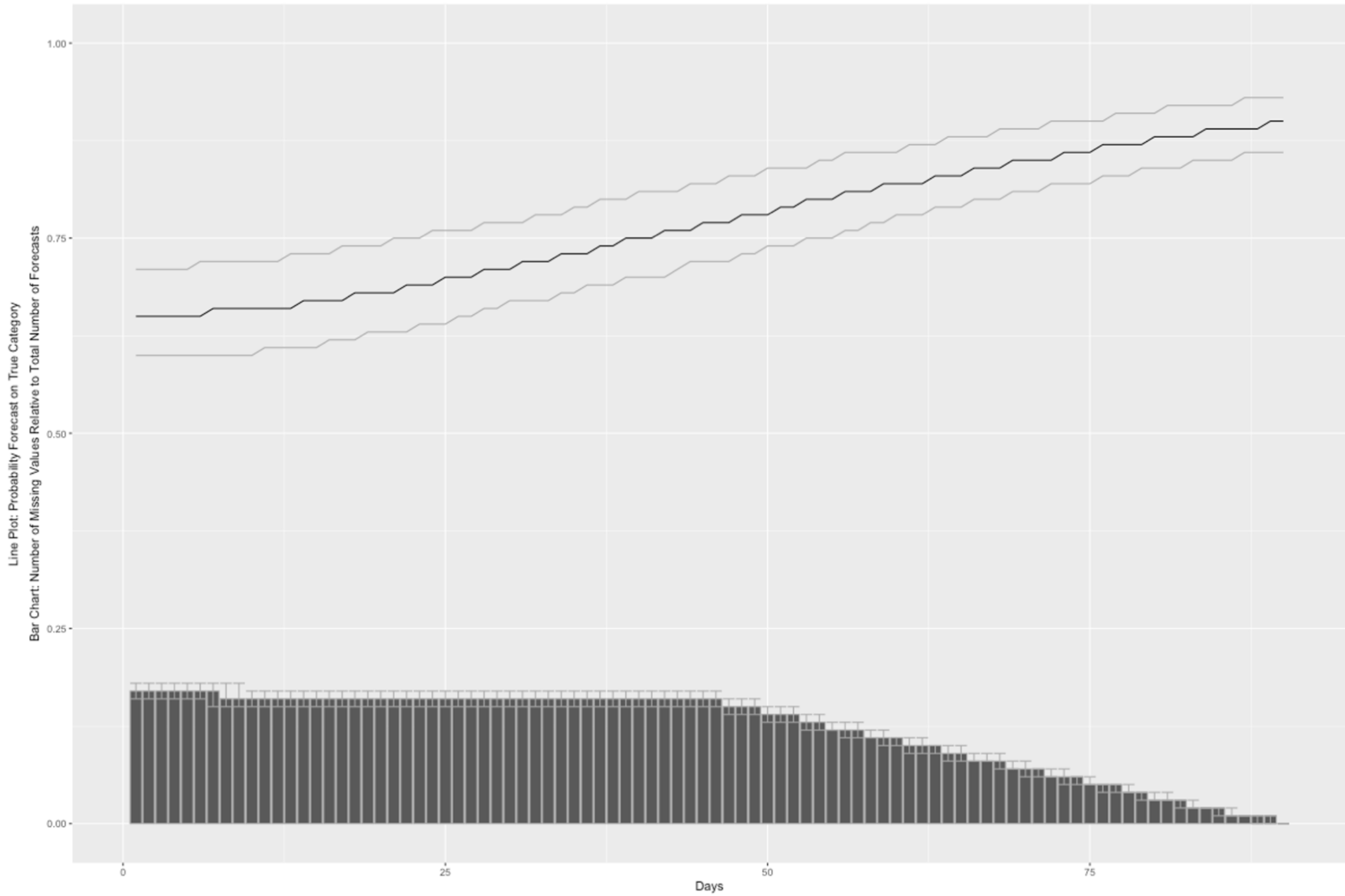
3.9.3 Appendix 3.C Exemplary Simulation Data for Various Parameter Values

Figure 3.C1 Exemplary Simulation Data for the Baseline (100 Trials)



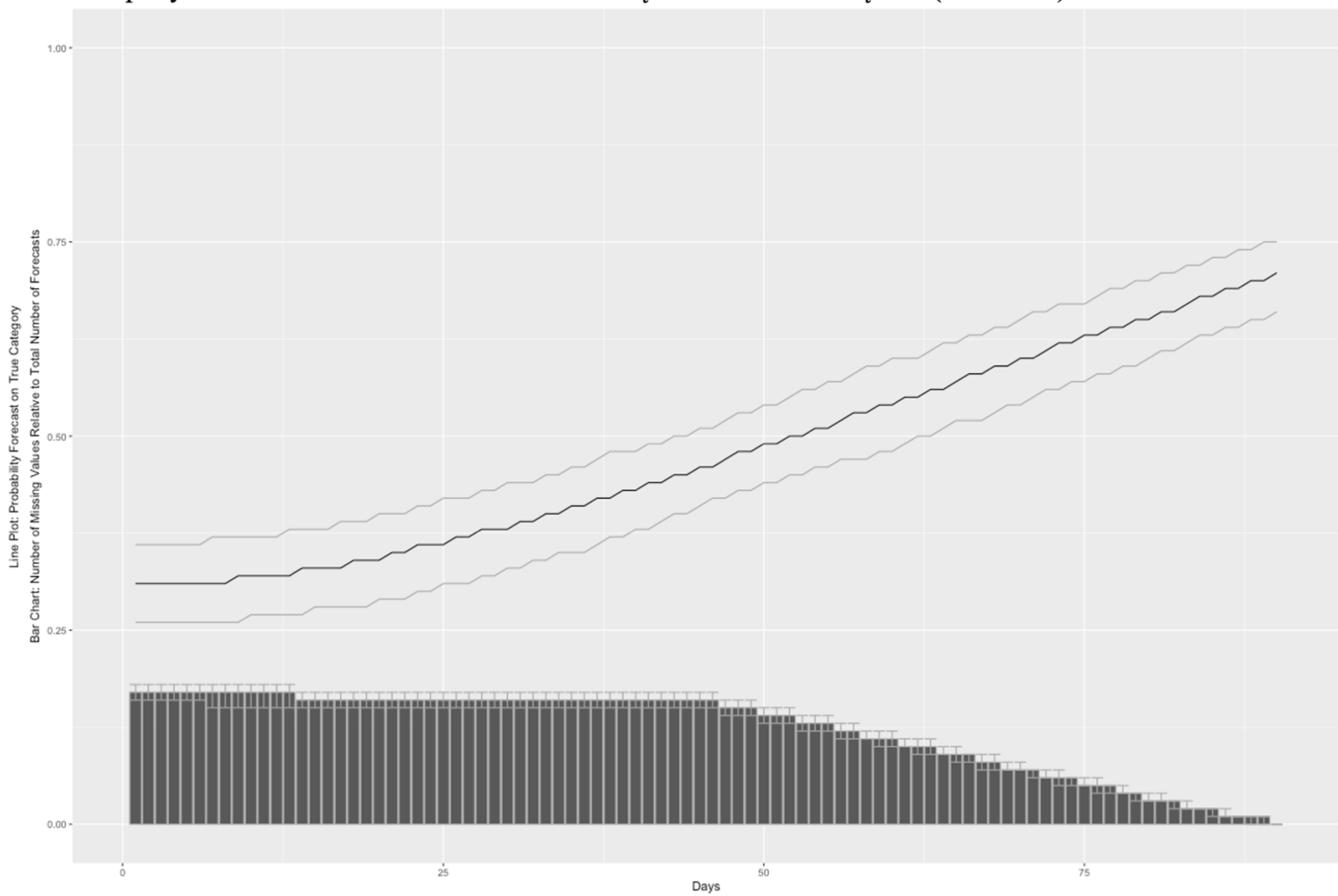
Note. The baseline simulation values correspond to the values described in section 5.1 with means $\sigma_i = -1$ (low strategic forecast delay), $\mu_i = -2$ (low miscellaneous forecast delay), and a linear signal trajectory. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C2 Exemplary Simulation Data for the Baseline with Mean $\theta_i = 1$ instead of $\theta_i = 0$ (100 Trials)



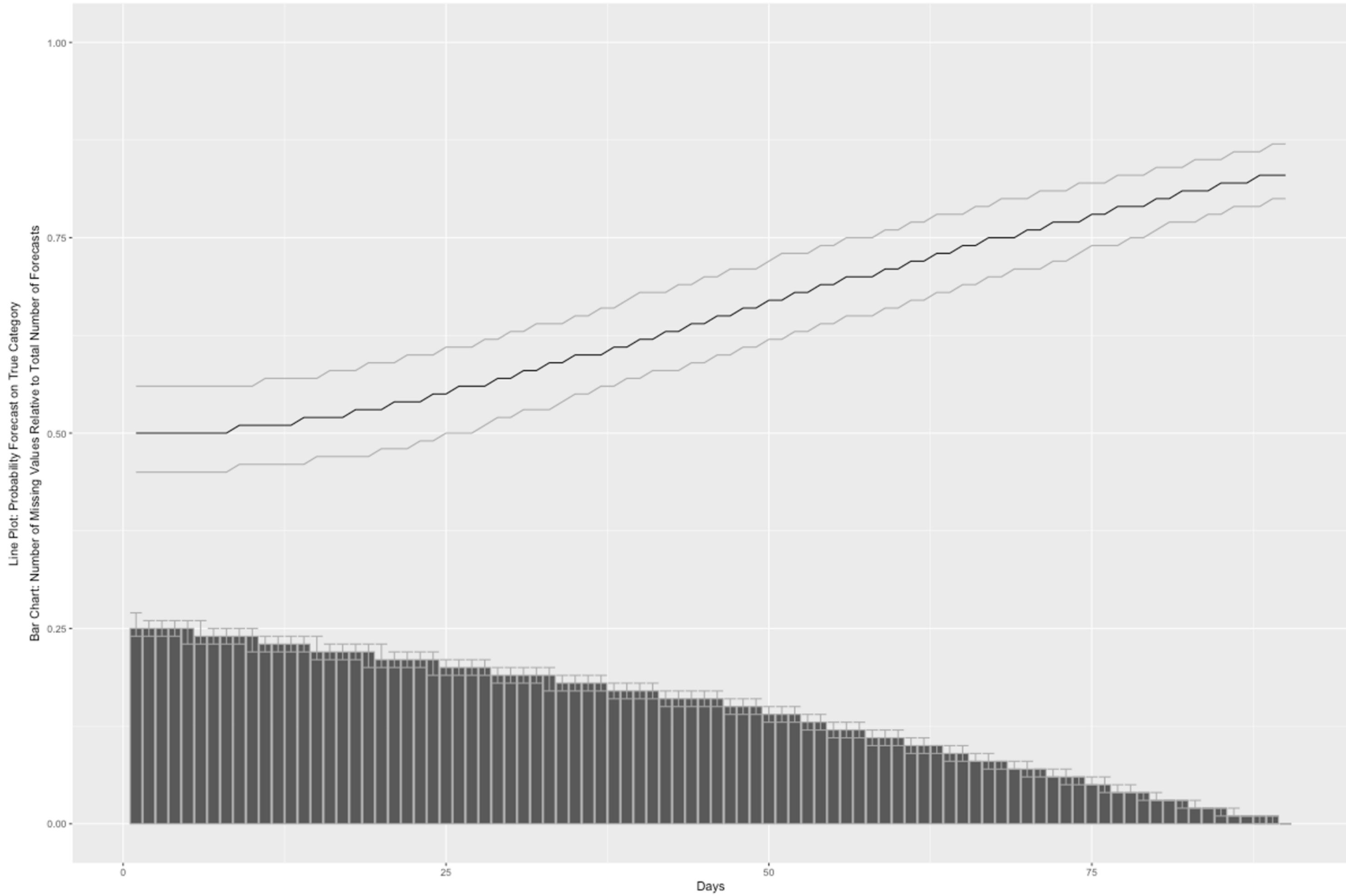
Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C3 Exemplary Simulation Data for the Baseline with $\theta_i = -1$ instead of $\theta_i = 0$ (100 Trials)



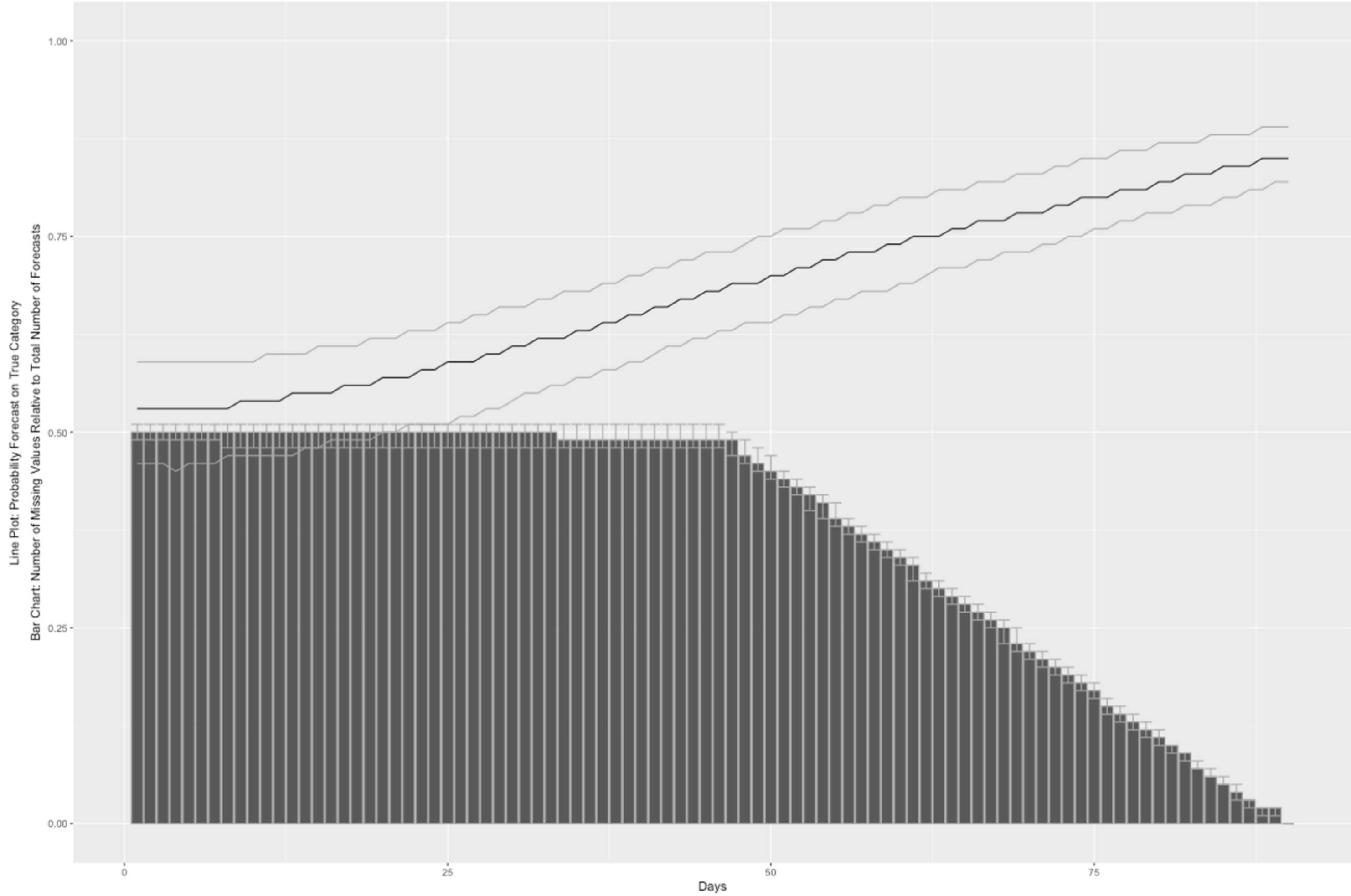
Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C4 Exemplary Simulation Data for the Baseline with $\mu_i = -1$ instead of $\mu_i = -2$ (100 Trials)



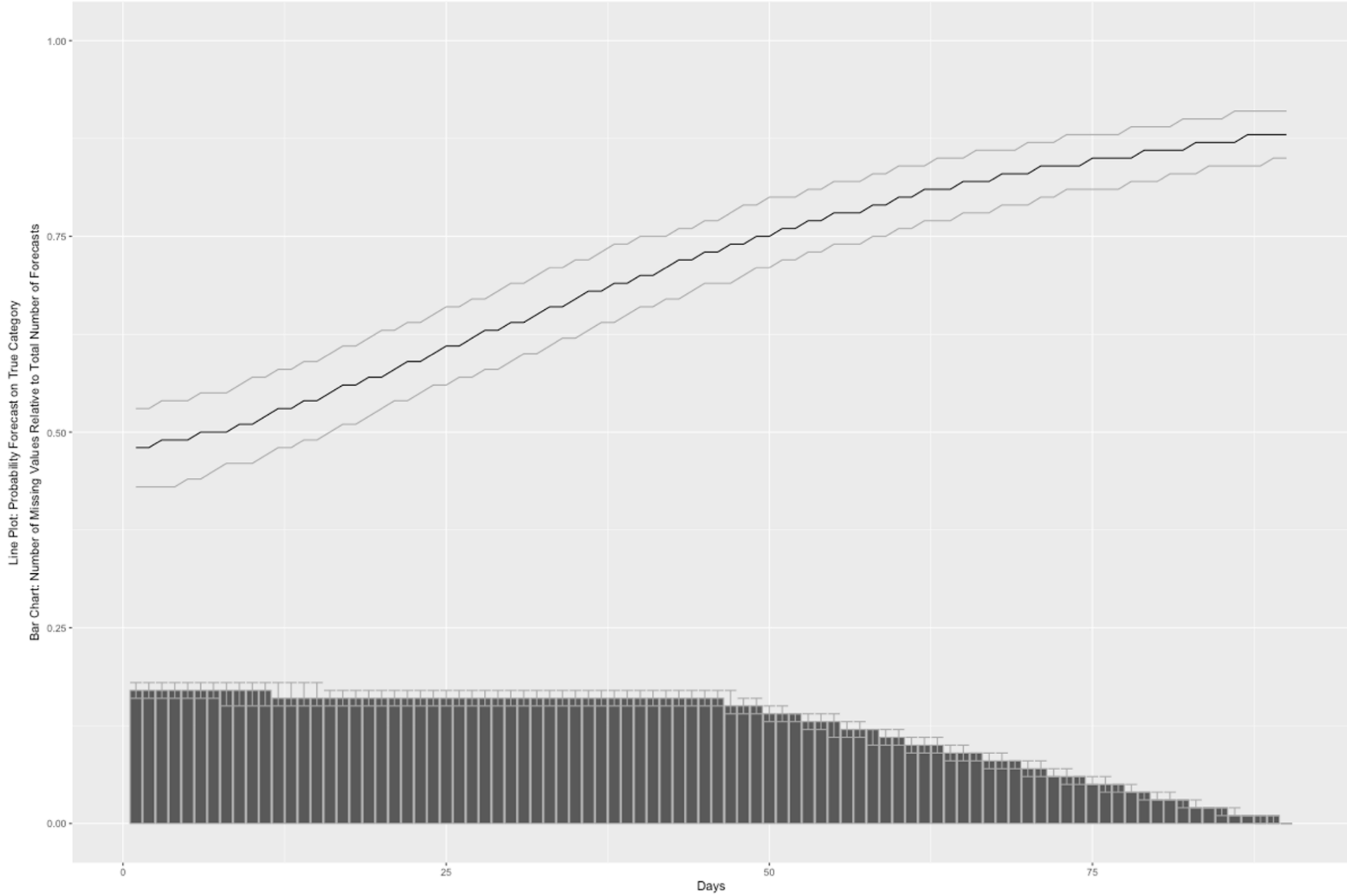
Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C5 Exemplary Simulation Data for the Baseline with $\sigma_i = 0$ instead of $\sigma_i = -1$ (100 Trials)



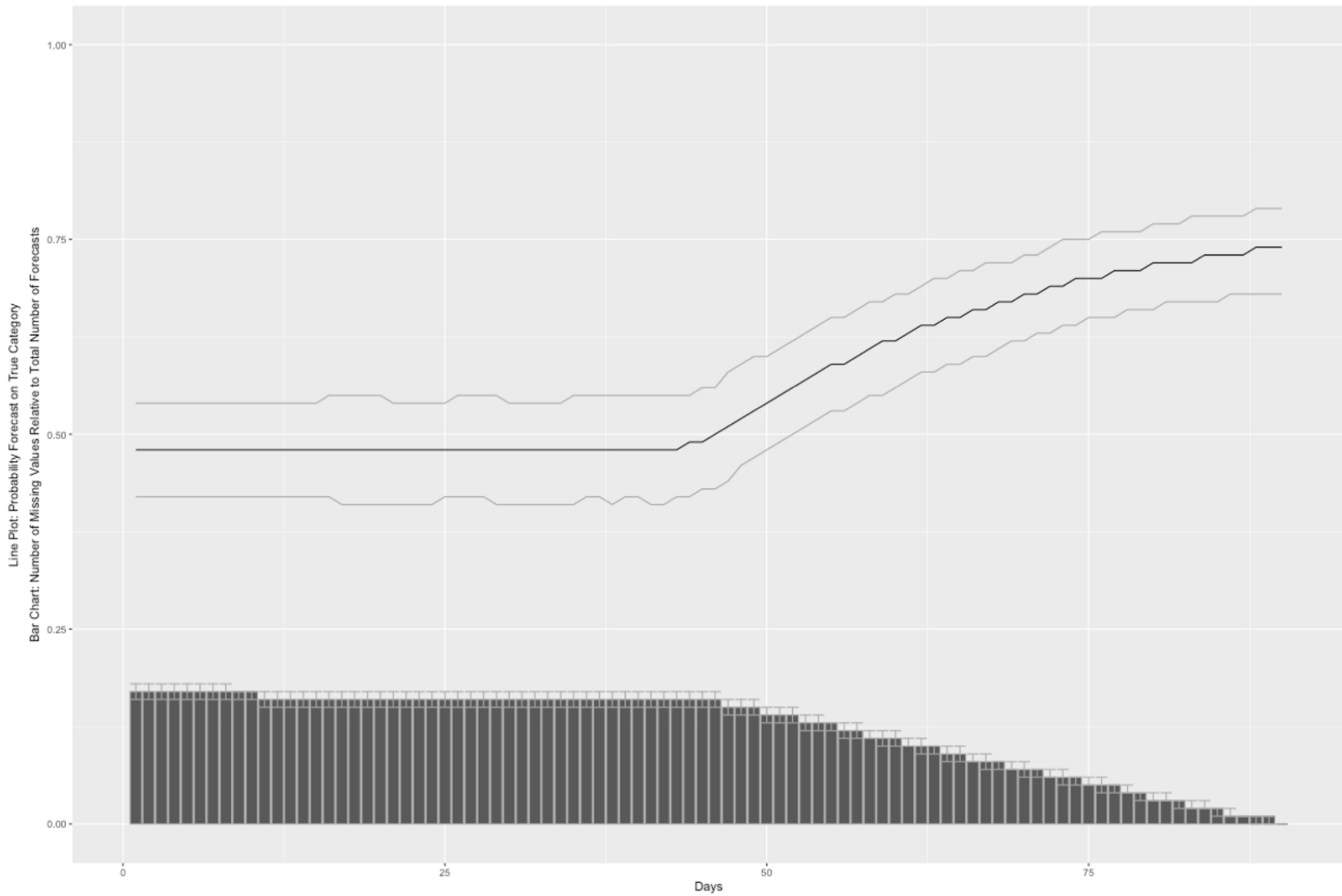
Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C6 Exemplary Simulation Data for the Baseline with a Square Root instead of Linear Signal Trajectory (100 Trials)



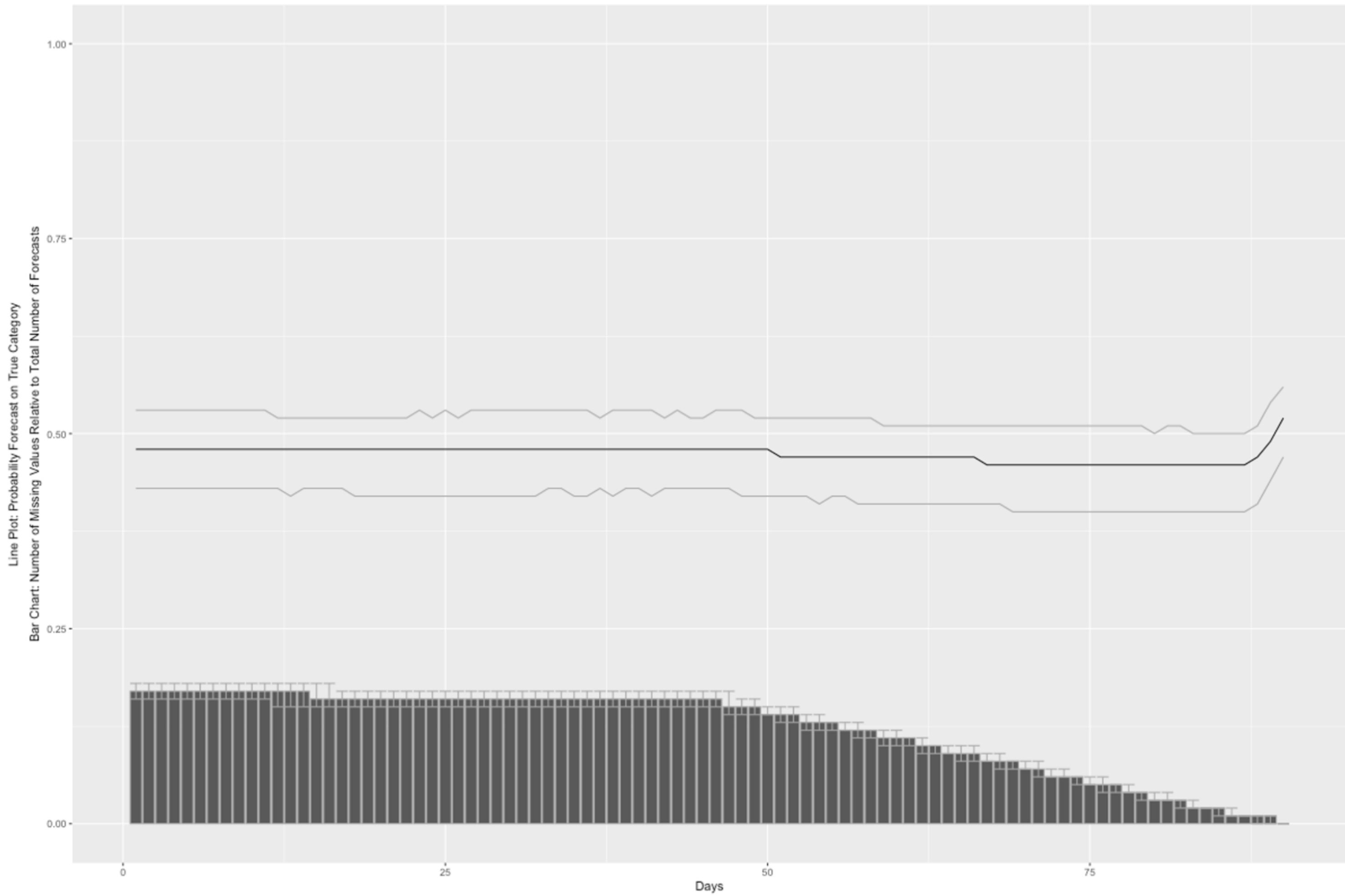
Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

Figure 3.C7 Exemplary Simulation Data for the Baseline with a Logistic instead of Linear Signal Trajectory (100 Trials)



Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

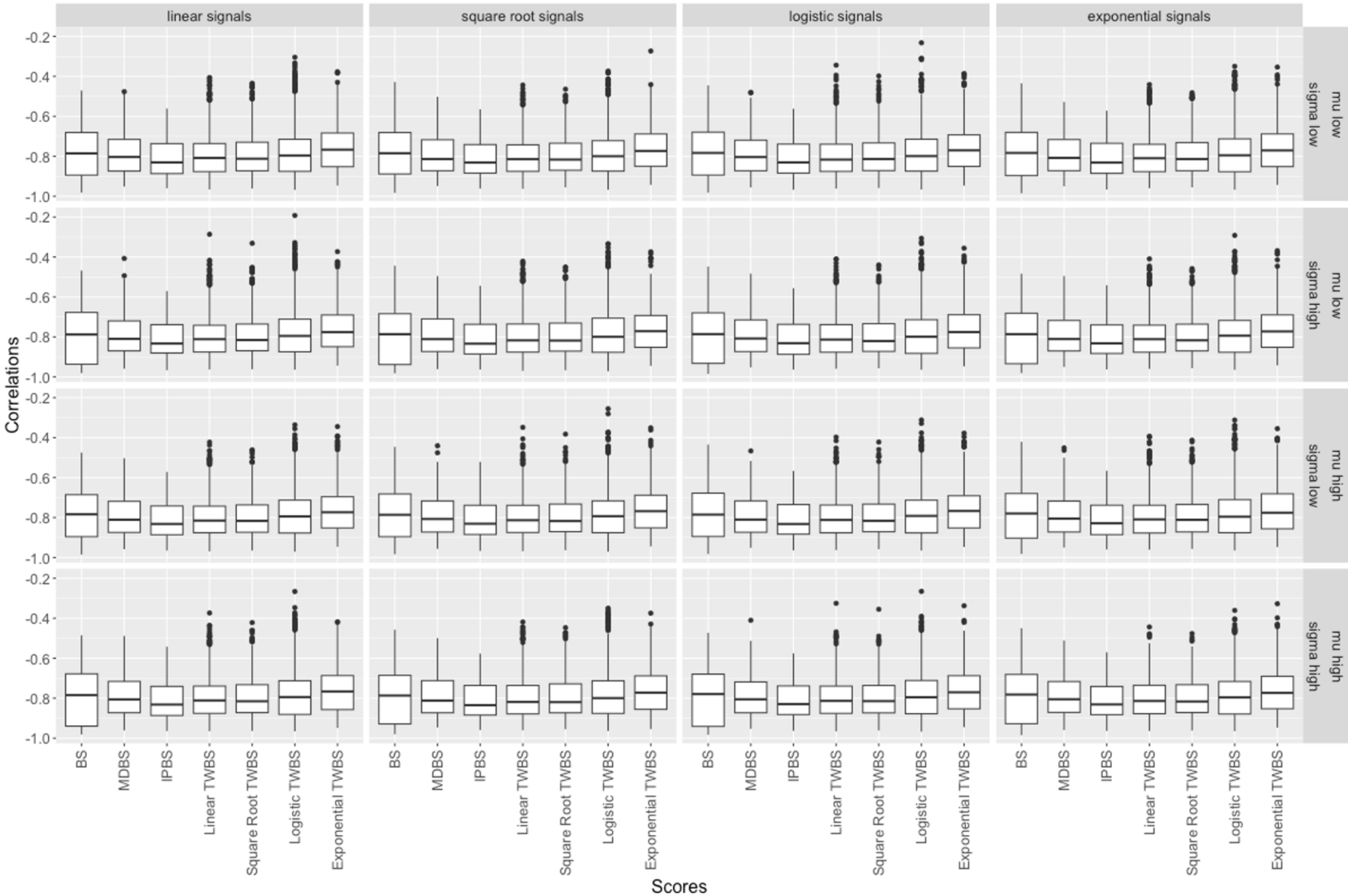
Figure 3.C8 Exemplary Simulation Data for the Baseline with an Exponential instead of Linear Signal Trajectory (100 Trials)



Note. The line plot shows the average probability forecast on the true category and its corresponding 90% confidence interval. The bar chart shows the average number of missing values relative to the total number of possible forecasts and the corresponding 90% confidence intervals.

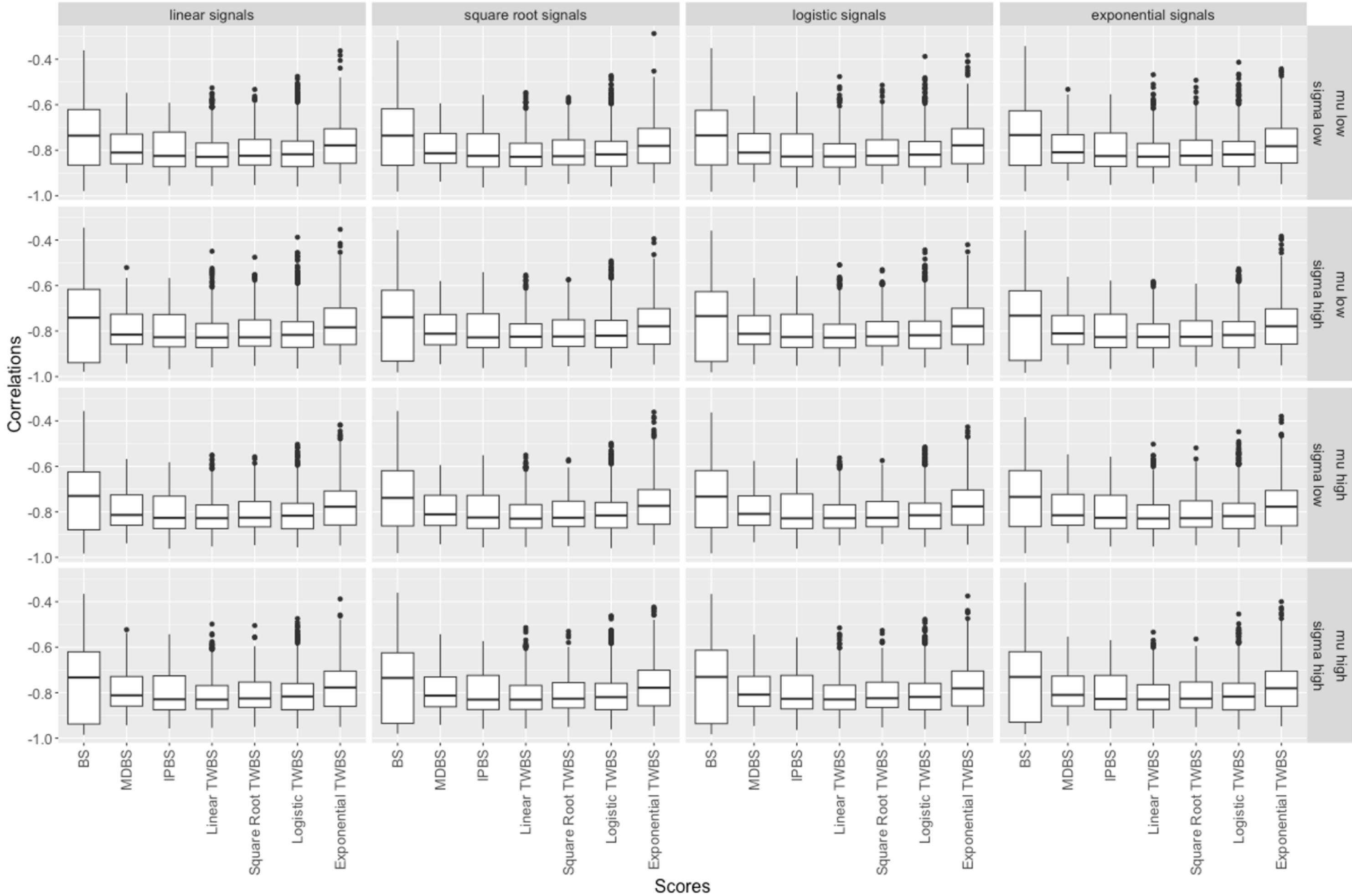
3.9.4 Appendix 3.D Robustness Checks of Simulation Study

Figure 3.D1 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition Assuming the Correlations between the Person Parameters to be 0



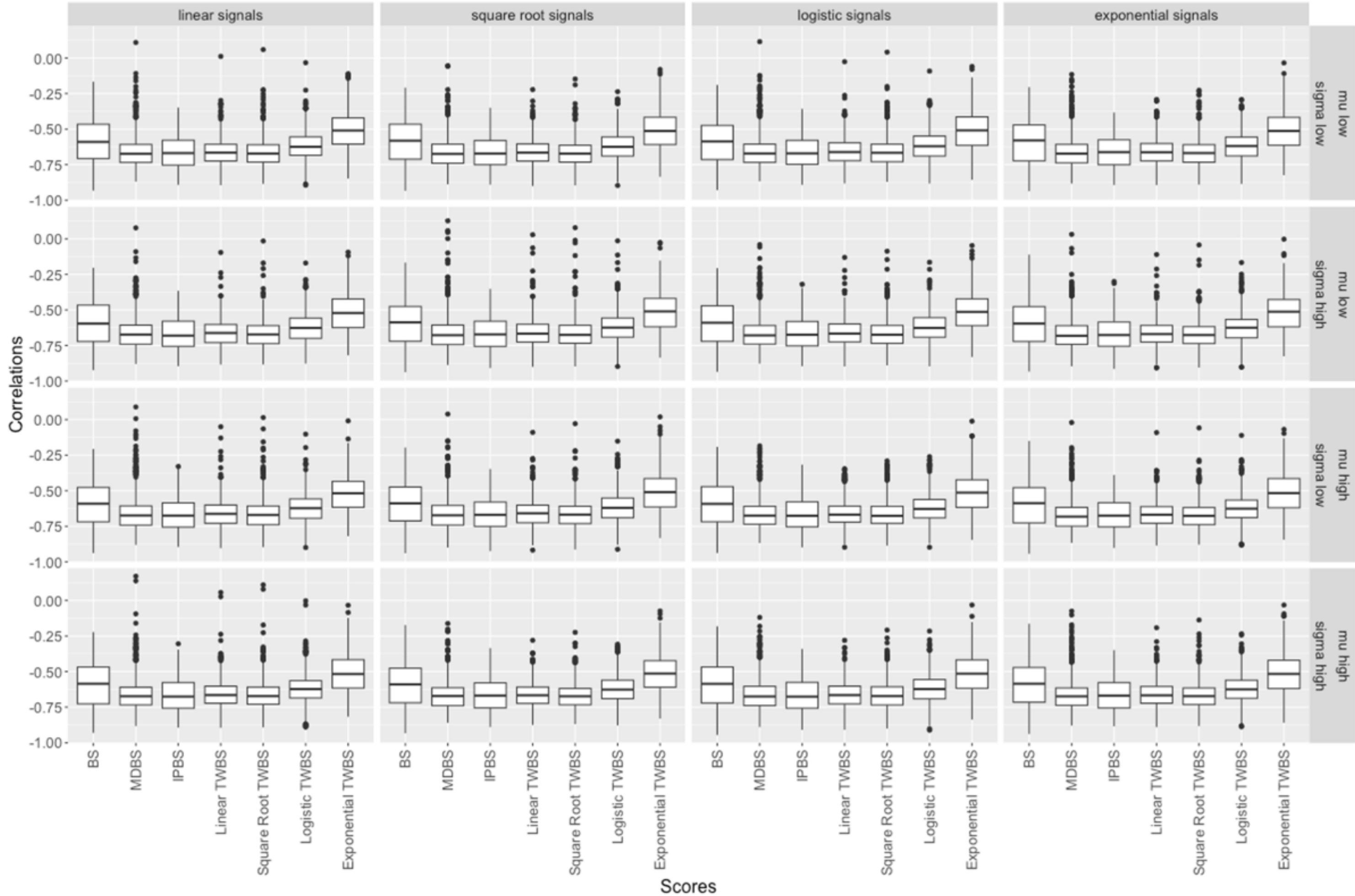
Note. All scoring rules measured the true foresight of forecasters θ_i equally well in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

Figure 3.D2 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition Assuming the Correlations between the Person Parameters to be 0.3



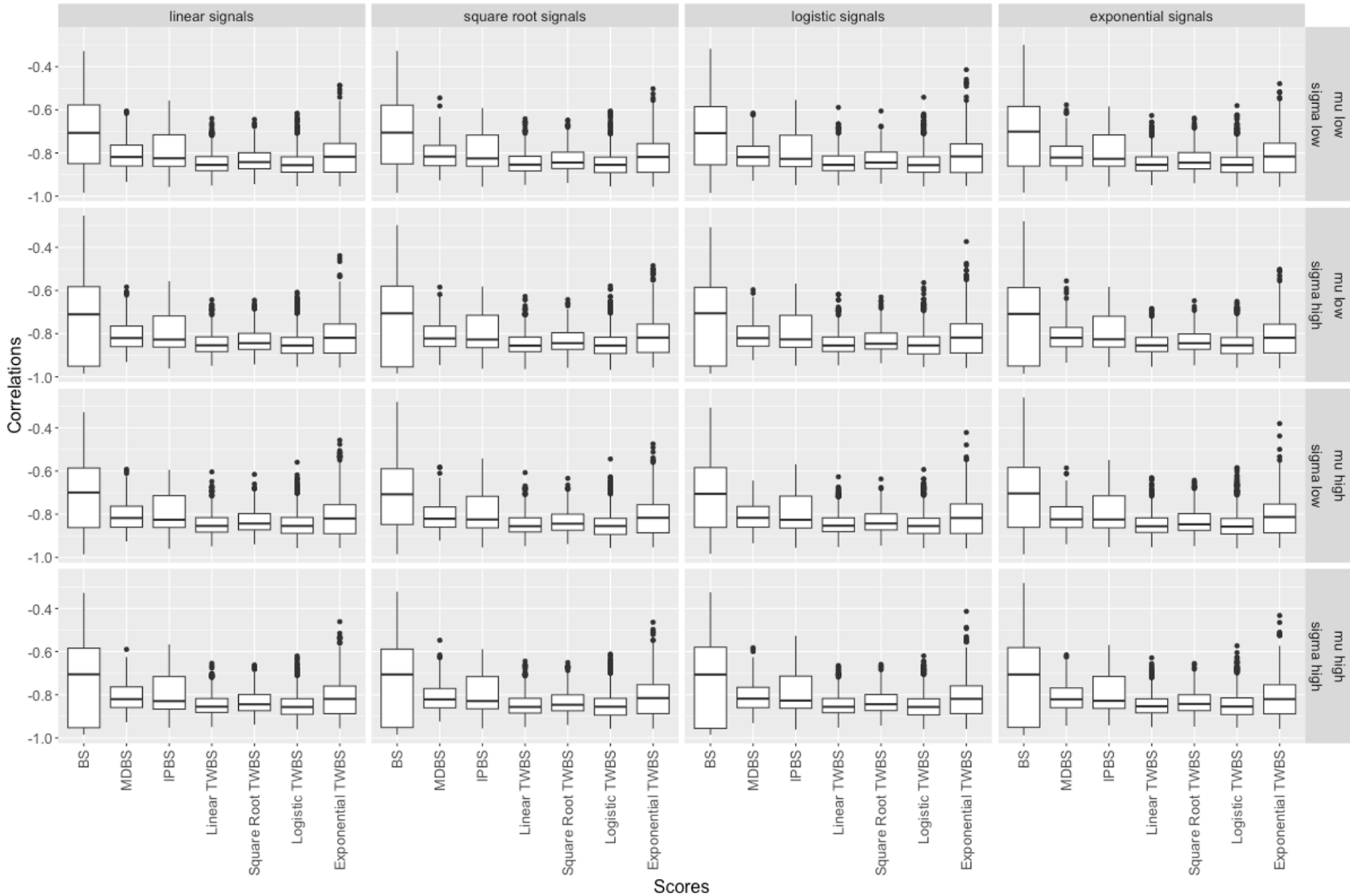
Note. The linear, square root, and logistic TWBS measured the true foresight of forecasters θ_i best in terms of consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

Figure 3.D3 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition for Forecasters Forecasting One instead of 10 Questions



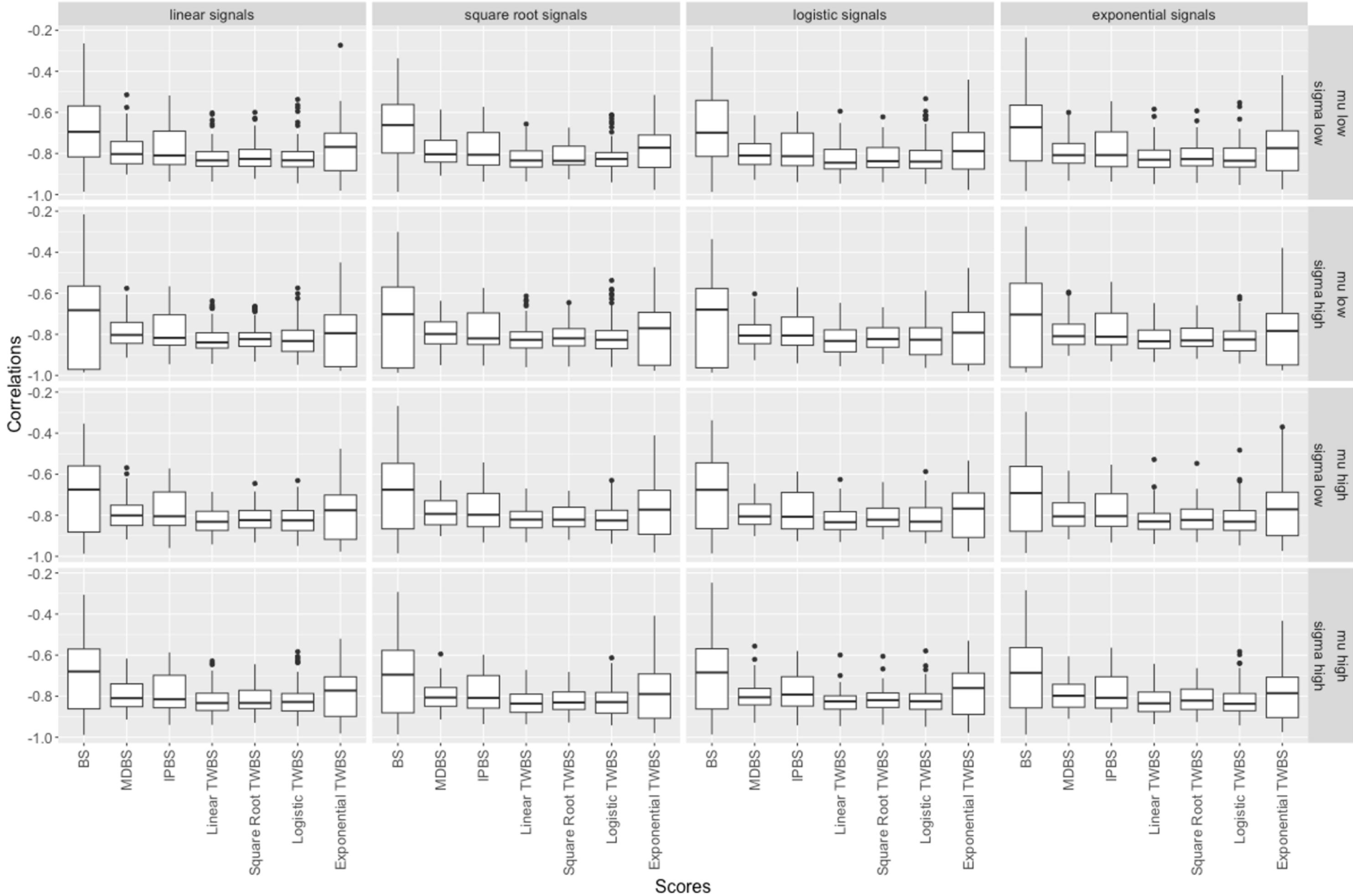
Note. All scoring rules with the exception of the BS, logistic TWBS, and exponential TWBS measured the true foresight of forecasters θ_i equally well in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

Figure 3.D4 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition for Forecasters Forecasting 20 instead of 10 Questions



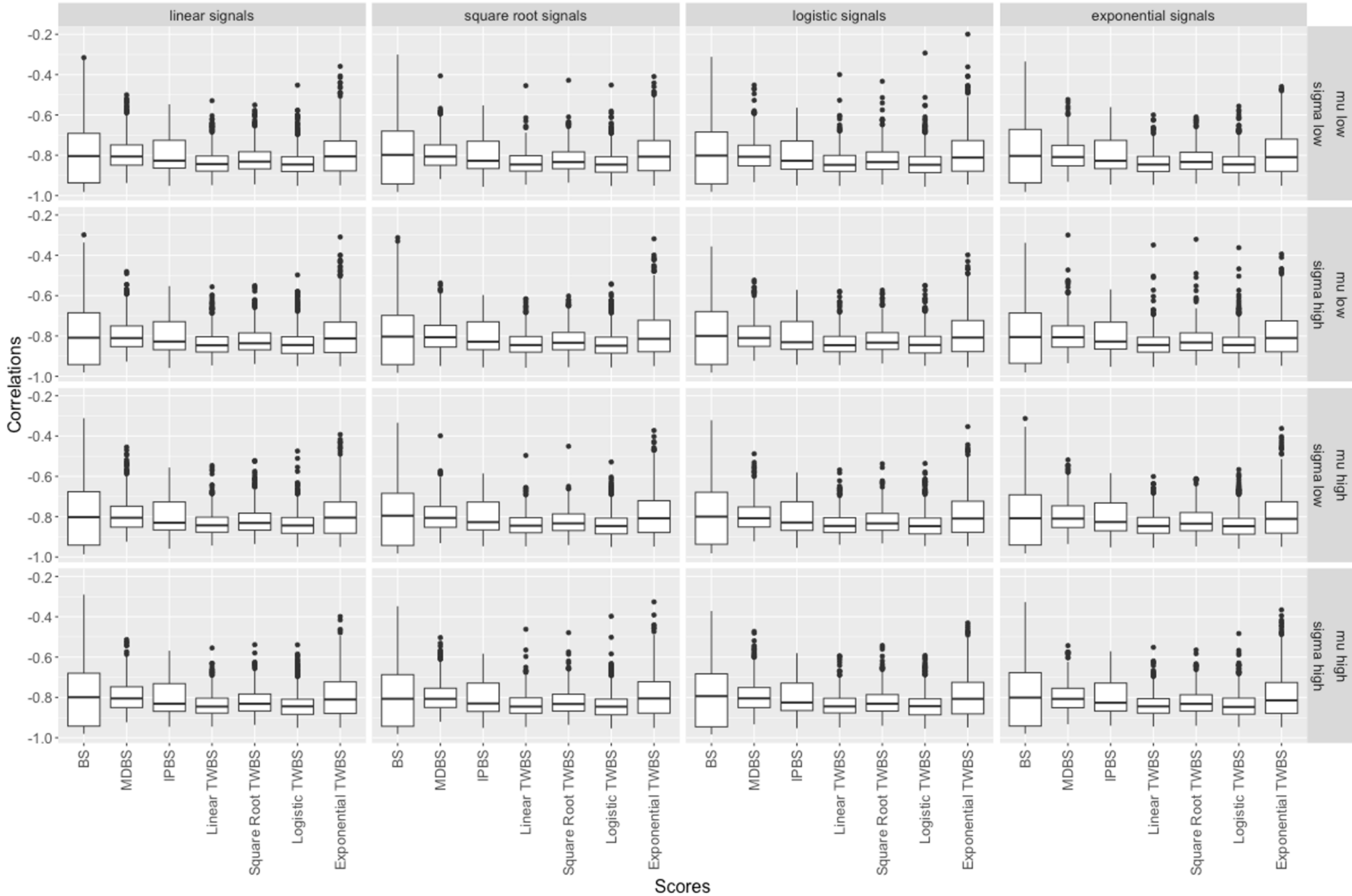
Note. The linear, square root, and logistic TWBS measured the true foresight of forecasters θ_i best in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

Figure 3.D5 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition for Questions Forecastable 365 instead of 90 Days



Note. The linear, square root, and logistic TWBS measured the true foresight of forecasters θ_i best in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

Figure 3.D6 Boxplots of the Correlations between the True Foresight of Forecasters θ_i and Each Scoring Rule Across All 1000 Trials by Simulation Condition for Ten Percent of Questions with a Decreasing Quality of Signals Until the Final Day of the Forecast



Note. The linear, square root, and logistic TWBS measured the true foresight of forecasters θ_i best in terms of average accuracy and consistency as indicated by the boxplots. More negative correlations indicate that a scoring rule measured the true foresight of forecasters θ_i more accurately. Linear signals = τ_{lin}; square root signals = τ_{sqrt}; logistic signals = τ_{logit}; exponential signals = τ_{exp}; sigma low: $\sigma_i = -1$; sigma high: $\sigma_i = 0$; mu low: $\mu_i = -2$; mu high: $\mu_i = -1$.

3.9.5 Appendix 3.E Analytical Approach in the Empirical Study

We computed all scoring rules based on the forecasting data so that we obtained each score for every question predicted by a forecaster. Then, we transformed each score to the real line by dividing each score by 2 (scale maximum), setting values of 0 and 1 to 0.0001 and 0.9999, applying the probit transformation, and standardizing the scores to ensure the comparability of the results across the scores. We used these unbounded standardized probit scores in all of our analyses because we used regression models assuming unbounded dependent variables. Furthermore, the resulting values represent z-values, which facilitate the interpretation of the scores and ensure the comparability of the effects by standardized predictors on the scores.

We merged these data with the data on the questions and empirical measures. When several measures for a psychological construct were collected during a year or across years, we computed the average of the percentage correct responses on each measure to obtain an aggregate measure for the corresponding construct (see Table 3.E1 for details). For example, numeracy was measured with five questions in years 1 and 2, whereas numeracy was operationalized with the Berlin Numeracy Test in years 3 and 4. Thereby, we obtained one measurement per construct for each forecaster if data on the construct was collected from the forecaster. We standardized the aggregate measures for the analyses to ensure the comparability of the effect sizes across predictors and scores in the analyses.

We fitted ordinary least squares (OLS) regressions to obtain a measure of linear association between the transformed scoring rules by regressing the scores on each other (e.g., the MDBS on the BS). Furthermore, we fitted OLS regressions to obtain a measure of linear association between each score and the empirical measures listed in Table 3.E1 by regressing each score on each empirical measure. We computed heteroscedasticity-consistent standard errors clustered by forecasters to account for the nested data structure resulting from the fact that forecasters usually forecast more than one question (Liang & Zeger, 1986). Furthermore, we included question and group fixed effects to account for differences in the scores attributable to forecasters deciding to forecast different questions and their assignment to different experimental groups during the tournament. Note that we did not include year fixed effects, as differences between years are indirectly controlled for by the question fixed effects as each question belonged to a specific year. We computed linear models with one predictor, because we did not have a theoretical justification for nonlinear relationships and the estimated regression weight is conceptually similar to the Pearson correlation coefficient, which is the standard measure of linear association for simple data structures. Due to the dependencies in the data, however, the Pearson correlation coefficient would yield biased estimates. All analyses were done in-sample.

Normal Q-Q plots of most models indicated that the errors were non-normally distributed, with heavier tails indicating that errors larger than $|2|$ deviated from the normal distribution. These heavy tails resulted from the fact that forecasters disproportionately often achieved extreme scores of 0 and 2. This suggests that forecasters likely cared about their performance relative to other forecasters and consequently, may have exaggerated their forecasts towards 0 and 1 to win the tournament (Lichtendahl & Winkler, 2007). Therefore, we implemented a nonparametric cluster bootstrap approach to compute 99.92% studentized *t* confidence intervals based on 10,000 bootstrap replications as measures of uncertainty and significance of the estimated parameters (Efron & Tibshirani, 1994). We computed 99.92% confidence intervals, which correspond to the Bonferroni corrected alpha level $\alpha \approx 0.08\%$ for 29 two-sided tests per score, to prevent an inflation of the type I error rate. For consistency, we also applied this confidence level to the regression coefficients, capturing the relationships between the scores. The cluster bootstrap approach samples the data at the cluster (i.e., forecaster) level thereby preserving the dependency structures within each sampled cluster resulting from the nested data structure. We opted for studentized *t* confidence intervals because they are pivotal, which means that they do not depend on unknown parameters (Davison & Hinkley, 1997).[16] We set the number of bootstrap replications to 10,000, because this allowed us to approximate the tails of the studentized parameter distributions reasonably well. Furthermore, this avoided the use of the most extreme bootstrapped parameters for the construction of the 99.92% studentized confidence intervals. The bootstrap replications were not constructed from the full data but from data that did not contain any missing values in the predictor of interest of the corresponding model. For each bootstrap sample, we saved the regression coefficient and standard error estimated based on the regression specification described above, which are required for the computation of the studentized *t* confidence intervals.

Nonparametric bootstrap approaches assume that the empirical distribution approximates the population distribution and the bootstrapped samples correspond to the empirical distribution (Davison & Hinkley, 1997). Although the first assumption cannot be tested, the large sample size suggests that the empirical distribution should

[16] We did not use bias-corrected and accelerated (BCa) confidence intervals as this involves jackknife estimation in Stata, which was computationally too costly due to the large dataset.

approximate the population distribution of geopolitical forecasters well. Even though it was practically unfeasible to check the distributions of the variables of each bootstrap sample, we compared tables and histograms of the variables based on independent randomly drawn samples with replacement to tables and histograms of these variables based on the empirical distribution to evaluate the second assumption. The tables and histograms indicated that the distributions corresponded to each other.

Table 3.E1. Overview and Descriptive Statistics of the Scores, Treatment Groups, and Empirical Measures of the Empirical Application to the Geopolitical Forecasting Tournaments

	N	*M*	*SD*	*Min*	*Max*	*Range*
Scores						
BS	414168	0.37	0.48	0.00	2.00	2.00
MDBS	414168	0.41	0.40	0.00	2.00	2.00
IPBS	414168	0.43	0.38	0.00	2.00	2.00
Linear TWBS	414168	0.68	0.60	0.00	2.00	2.00
Square Root TWBS	414168	0.76	0.61	0.00	2.00	2.00
BS*	414168	0.00	1.00	-1.91	3.94	5.85
MDBS*	414168	0.00	1.00	-3.62	6.02	9.64
IPBS*	414168	0.00	1.00	-3.89	6.21	10.10
Linear TWBS*	414168	0.00	1.00	-2.83	3.79	6.62
Square Root TWBS*	414168	0.00	1.00	-3.01	3.75	6.77
Treatment Groups						
Training	45203					
Teaming	131566					
Tracking	25163					
Actively Open-Minded Thinking	240480	0.00	1.00	-8.51	1.82	10.34
Analytic Thinking	157899	0.00	1.00	-6.94	2.07	9.01
Cognitive Reflection Test	257882	0.00	1.00	-2.56	0.89	3.45
Cognitive Reflection Test Extended	262028	0.00	1.00	-3.96	1.08	5.05
Conscientiousness	205474	0.00	1.00	-4.04	2.02	6.06
Extraversion	194555	0.00	1.00	-2.53	2.61	5.13
Extrinsic Motivation	196092	0.00	1.00	-0.99	2.54	3.53
Fox-Hedgehog	205433	0.00	1.00	-4.38	4.45	8.83
Grit	254112	0.00	1.00	-3.68	2.43	6.11
Intrinsic Motivation	196092	0.00	1.00	-6.45	1.23	7.68
Intuitive Thinking	157199	0.00	1.00	-4.37	3.08	7.45
Need for Closure	188540	0.00	1.00	-3.55	3.03	6.58
Need for Cognition	287379	0.00	1.00	-4.35	1.97	6.32
Numeracy	302129	0.00	1.00	-2.76	0.86	3.62
Openness	205433	0.00	1.00	-6.38	1.79	8.17
P-Beauty Contests	194838	0.00	1.00	-3.50	3.33	6.83

Prosocial Motivation	196092	0.00	1.00	-3.01	1.65	4.67
Raven Progressive Matrices	300575	0.00	1.00	-3.11	1.54	4.65
Shipley-2 Abstraction	195662	0.00	1.00	-4.65	1.84	6.50
Shipley-2 Vocabulary	195711	0.00	1.00	-13.57	1.14	14.71
Update Frequency	414168	0.00	1.00	-0.27	46.16	46.44
Update Magnitude	172338	0.00	1.00	-0.85	4.89	5.75
Working Memory	197761	0.00	1.00	-4.55	0.91	5.47
Working Memory Math	197761	0.00	1.00	-6.12	0.55	6.67
Working Memory Word	197761	0.00	1.00	-2.62	0.89	3.51

Note. N = number of scored questions; M = mean; SD = standard deviation; *Min* = minimum; *Max* = maximum. * indicates probit transformed standardized scores. Control group contains 41,315 observations. Political Knowledge represents an aggregate measure consisting of 35 items (year 1), 50 items (year 2), 55 items (year 3), 15 multiple choice questions (years 3 and 4), and 15 items of general political knowledge (year 4). Numeracy represents an aggregate measure consisting of five numeracy questions (years 1 and 2) and the Berlin Numeracy Test (years 3 and 4). Percentage correct scores on the measures were averaged within each year and then averaged across all years to obtain one value per measure for each forecaster. See online supplement of the data available at https://dataverse.harvard.edu/dataverse/gjp for detailed information on the measures. All empirical measures are z-standardized and represent probit values.

3.9.6 Appendix 3.F Precise Quantitative Results

Table 3.F1 Precise Quantitative Results of the OLS Linear Regressions and the 99.92% Studentized t Bootstrap Confidence Intervals

Single Predictor	BS				MDBS				IPBS				Linear TWBS				Square Root TWBS			
	b	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.
Scores																				
MDBS	0.66	0.65	0.67	*																
IPBS	0.63	0.61	0.64	*	0.94	0.94	0.94	*												
Linear TWBS	0.69	0.67	0.70	*	0.94	0.93	0.95	*	0.99	0.98	1.00									
Square Root TWBS	0.66	0.65	0.68	*	0.95	0.94	0.95	*	1.00	1.00	1.00		0.99	0.99	0.99	*				
Treatment Groups																				
Training	-0.06	-0.11	-0.01	*	-0.01	-0.05	0.03		0.00	-0.04	0.04		-0.01	-0.05	0.04		-0.01	-0.05	0.04	
Teaming	-0.15	-0.17	-0.12	*	-0.02	-0.04	0.01		0.00	-0.03	0.03		-0.03	-0.05	0.01		-0.01	-0.04	0.02	
Tracking	-0.56	-0.61	-0.50	*	-0.73	-0.80	-0.66	*	-0.78	-0.87	-0.70	*	-0.85	-0.93	-0.78	*	-0.82	-0.90	-0.75	*
Empirical Measures																				
Actively Open-Minded Thinking	-0.01	-0.03	0.00		-0.02	-0.03	-0.00	*	-0.02	-0.03	-0.01	*	-0.02	-0.04	-0.01	*	-0.02	-0.04	-0.01	*
Analytic Thinking	0.00	-0.02	0.02		0.00	-0.02	0.03		0.00	-0.03	0.03		0.00	-0.03	0.03		0.00	-0.03	0.03	
Conscientiousness	0.01	-0.00	0.03		-0.00	-0.02	0.02		0.00	-0.02	0.01		0.00	-0.02	0.02		-0.00	-0.02	0.02	
Extended Cognitive Reflection Test	-0.04	-0.06	-0.03	*	-0.04	-0.06	-0.03	*	-0.04	-0.06	-0.03	*	-0.05	-0.06	-0.03	*	-0.05	-0.06	-0.03	*
Extraversion	0.00	-0.02	0.02		0.01	-0.01	0.03		0.02	-0.01	0.03		0.02	-0.00	0.03		0.02	-0.00	0.03	
Extrinsic Motivation	0.03	0.01	0.05	*	0.03	0.01	0.05	*	0.04	0.01	0.05	*	0.05	0.02	0.07	*	0.04	0.02	0.06	*
Fox-Hedgehog	0.02	-0.00	0.03		0.00	-0.02	0.02		0.00	-0.02	0.02		0.00	-0.02	0.02		0.00	-0.02	0.02	
Grit	0.01	-0.00	0.03		0.00	-0.01	0.02		0.00	-0.02	0.02		0.00	-0.01	0.02		0.00	-0.01	0.02	
Intrinsic Motivation	-0.03	-0.05	-0.01	*	-0.02	-0.04	-0.00	*	-0.03	-0.05	-0.00	*	-0.03	-0.05	-0.01	*	-0.03	-0.05	-0.01	*
Intuitive Thinking	0.00	-0.02	0.02		0.01	-0.01	0.03		0.01	-0.01	0.03		0.01	-0.01	0.03		0.01	-0.01	0.03	
Need for Closure	0.01	-0.01	0.02		-0.01	-0.03	0.01		-0.01	-0.03	0.01		-0.01	-0.03	0.01		-0.01	-0.03	0.01	
Need for Cognition	-0.02	-0.04	-0.01	*	-0.01	-0.03	-0.00	*	-0.01	-0.03	-0.00	*	-0.02	-0.03	-0.00	*	-0.02	-0.03	-0.00	*
Numeracy	-0.03	-0.05	-0.01	*	-0.03	-0.05	-0.02	*	-0.04	-0.05	-0.02	*	-0.04	-0.05	-0.02	*	-0.04	-0.05	-0.02	*
Openness	0.00	-0.02	0.02		0.01	-0.01	0.02		0.01	-0.01	0.03		0.00	-0.01	0.02		0.01	-0.01	0.02	
P-Beauty Contest	-0.01	-0.03	0.01		-0.00	-0.02	0.01		0.00	-0.02	0.02		0.00	-0.02	0.02		0.00	-0.02	0.02	
Political Knowledge	-0.03	-0.05	-0.02	*	-0.04	-0.05	-0.03	*	-0.05	-0.06	-0.04	*	-0.05	-0.06	-0.04	*	-0.05	-0.06	-0.04	*
Prosocial Motivation	0.01	-0.02	0.03		0.00	-0.02	0.02		0.00	-0.02	0.02		0.00	-0.02	0.02		0.00	-0.02	0.02	
Raven Progressive Matrices	-0.02	-0.03	-0.00	*	-0.03	-0.04	-0.02	*	-0.03	-0.04	-0.02	*	-0.03	-0.05	-0.02	*	-0.03	-0.04	-0.02	*
Shipley-2 Abstraction	-0.03	-0.05	-0.01	*	-0.04	-0.06	-0.02	*	-0.04	-0.06	-0.03	*	-0.05	-0.06	-0.03	*	-0.04	-0.06	-0.03	*
Shipley-2 Vocabulary	-0.02	-0.04	0.01		-0.02	-0.04	-0.01	*	-0.03	-0.05	-0.01	*	-0.03	-0.05	-0.01	*	-0.03	-0.05	-0.01	*
Standard Cognitive Reflection Test	-0.03	-0.05	-0.02	*	-0.03	-0.05	-0.02	*	-0.03	-0.05	-0.02	*	-0.04	-0.05	-0.02	*	-0.04	-0.05	-0.02	*
Update Frequency	0.03	-0.01	0.05		-0.03	-0.05	-0.02	*	-0.05	-0.07	-0.02	*	-0.06	-0.09	-0.03	*	-0.05	-0.08	-0.03	*
Update Magnitude	0.27	0.26	0.28	*	0.25	0.24	0.26	*	0.25	0.24	0.26	*	0.23	0.22	0.24	*	0.24	0.23	0.25	*
Working Memory	-0.02	-0.04	-0.00	*	-0.01	-0.03	0.01		-0.01	-0.02	0.01		-0.01	-0.02	0.01		-0.01	-0.02	0.01	
Working Memory Math	-0.02	-0.04	0.00		-0.01	-0.03	0.01		-0.01	-0.03	0.01		-0.01	-0.03	0.02		-0.01	-0.03	0.01	
Working Memory Word	-0.02	-0.04	-0.00	*	-0.01	-0.02	0.01		-0.00	-0.02	0.01		-0.00	-0.02	0.01		-0.00	-0.02	0.01	

Note. b = regression coefficient; Low = lower limit of the 99.92% studentized t bootstrap confidence interval; Up = upper limit of the 99.92% studentized t bootstrap confidence interval; Sig. = significance. * indicates coefficients significant at the $\alpha = 0.0008$ level based on 99.92% studentized t bootstrap confidence intervals. All values represent z-values.

3.9.7 Appendix 3.G Precise Quantitative Results of the Sensitivity Analyses

Table 3.G1 Precise Quantitative Results of the OLS Linear Regressions and the Bootstrap Confidence Intervals

Single Predictor	BS				MDBS				IPBS				Linear TWBS				Square Root TWBS			
	b	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.
1st Half of Forecasters																				
Relationship Between the Linear TWBS and the Scores																				
Linear TWBS	0.69	0.67	0.71	*	0.94	0.93	0.96	*	0.99	0.98	1.01						0.99	0.98	0.99	*
Treatment Groups																				
Training	-0.11	-0.19	-0.03	*	-0.02	-0.08	0.05		0.00	-0.07	0.08		-0.01	-0.09	0.06		-0.01	-0.08	0.07	
Teaming	-0.13	-0.17	-0.09	*	0.00	-0.03	0.04		0.03	-0.01	0.08		0.00	-0.04	0.05		0.01	-0.03	0.06	
Tracking	-0.58	-0.66	-0.49	*	-0.77	-0.89	-0.68	*	-0.83	-0.97	-0.72	*	-0.89	-1.01	-0.79	*	-0.87	-1.00	-0.76	*
Differences Between the Linear Associations of the Tracking Group Relative to the Linear TWBS																				
Tracking	-0.31	-0.41	-0.22	†	-0.12	-0.13	-0.10	†	-0.06	-0.08	-0.05	†					-0.03	-0.03	-0.02	†
2nd Half of Forecasters																				
Relationship Between the Linear TWBS and the Scores																				
Linear TWBS	0.68	0..67	0.70	*	0.94	0.93	0.95	*	0.99	0.98	1.00						0.99	0.99	0.99	*
Treatment Groups																				
Training	-0.01	-0.08	0.05		-0.01	-0.06	0.04		0.00	-0.05	0.05		0.00	-0.06	0.05		0.00	-0.05	0.05	
Teaming	-0.16	-0.20	-0.12	*	-0.04	-0.07	-0.00	*	-0.02	-0.06	0.02		-0.05	-0.09	-0.00	*	-0.04	-0.07	0.01	
Tracking	-0.54	-0.64	-0.46	*	-0.67	-0.78	-0.59	*	-0.72	-0.83	-0.62	*	-0.80	-0.91	-0.71	*	-0.77	-0.88	-0.67	*
Differences Between the Linear Associations of the Tracking Group Relative to the Linear TWBS																				
Tracking	-0.27	-0.37	-0.16	†	-0.13	-0.14	-0.11	†	-0.08	-0.10	-0.07	†					-0.03	-0.04	-0.03	†

Note. Results are based on a random split of the data into two halves. b = regression coefficient; Low = lower limit of the bootstrap confidence interval; Up = upper limit of the bootstrap confidence interval; Sig. = significance. * indicates coefficients significant at the $\alpha = 0.0008$ level based on 99.92% studentized t bootstrap confidence intervals. † indicates coefficients significant at the $\alpha = 0.025$ level based on 95% normal confidence intervals. All values represent z-values. The results of the sensitivity analyses are qualitatively identical to the main analyses except for three linear associations in the 2nd half of forecasters: The relationship between (1) the BS and the training group, (2) the MDBS and the teaming group, and (3) the linear TWBS and the teaming group. (1) is statistically nonsignificant and (2) as well as (3) are significantly negative, albeit the effect sizes are very small. While (1) may be interpreted as evidence that training may not improve foresight when measured by the BS, (2) and (3) may be read as support for the claim that teaming may also improve foresight when measured by the MDBS and linear TWBS. However, these findings are inconsistent with the results based on the IPBS and square root TWBS in the 2nd half of forecasters, the findings based on the 1st half of forecasters, and the results based on all data. Taken together, the findings therefore suggest that training improves foresight conceptualized as the ability to forecast future states of the world accurately and teaming does not enhance foresight conceptualized as the ability to forecast future states of the world early and consistently accurate over time.

Table 3.G2 Precise Quantitative Results of the OLS Linear Regressions and the Bootstrap Confidence Intervals

Single Predictor	BS				MDBS				IPBS				Linear TWBS				Square Root TWBS			
	b	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.	*b*	Low	Up	Sig.
Scores																				
MDBS	0.69	0.65	0.73	*																
IPBS	0.65	0.61	0.70	*	0.94	0.94	0.95	*												
Linear TWBS	0.71	0.67	0.76	*	0.93	0.91	0.96	*	0.98	0.96	1.00									
Square Root TWBS	0.69	0.64	0.74	*	0.95	0.93	0.96	*	1.00	0.99	1.01		0.99	0.98	1.00					
Treatment Groups																				
Training	-0.13	-0.77	0.28		-0.02	-0.25	0.23		0.01	-0.24	0.24		0.01	-0.24	0.26		0.01	-0.25	0.25	
Teaming	-0.22	-0.39	-0.04	*	-0.02	-0.15	0.11		0.01	-0.14	0.17		-0.02	-0.18	0.12		-0.01	-0.16	0.14	
Tracking	-0.38	-5.22	1.30		-0.63	-13.2	2.25		-0.68	-28.7	5.21		-0.73	-19.4	4.07		-0.71	-4.77	5.04	
Empirical Measures																				
Actively Open-Minded Thinking	-0.04	-0.18	0.11		-0.01	-0.14	0.14		0.00	-0.14	0.17		-0.01	-0.17	0.15		-0.01	-0.15	0.15	
Analytic Thinking	-0.09	-0.44	0.59		-0.06	-0.58	0.56		-0.05	-0.55	0.69		-0.05	-0.62	0.45		-0.05	-0.61	0.52	
Conscientiousness	-0.04	-0.32	0.18		-0.04	-0.31	0.09		-0.05	-0.27	0.07		-0.04	-0.30	0.11		-0.04	-0.32	0.10	
Extended Cognitive Reflection Test	-0.07	-0.31	0.11		-0.02	-0.09	0.04		-0.03	-0.10	0.04		-0.04	-0.10	0.05		-0.03	-0.10	0.04	
Extraversion	-0.07	-0.55	0.10		0.00	-0.21	0.22		0.01	-0.25	0.27		0.00	-0.35	0.33		0.01	-0.27	0.29	
Extrinsic Motivation	0.10	-0.17	0.58		0.03	-0.23	0.36		0.03	-0.25	0.34		0.06	-0.21	0.38		0.05	-0.21	0.38	
Fox-Hedgehog	0.06	-0.14	0.20		0.08	-0.09	0.32		0.06	-0.14	0.32		0.07	-0.16	0.36		0.07	-0.17	0.33	
Grit	-0.05	-0.30	0.09		-0.05	-0.27	0.11		-0.05	-0.31	0.13		-0.05	-0.31	0.13		-0.05	-0.30	0.11	
Intrinsic Motivation	-0.12	-0.69	0.30		-0.05	-0.81	0.45		-0.05	-0.80	0.43		-0.04	-0.94	0.56		-0.04	-0.78	0.60	
Intuitive Thinking	-0.04	-1.13	1.00		-0.12	-1.23	0.32		-0.12	-1.27	0.41		-0.11	-1.39	0.67		-0.12	-1.13	0.55	
Need for Closure	-0.05	-0.38	0.21		-0.04	-0.14	0.10		-0.05	-0.15	0.13		-0.06	-0.18	0.10		-0.06	-0.16	0.10	
Need for Cognition	0.02	-0.15	0.29		0.04	-0.06	0.17		0.05	-0.05	0.14		0.06	-0.06	0.18		0.05	-0.06	0.16	
Numeracy	-0.05	-0.28	0.13		0.02	-0.15	0.25		0.02	-0.14	0.28		0.00	-0.16	0.25		0.01	-0.16	0.24	
Openness	-0.11	-0.30	0.04		-0.04	-0.24	0.11		-0.02	-0.27	0.12		-0.03	-0.25	0.16		-0.02	-0.25	0.13	
P-Beauty Contest	-0.03	-0.50	0.36		-0.05	-0.35	0.30		-0.05	-0.38	0.29		-0.05	-0.39	0.46		-0.05	-0.41	0.32	
Political Knowledge	0.00	-0.11	0.12		-0.03	-0.09	0.05		-0.03	-0.11	0.05		-0.03	-0.11	0.05		-0.03	-0.11	0.06	
Prosocial Motivation	0.06	-0.13	0.19		0.00	-0.14	0.15		0.00	-0.16	0.16		0.00	-0.17	0.17		0.00	-5.11	5.06	
Raven Progressive Matrices	-0.09	-0.26	0.11		-0.02	-0.14	0.18		-0.01	-0.12	0.18		-0.03	-0.17	0.17		-0.02	-0.14	0.17	
Shipley-2 Abstraction	-0.08	-0.71	0.27		0.02	-0.51	0.55		0.04	-0.60	0.62		0.02	-0.70	0.88		0.03	-0.57	0.63	
Shipley-2 Vocabulary	0.06	-0.38	0.57		0.11	-0.24	0.70		0.13	-0.30	0.70		0.11	-0.38	0.84		0.12	-0.50	0.90	
Standard Cognitive Reflection Test	-0.07	-0.28	0.03		-0.04	-0.12	0.06		-0.04	-0.12	0.06		-0.04	-0.13	0.06		-0.04	-0.12	0.06	
Update Frequency	0.13	-0.01	0.28		-0.06	-0.17	0.07		-0.10	-0.20	0.04		-0.14	-0.24	0.03		-0.12	-0.22	0.03	
Update Magnitude	0.28	0.23	0.34	*	0.26	0.20	0.39	*	0.26	0.20	0.39	*	0.23	0.17	0.35	*	0.25	0.19	0.37	*
Working Memory	-0.03	-0.78	0.40		0.01	-0.58	0.28		0.02	-0.43	0.28		0.01	-0.43	0.25		0.01	-0.51	0.25	
Working Memory Math	0.10	-0.39	0.46		0.14	-0.04	0.54		0.14	-0.10	0.52		0.13	-0.13	0.59		0.13	-0.08	0.55	
Working Memory Word	-0.07	-0.50	0.24		-0.03	-0.25	0.14		-0.03	-0.23	0.17		-0.04	-0.35	0.14		-0.04	-0.34	0.16	
Differences Between the Linear Associations of the Tracking Group Relative to Linear TWBS																				
Tracking	-0.34	-0.68	-0.01	†	-0.09	-0.15	-0.02	†	-0.04	-0.09	0.00						-0.01	-0.04	0.07	

Note. Results are based on a random sample of 300 forecasters from the full data. b = regression coefficient; Low = lower limit of the bootstrap confidence interval; Up = upper limit of the bootstrap confidence interval; Sig. = significance. * indicates coefficients significant at the $\alpha = 0.0008$ level based on 99.92% studentized t bootstrap confidence intervals. † indicates coefficients significant at the $\alpha = 0.025$ level based on 95% normal confidence intervals. All values represent z-values. Note that the effective sample size of each variable is usually lower than 300 as not all variables were observed for each forecaster. The IPBS, linear TWBS, and square root TWBS become statistically indistinguishable at low sample sizes. The training ($n = 41$) and tracking groups ($n = 24$) have substantially lower sample sizes as only a fraction of all forecasters was assigned to these groups. This resulted in very large confidence intervals of the corresponding regression coefficients and prevents any meaningful conclusions about the effect sizes of these groups. Strikingly, the results of this sensitivity analysis suggest that all variables except for the update magnitude are statistically nonsignificant. This is consistent with the finding that the effect sizes of the variables are generally very small. Overall, the sensitivity analyses further corroborated our main analyses while showing that the differences between the IPBS, linear TWBS, and square root TWBS may only be detectable in larger samples. However, this does not mean that researchers and practitioners should be indifferent between these scores since both the simulation study, the full data results, and the split data results suggest that there are differences between these scores. Furthermore, the linear TWBS still assigns better scores to superforecasters and rewards forecasters with superior foresight more than the other scoring rules. Moreover, the linear TWBS still matches foresight conceptualized as the ability to forecast future states of the world early and consistently accurate over time best.