The randomness in prediction tournaments

Ella de Vries

Student number: 5300592
Supervisor: Dr. J. Söhl

Thesis committee: Dr. J. Söhl, Dr. E. Emsiz
Date of defense: Friday June 26, 2026 at 10:00

Technische Universiteit Delft

TU Delft

Abstract

Prediction tournaments are competitions in which participants report probabilistic forecasts about uncertain future events, after this forecasters are ranked based on their scores conducted with the help of a scoring rule. The main objective of such a tournament is to select the most accurate forecaster as the winner. However, a fundamental problem known as the prediction tournament paradox shows that in standard winner-take-all competitions, the most accurate forecaster does not have the highest probability of winning. The reasoning behind this paradox is that extreme predictions introduce higher variance in realized scores, which can lead to a winning score despite being less accurate on average. To address this paradox the design of the tournaments mechanism is important. For this mechanism Incentive-compatibility is needed to go against forecasters that want to strategize and for rewarding the accurate forecasters. This means that truthful reporting must be the best strategy for winning a prediction tournament.

This thesis analyzes and compares four forecasting competition mechanisms: the standard deterministic mechanism, the Event Lotteries Forecasting Competition mechanism (ELF), the Independent Event Lotteries Forecasting mechanism (I-ELF), and the Wisdom of the Most Accurate Crowd mechanism (WOMAC). The deterministic mechanism always chooses the forecaster with the highest score. However this is where the prediction tournament paradox appears, such that the most accurate forecaster is not always chosen as the tournament winner. This is why the other mechanisms are introduced. ELF and I-ELF add a amount of randomness in choosing the winner, which makes these mechanisms incentive compatible, although the forecaster with the highest score does not always win. The last mechanism which is introduced is WOMAC, this mechanism scores forecasters against a reference prediction made from other forecasters predictions, letting the forecaster with the highest score win and having Bayes-Nash incentive compatibility. The disadvantage of this mechanism is that there is a amount of randomness added by scoring forecasters against a reference prediction and not against the true probabilities. To select the best mechanism to use in a prediction tournament, simulations are made for comparisons. These simulations are made with the help of the point mass noise model for realistic forecasting errors. In this model, each forecasters prediction deviates from the true probability by a fixed amount, either $+\sigma$ or $-\sigma$, with probability $\frac{1}{2}$. This noise structure captures the idea of bounded forecasting errors while maintaining control over the variance introduced in predictions. The mechanisms are evaluated on two criteria: the probability of selecting the most accurate forecaster and the degree of randomness introduced in winner selection, quantified using the expectation of the winner's rank. The results show that while ELF and I-ELF achieve strict dominant strategy incentive compatibility, both mechanisms introduce substantial randomness into winner selection, particularly when the accuracy gap between forecasters is small. The I-ELF mechanism was designed by Witkowski et al. (2021) to reduce this randomness as the number of events grows, and a lower bound on the required number of events is derived using Hoeffding's inequality. After conducting simulations in this thesis, it is found that for this bound an unrealistic high number of events is needed. Simulations were also conducted independently of this bound, examining how varying the number of events affects the performance of this mechanism directly. These simulations confirmed that an unrealistically large number of events would be required to reduce randomness enough to guarantee a desired probability of the best forecaster winning. The WOMAC mechanism, which scores forecasters against a reference prediction constructed from the other forecasters rather than against the realized outcome, achieves Bayes-Nash incentive compatibility and consistently selects the best forecaster with higher probability and less randomness than ELF and I-ELF across all simulated settings.

The findings suggest that for organizations designing prediction tournaments under the given conditions, WOMAC represents the most practical choice, offering the best trade-off between incentive compatibility and reliable identification of the most accurate forecaster.

Layman's Summary

Most people are familiar with predicting the outcomes of football matches or other sports events, but prediction tournaments are also widely used by companies seeking insight into future developments or looking to harness the expertise of specialists in a given field. A prediction tournament is a competition in which participants make probabilistic predictions about future events. Once outcomes are known, the predictions are evaluated using a scoring rule, and participants are ranked based on their total score. The ultimate goal is to identify the most accurate forecaster.

In a standard prediction tournament, where the winner is simply the participant with the highest score, a problem arises called the prediction tournament paradox. This paradox refers to the fact that the most accurate forecaster does not have the highest chance of being selected as the winner. In the case of binary events, those with a Yes or No outcome, where a prediction consists of assigning a probability to how likely something is to occur, more extreme or risky predictions can occasionally achieve higher scores, even if they are less accurate on average. To address this paradox, alternative mechanisms have been proposed. One approach selects the winner randomly, weighted by how well participants scored. However, introducing randomness into the selection process is far from ideal. Another mechanism evaluates participants not against the actual outcome, but against the average prediction of all other forecasters. While imperfect, this approach tends to perform considerably better than randomly selecting a winner with a probability based on their scores.

Since it is important for companies and organizations to reliably identify the most accurate forecaster, this thesis analyzes these different mechanisms, compares their properties, and provides guidance on which mechanism is most suitable under different conditions. The approach of scoring participants against the average of the other participants seems to work the best in selecting the best forecaster and is recommended to organizers of prediction tournaments.

Contents

1

Introduction

Prediction tournaments have become an increasingly widely used tool for eliciting probabilistic forecasts from a group of participants. Their usage is across a wide range of settings, from geopolitical forecasting platforms such as the Good Judgment Project (Good) to machine learning competitions on Kaggle (Kaggle) or the million dollar competition of netflix, as well as entertainment competitions on soccer match outcomes. In a prediction tournament, participants report probability distributions over uncertain future events, which are then evaluated and ranked according to a scoring rule once outcomes are realized. The most important point these competitions seek is to identify the most accurate forecaster and get information out of the predictions. A natural assumption for these tournaments is that the participant whose predictions are closest to the true probabilities of events should be crowned as the winner. However, this is where something counter-intuitive happens, which creates a large issue for prediction tournaments. In standard winner-take-all tournaments, the prize is awarded to the forecaster with the highest overall score. More extreme or risky predictions can get higher realized scores, which can occasionally result in the highest score overall, even when a forecaster was less accurate on average. This phenomenon, known as the prediction tournament paradox, implies that the most accurate forecaster does not have the highest probability of being selected as the winner, undermining the very purpose of the competition.

Prediction tournaments are searching for the perfect fix for the paradox above. This requires moving beyond standard winner-take-all mechanisms. Witkowski et al. (2021) established that no deterministic mechanism can be strictly incentive-compatible. A mechanism is incentive-compatible if each forecaster maximizes their expected reward by reporting their true beliefs, rather than strategically exaggerating or understating them. This means that a randomized element should be added to let truthful reporting of beliefs always be the most optimal strategy for every forecaster. The Event Lotteries Forecasting Competition Mechanism (ELF) and its extension, the Independent Event Lotteries Forecasting (I-ELF) mechanism, both found by Witkowski et al. (2021), are therefore introduced and researched. These mechanisms give probabilities of winning to forecasters, resulting in a random choice of winner. This randomness is external to the realized outcomes of the events. The chances of winning of forecasters are influenced by their reported beliefs, not by the randomness of the events themselves, which is always present. This results in incentive compatibility, encouraging forecasters to report truthfully even when the outcomes are uncertain. This added randomness is however not the most optimal choice, therefore the WOMAC mechanism is introduced by Srinivasan et al. (2025). This mechanism instead scores forecasters against a reference prediction derived from the remaining participants, achieving a weaker but practically meaningful form of incentive compatibility without relying on explicit randomization in winner selection.

The researched mechanisms address the paradox to a certain degree; however, the introduction of randomness raises a further question: how much randomness is strictly necessary and can it be reduced? A mechanism that selects the winner randomly provides limited value to organizations relying on prediction tournaments to identify a valuable forecaster. A mechanism that is not randomized enough may fail to incentivize truthful reporting. In this thesis, the trade-off between incentive com-

patibility and the minimization of randomness in forecasting competition mechanisms is presented with the research question being: *Among the class of incentive-compatible forecasting competition mechanisms, which mechanism selects the most accurate forecaster with the highest probability while using the least amount of randomness?*.

To answer the research question, multiple simulations need to be conducted and the mechanisms need to be compared on all levels. First, the necessary theoretical background is established, covering proper scoring rules and the principles of mechanism design. Next, to simulate realistic forecasting behavior, a noise model is introduced that captures the errors forecasters make in real life relative to the true probabilities. Although the theoretical properties of the mechanism studied in this thesis are established in prior work, comparisons through simulation have not received enough attention. This thesis addresses that gap directly, using simulation to illustrate the practical behavior of each mechanism. Simulations of the different mechanisms then follow to better understand them. These simulations can take many forms, from detecting randomness to demonstrating incentive compatibility and providing visual representations of scoring rules and noise models. Beyond comparing mechanisms at a fixed scale, the influence of the number of events on mechanism performance is also examined, as the capacity of a tournament, that is, how many questions forecasters are asked to predict, plays a crucial role in determining how reliably the most accurate forecaster is identified. Understanding how the number of events influences mechanism performance is therefore of direct relevance to organizations designing prediction tournaments, as it may inform practical decisions about tournament structure. Ultimately, knowing which mechanism best balances incentive compatibility with a reliable identification of the most accurate forecaster is of direct and practical importance to any organization seeking to extract genuine insight from a prediction tournament.

2

Background of prediction tournaments

Prediction tournaments rely on probabilistic forecasts of uncertain events. Before building a prediction tournament, some basic probabilities have to be introduced.

Starting with a random variable X representing the outcome of an event. In the simplest case, X takes values in $\{0, 1\}$, where $X = 1$ indicates that the event occurs and $X = 0$ otherwise. A probability distribution assigns probabilities to all possible outcomes of X. For a binary event, a distribution of X can be described by a Bernoulli distribution with parameter $p \in [0, 1]$, denoted by

$$X \sim \mathrm{Ber}(p).$$

This means that

$$\mathbb{P}(X = 1) = p, \qquad \mathbb{P}(X = 0) = 1 - p.$$

In a prediction tournament, a forecaster reports a probability distribution representing their subjective belief about the outcome of events. The quality of the forecast is then evaluated with the help of scoring rules that reward accuracy and, in the case of proper scoring rules, incentivize truthful reporting of beliefs.

In many applications, forecasts depend on available information. The conditional probability of an event given some information I is denoted by

$$\mathbb{P}(X = x \mid I).$$

This allows forecasters to update their beliefs when new information becomes available. Furthermore, only prediction tournaments with independent events are evaluated throughout this thesis.

2.1. Proper scoring rules

For a prediction tournament to have a winner, scoring rules must be introduced (Dawid and Musio, 2014). A scoring rule is a function that takes two inputs, a forecaster's reported probability and the actual outcome of an event, and returns a numerical score. For instance, when a forecaster believes there is a 70% chance of rain tomorrow and it does rain, the scoring rule evaluates how good this prediction was.

Formally, for a binary event $X \in \{0, 1\}$, a scoring rule R takes a reported probability $y \in [0, 1]$ and a realized outcome $x \in \{0, 1\}$ and returns a real number $R(y, x)$. Higher scores indicate better predictions.

Now suppose a forecaster holds a subjective belief represented by a probability distribution $\mathbb{P}$ over the outcome X. The expected score of reporting y under belief $\mathbb{P}$ is defined as

$$R(y;\, \mathbb{P}) := \mathbb{E}_{X \sim \mathbb{P}}\big[R(y, X)\big].$$

Rational behavior can be formalized as maximizing expected utility. Because of this, expected scores are used for prediction tournaments (Good, 1951). For binary events, the forecaster's belief is fully

characterized by a single number $p \in [0,1]$, corresponding to the Bernoulli distribution $\mathrm{Ber}(p)$ with $\mathbb{P}(X=1)=p$. We therefore write

$$R\big(y;\,\mathrm{Ber}(p)\big) = \mathbb{E}_{X\sim Ber(p)}\big[R(y,X)\big] = \sum_{x\in\{0,1\}} p(x)\,R(y,x) = p\cdot R(y,1) + (1-p)\cdot R(y,0),$$

Representing the average score a forecaster with true belief p expects to receive when reporting y, if the experiment were repeated many times.

Definition 1 (Waggoner, 2017) *A scoring rule is a function $R : [0,1]\times\{0,1\} \to \mathbb{R}$. It is called **proper** if truthfulness maximizes expected score, i.e., for all beliefs p and all reports $y \neq p$,*

$$R(p;p) \geq R(y;p).$$

*It is called **strictly proper** if truthfulness uniquely maximizes expected score, i.e., for all p and all $y \neq p$,*

$$R(p;p) > R(y;p).$$

Let δ_x denote the distribution that assigns probability one to outcome x. If a forecaster believes δ_x, then the outcome is certain and the score is $R(y,x)$. Hence,

$$R(y;\delta_x) = R(y,x).$$

Proper scoring rules incentivize truthful reporting of beliefs, whereas improper scoring rules encourage extreme forecasts (close to 0 or 1). For tournaments to be incentive-compatible, these *proper* scoring rules are needed. (Palfrey and Wang, 2009)

Definition 2 *A scoring rule is **bounded** if:*

there exist numbers a, b such that $a \leq R(y,x) \leq b$ for all y, x.

With the help of these definitions, multiple scoring rules can be defined and their properness and boundedness can be analyzed.

2.1.1. Zero-one score

The zero-one score is defined as

$$R(y,x) = \begin{cases} 1 & \text{if the forecast matches the outcome exactly,} \\ 0 & \text{otherwise.} \end{cases}$$

The zero-one scoring rule, in a binary event outcome setting, can be seen as the most natural way to evaluate a forecast, forecasters are rewarded when they are right and receive nothing when they are wrong. This simplicity is its main appeal; the rule is intuitive and requires no mathematical background to understand or explain. This makes the scoring rule accessible to a broad audience.

Despite its intuitive appeal, the zero-one rule has a fundamental flaw that makes it unsuitable for probabilistic forecasting, namely that it is not a proper scoring rule. To see why, consider a forecaster who genuinely believes an event occurs with probability 0.6. Under the zero-one rule, she maximizes her expected score not by reporting 0.6, but by reporting 1. The rule therefore incentivizes forecasters to report deterministic guesses rather than honest probability assessments, entirely defeating the purpose of eliciting probabilistic forecasts. Furthermore, a forecaster who wanted to report 0.51 for an event that occurs with probability 0.9 is treated identically to one who wanted to report 0.99, although the 0.99 is closer to the truth. This means that the rule ignores all the information contained in a probabilistic forecast beyond the binary question of which outcome was most likely.

2.1.2. Quadratic scoring rule (Brier score)

For a binary event $x \in \{0, 1\}$ and a reported probability $y \in [0, 1]$, the quadratic scoring rule, introduced by Brier (1950), is defined as

$$R_{\text{Quadratic}}(y, x) = 1 - (y - x)^2. \tag{2.1}$$

For example, if a forecaster reports $y = 0.7$ and the event occurs ($x = 1$), her score is $1 - (0.7 - 1)^2 = 0.91$. If the event does not occur ($x = 0$), her score is $1 - (0.7 - 0)^2 = 0.51$. The expected score for a forecaster with true belief p who reports y is therefore:

$$R_{\text{Quadratic}}(y;\, p) = p \cdot \left(1 - (y - 1)^2\right) + (1 - p) \cdot \left(1 - y^2\right) = 1 - (y - p)^2 - p(1 - p),$$

This is the form used throughout this thesis. The rule can however be generalized to continuous outcomes. In that setting, a forecaster submits a full probability distribution $\mathbb{Q}$ with density $p(x)$ over the real line, and the quadratic scoring rule becomes (Matheson and Winkler, 1976):

$$R_{\text{Quadratic}}(\mathbb{Q}, y) = -2p(y) + \int_{-\infty}^{\infty} p(x)^2 \, dx$$

Where y is the realized outcome, $p(y)$ is the density that the forecaster assigned to that outcome, and the integral penalizes distributions that are overly spread out. The Brier score is introduced for binary events, in continuous cases the CRPS score introduced in section 2.1.7 works better.

The quadratic scoring rule has several appealing properties. It is straightforward to understand and compute, and as a strictly proper scoring rule it incentivizes truthful reporting. Being bounded is an additional practical advantage, since it guarantees that scores never become extremely large or extremely small regardless of the given forecast. A further strength is that large mistakes are penalized more severely than small ones, a property that aligns well with the desired prediction tournament.

However, there are also some small limitations to this scoring rule. The choice of the quadratic function is to some extent arbitrary, because there is no deeper theoretical reason to prefer squared over other functions. By comparing the quadratic rule to the logarithmic scoring rule (introduced in section 2.1.3) it is found that quadratic is less sensitive to extreme overconfidence. As the quadratic rule is bounded, a forecaster who reports a probability of zero for an event that subsequently occurs receives a finite penalty, whereas the logarithmic rule would assign a score of $-\infty$ in the same situation. In competition settings, boundedness is highly recommended, which makes the quadratic rule's weakness an advantage.

2.1.3. Logarithmic scoring rule

In the binary setting, the outcome space is in $\{0, 1\}$. A reported probability $q \in [0, 1]$ corresponds to the Bernoulli distribution. The logarithmic score (Good, 1951) for binary settings is therefore:

$$R_{\text{Log}}(q, x) = \begin{cases} -\log(q), & \text{if } x = 1, \\ -\log(1 - q), & \text{if } x = 0. \end{cases}$$

The logarithmic scoring rule in continuous setting, with $\mathbb{Q}$ being the forecasters submitted probability distribution and $p(y)$ the density assigned to outcome y, is given by (Matheson and Winkler, 1976):

$$R_{\text{Log}}(\mathbb{Q}, y) = -\log(p(y))$$

Again, the binary case is preferred, since the score is made for binary cases, the binary form of the logarithmic scoring rule is simpler and broader used in reality.

Just like the quadratic rule, the logarithmic rule is strictly proper, and therefore incentivizes truthful reporting. Its most distinctive feature is that it strongly penalizes overconfident wrong predictions. If a forecaster reports probability 0 for an event that actually occurs, then the score is

$$\log(0) = -\infty.$$

Therefore, the logarithmic rule is unbounded below, meaning that scores can become arbitrarily negative. While this is defended theoretically, one should never be infinitely certain about an uncertain event. It makes the rule difficult to work with in practice. A single extreme misreport can dominate all other scores, giving a meaningless comparison between forecasters.

However, the logarithmic scoring rule has a natural connection to information theory, as it is equivalent to the Shannon entropy and Kullback–Leibler divergence. This interpretation provides a strong theoretical justification for its use (Gneiting, 2007).

2.1.4. Spherical scoring rule

The spherical scoring rule for binary outcomes is defined as:

$$S_{\text{spherical}}(y, x) = \frac{x^y (1-x)^{1-y}}{\sqrt{x^2 + (1-x)^2}}.$$

In the case $x = 0$ and $y = 0$ occur, 0^0 is defined as 1 in scoring rules. This is because when $x = 0$ (the event did not occur), the contribution of x^y should not zero out the score. With continuous outcomes with probability density $p(y)$ (Matheson and Winkler, 1976) the spherical rule is defined as:

$$R_{\text{Sphere}}(\mathbb{Q}, y) = -\frac{p(y)}{\sqrt{\int_{-\infty}^{\infty} p(x)^2 \, dx}}$$

The spherical scoring rule is a strictly proper and bounded scoring rule that evaluates probabilistic forecasts by normalizing the reported probability vector. In the binary case, a proper scoring rule R is normalized if it is bounded between 0 and 1, and if $R(0,0) = R(1,1) = 1$ and $R(y,x) = 0 \quad$ for some $y \in [0,1]$ and $x \in \{0,1\}$ (Witkowski et al., 2021). This normalization ensures that forecasts are assessed relative to their overall distribution, rewarding accurate predictions while maintaining stability in the scoring process.

An important advantage of the spherical scoring rule is that, due to its boundedness, scores remain finite and well-behaved for all possible forecasts. However, compared to more commonly used scoring rules such as the quadratic or logarithmic score, the spherical scoring rule is less intuitive and can be more difficult to interpret. As a result, it is less frequently used and may require additional explanation when presented in practical settings.

2.1.5. Pseudospherical Score

The pseudospherical score is defined as:

$$S_{\text{Pseudosphere}}(\mathbb{Q}, y) = -\left(\frac{p(y)}{\sqrt[\beta]{\int_{-\infty}^{\infty} p(x)^\beta dx}} \right)^{\beta - 1}$$

The parameter β in scoring rules is a tuning parameter that adjusts the sensitivity of the scoring rule to different aspects of the forecast. The pseudospherical scoring rule is a generalization of the spherical scoring rule, which uses $\beta = 2$ (van der Eng, 2025). The role of β is to control how much weight is given to different probability values, especially extreme predictions (near 0 or 1). The following β values give the following results:

- $\beta > 1$: Punishes extreme predictions more, as it increases the weight of higher or lower probabilities in the score.
- $\beta < 1$: Makes the rule more lenient towards extreme predictions, as it decreases the penalty for very confident forecasts

The pseudospherical scoring rule generalizes the spherical scoring rule with the parameter $\beta > 1$. This additional flexibility is its biggest advantage. By changing β, the system designer can adjust how

aggressively the scoring rule penalizes inaccurate forecasts, making it adaptable to different contexts. Following from the spherical rule, the pseudospherical rule is strictly proper for any fixed $\beta > 1$, and therefore incentivizes truthful reporting, which is needed for the desired prediction tournament

However, the pseudosperical scoring rule does not only have strengths. The parameter β adds a certain amount of complexity that is absent from simpler rules such as the Brier score. A designer must choose a value of β before the mechanism can be applied, and different choices lead to different rankings of forecasters. This makes the results harder to interpret and more difficult to compare across settings. Moreover, the behavior of the power function within the pseudospherical rule is less intuitive than the quadratic rule or spherical rule, making it more difficult to communicate the mechanism to forecasters.

2.1.6. Power Score

The power score is defined as:

$$R_{\text{Power}}(\mathbb{Q}, y) = -\beta p(y)^{\beta-1} + (\beta - 1) \int_{-\infty}^{\infty} p(x)^{\beta} dx$$

Where $\beta > 1$ is a tuning parameter. Like the pseudospherical rule, the power score belongs to the broader family of parameterized proper scoring rules and is strictly proper for any fixed $\beta > 1$, meaning that truthful reporting is always the optimal strategy for forecasters who wish to maximize their expected score. Again, its most distinctive feature is the flexibility received by the parameter β. However, this feature makes the rule more complex, because different values of β can lead to substantially different rankings of forecasters, making results difficult to interpret and compare

Furthermore, the power score is unbounded below, meaning that poor forecasts can receive arbitrarily large penalties, which was also seen in the logarithmic rule. Therefore, the power score is not desirable for prediction tournaments.

2.1.7. Continuous Ranked Probability Score (CRPS)

The continuous ranked probability score is defined as:

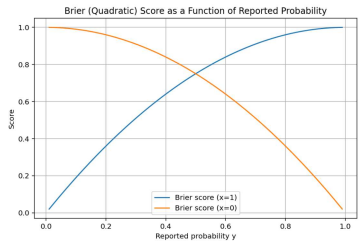
$$R_{\text{CRPS}}(\mathbb{Q}, y) = \int_{-\infty}^{\infty} \left(F_{\mathbb{Q}}(x) - \mathbb{1}\{y \leq x\}\right)^2 \, dx$$

Where $F_{\mathbb{Q}}$ is the cumulative distribution function of the probabilistic forecast $\mathbb{Q}$. Furthermore, $\mathbb{1}\{y \leq x\}$ is the indicator function equal to one if the observed outcome y is at most x and zero otherwise.

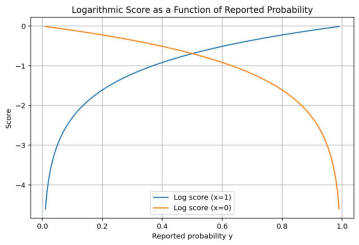
Intuitively, the CRPS measures how far the predicted distribution is from the actual result across all possible values. It can be seen as an extension of the absolute error, where instead of comparing single values, entire probability distributions are compared. The CRPS has several appealing properties, starting with its strictly properness resulting in incentivization of truthful reporting. A forecast that is centered on the correct outcome but is overconfident will be penalized, as will one that is statistically correct on average but not very informative. This makes the CRPS particularly valuable in applications where the forecast cannot be fully summarized by a single number, such as financial risk assessment or weather forecasting where decision makers need to know not just the most likely temperature but also how uncertain that prediction is.

The CRPS is however not without limitations. It is more computationally demanding than simpler scoring rules, particularly when the forecast distribution is complex or when it must be evaluated on large datasets. Moreover, the need for the forecasters to submit a predictive distribution instead of a single point prediction or a scalar probability makes it harder for the forecasters.

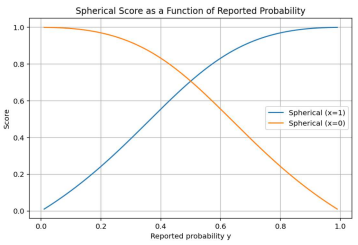
In figure 2.1, plots of the different scoring rules are presented. Here the reported probability y is set against the score.



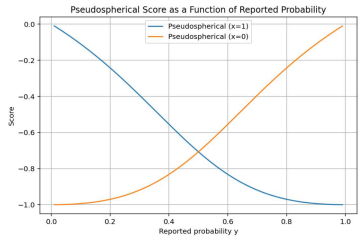
(a) Brier Score



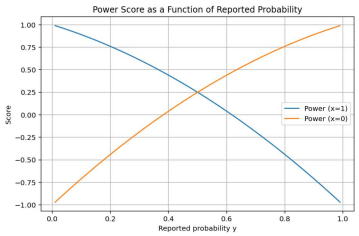
(b) Logarithmic Score



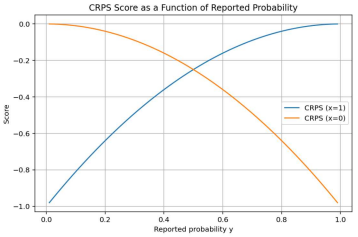
(c) Spherical score



(d) Pseudo-spherical Score



(e) Power Score



(f) CRPS

Figure 2.1: Comparison of six scoring rules as a function of reported probability.

2.2. Mechanism design

When a proper scoring rule is used to rank forecasters, a natural approach is to award the prize to the forecaster who achieves the highest cumulative score. This mechanism is simple, transparent, and widely used in practice. Examples include the Good Judgment (Good) Project and Kaggle's machine learning competitions (Kaggle). However, simplicity does not imply good incentive properties.

A central challenge in mechanism design is that rules must be designed with the resulting behavior in mind. In classical game theory, a game is specified first and the behavior of players that follow from those rules is then analyzed. In the design of forecasting competitions, this order must be reversed: the desired behavior (truthful reporting) is fixed as an objective, and the mechanism must then be constructed to make that behavior optimal for every forecaster. This property is known as incentive compatibility, or equivalently, strategy-proofness. Formally, a mechanism is strictly incentive compatible if truthful reporting is a dominant strategy. That is, reporting one's true beliefs maximizes the probability of winning regardless of what other forecasters report (Witkowski et al., 2021). Personal beliefs can be represented as probabilities. Forecasters' predictions are subjective estimates of likelihood, which the mechanisms then aggregate (Savage, 1971).

It is natural to ask whether any deterministic mechanism can achieve this. A deterministic mechanism is one in which the winner is a fixed, non-random function of the submitted forecasts and the realized outcomes. For instance, the rule "the forecaster with the highest score wins." Such mechanisms are attractive for their transparency and ease of implementation. Unfortunately, Witkowski et al. establish the following impossibility result, Considering a deterministic forecasting mechanism where $m \geq 1$ independent events have binary outcomes.

Theorem 1 (Witkowski et al., 2021) *No deterministic forecasting competition mechanism is strictly incentive compatible.*

The intuition behind this result is as follows. Consider any deterministic mechanism and fix the reports of all forecasters except forecaster i, given by y_{-i}. Given these fixed reports, the set of outcomes for which forecaster i wins is completely determined by their own report y_i. If the reports of the other forecasters are not known, a forecaster chooses a report without conditioning on y_{-i} and instead maximizes the expected probability of winning over all possible reports from the others. When the outcome space is finite (there are only 2^m possible realizations of m binary events), there are only finitely many distinct winning sets available to any forecaster i. Concluding that the forecaster's reporting space is continuous, by the pigeonhole principle there always exist two distinct reports that produce the same

winning set. It follows that the forecaster is indifferent between these two reports, which violates the strict preference for truthful reporting required by incentive compatibility (Witkowski et al., 2021). Since for fixed y_{-i} there exist multiple reports that induce the same winning set, these reports also yield the same conditional winning probability. Averaging over all possible y_{-i} does not eliminate this indifference, because in each scenario, multiple reports of forecaster i still yield the same winning outcome.

One may ask whether this argument extends to infinite outcome spaces, such as when predicting continuous random variables. The underlying issue remains. Since the mechanism is deterministic and the winner is determined by comparing scores, small changes in a forecaster's report typically do not change the winner. As a result, the probability of winning is constant over intervals of reports and changes only at certain boundary points. This creates flat regions in the forecaster's objective, meaning that small deviations from the truthful report do not affect the probability of winning. Even if the score differences are very small and continuous, a forecaster can slightly over or under report and still remain the top scorer. Since a deterministic mechanism chooses the forecaster with the highest score as the winner, strict incentive compatibility still fails.

This impossibility result motivates a departure from deterministic mechanisms. If no deterministic rule can guarantee truthful reporting, attention must turn to a mechanism that selects a winner in a different way, while keeping the requirement that the tournament should still favor better forecasters.

3

The prediction tournament paradox

The goal of a prediction tournament is naturally to let the most accurate forecaster be the winner. Intuitively, It is expected that the forecaster with the prediction closest to the truth should be the winner of such a prediction tournament. However, this intuition does not generally hold. This phenomenon is known as the *prediction tournament paradox* (Aldous, 2019).

In a standard prediction tournament, widely used in many prediction tournaments such as the good judgment project (Tetlock and Gardner, 2015), each forecaster $i \in \{1, \dots, n\}$ reports a probability $y_i \in [0, 1]$ for a binary event $X \in \{0, 1\}$. After the outcome is realized, each forecaster receives a score according to a proper scoring rule $R(y_i, X)$. The winner is determined by

$$i^* = \arg \max_{i \in \{1,\dots,n\}} R(y_i, X).$$

For multiple events $k = 1, \dots, m$, the winner is given by

$$i^* = \arg \max_{i \in \{1,\dots,n\}} \sum_{k=1}^{m} R(y_{i,k}, X_k).$$

This mechanism gives a winner-take-all outcome where the forecaster with the highest realized score wins the entire competition. Suppose forecaster i has belief $q \in [0, 1]$ about the probability that $X = 1$. Under a proper scoring rule, truthful reporting maximizes expected score:

$$\mathbb{E}_q[R(q, X)] \geq \mathbb{E}_q[R(y, X)] \quad \text{for all } y \neq q.$$

However, in a winner-take-all tournament, the objective is not to maximize expected score, but rather to maximize the probability of winning:

$$\arg\max_{y_i} \ \mathbb{P}\left(R(y_i, X) \geq R(y_j, X) \ \forall j \neq i\right).$$

These two objectives are not equivalent. In particular, maximizing the probability of winning may incentivize forecasts that differ from their true belief q. Proper scoring rules ensure that truthful reporting maximizes expected score, however they do not guarantee that the most accurate forecaster has the highest probability of receiving the highest realized score in a finite number of events. More extreme or risky predictions can lead to higher variability in scores. These scores may be less accurate on average, but can lead to very high scores when favorable to the outcome and actually increasing the probability of having a higher score than all other participants.

Example 1 *Consider two forecasters, $i = 1, 2$, and a single binary event $X \in \{0, 1\}$. Suppose both forecasters have the same true belief:*

$$q = \mathbb{P}(X = 1) = 0.7.$$

Assume forecaster 1 reports truthfully:

$$y_1 = q = 0.7,$$

While forecaster 2 considers reporting an extreme prediction $y_2 \in \{0.7, 1\}$. *with the use of the quadratic scoring rule the scores of the two forecasters are realized.*

$$R(y, x) = 1 - (y - x)^2.$$

If $y_2 = 0.7$, *then both forecasters receive identical scores for any realization of* X, *so each forecaster wins with probability*

$$\mathbb{P}(win) = \frac{1}{2}.$$

Now suppose forecaster 2 reports $y_2 = 1$. *If* $X = 1$:

$$R(1,1) = 1, \quad R(0.7,1) = 1 - (0.3)^2 = 0.91.$$

resulting in a win for forecaster 2.

If $X = 0$:

$$R(1,0) = 1 - (1)^2 = 0, \quad R(0.7,0) = 1 - (0.7)^2 = 0.51.$$

resulting in a win for forecaster 1.

The probability that forecaster 2 wins is

$$\mathbb{P}(win) = \mathbb{P}(X = 1) = 0.7.$$

obtaining the following:

$$\mathbb{P}(\textit{win with truthful report}) = 0.5,$$

$$\mathbb{P}(\textit{win with extreme report}) = 0.7.$$

Thus, although $y_2 = 0.7$ *maximizes expected score, the deviation* $y_2 = 1$ *strictly increases the probability of winning when* y_1 *chooses to report thruthfully.*

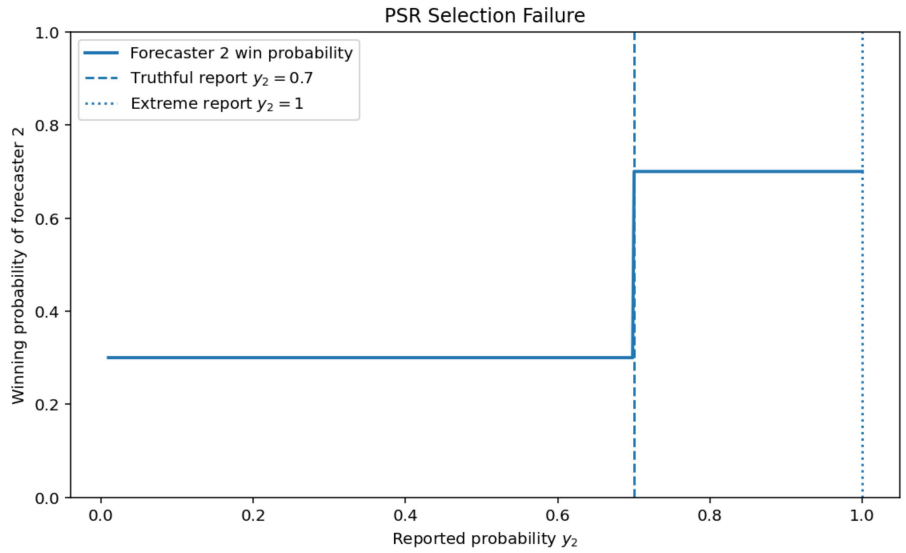


Figure 3.1: Winning probability of forecaster 2 when forecaster 1 reported 0.7

In Figure 3.1 it can be seen that reporting thruthfully does not result is the highest probability of winning. The prediction tournament is a result of a mismatch between avergae accuracy and winning probability. The standard Winner-take-all competitions are therefore flawed and in need of change.

3.1. ELF mechanism

As a result of the prediction tournament paradox the choice in mechanisms are studied. Since by theorem 1 of Witkowski et al. (2021) no deterministic mechanism can achieve strict incentive compatibility, a randomized mechanism may be the solution. In this context, a forecaster is considered more accurate if their forecast obtains a higher expected score under the proper scoring rule. Thus, if forecaster i has a higher expected score than forecaster j, an accuracy-based mechanism should give forecaster i a higher probability of winning. Randomized mechanisms are useful because they can connect winning probabilities directly to the accuracy of a forecast. Instead of assigning the full prize deterministically to one forecaster, a randomized mechanism assigns each forecaster a probability of winning that reflects her relative performance. In this way, the mechanism preserves truthful incentives while still favoring more accurate forecasters. Kilgour (2003) introduces the idea that participants' scores can be compared against others' reports (or some refrence) to incentivize accuracy, resulting in the making of a different mechanism for prediction tournaments.

The solution proposed by Witkowski et al. (2021) is to normalize scores additively rather than multiplicatively. The resulting mechanism is called the Event Lotteries Forecasting Competition Mechanism (ELF) denoted by M_{ELF}. Let m denote the number of events in the tournament, n the number of forecasters and $R(y_i, x)$ a bounded strictly proper scoring rule for forecaster i with prediction $y_i \in [0, 1]$ and realized outcome $x \in \{0, 1\}$ for binary events, which is looked at mainly throughout this thesis. The probability that forecaster i is selected as winner is given by:

$$f_i(y_1, \dots, y_n, x) = \frac{1}{n} + \frac{1}{n}\left(R(y_i, x) - \frac{1}{n-1}\sum_{j \neq i} R(y_j, x)\right) \tag{3.1}$$

Every forecaster starts with probability $\frac{1}{n}$ and it is added a probability of winning based on how well they did relative to everyone else for that event. A winner is then drawn from the probability distribution of the probabilities of all forecasters.

Theorem 2 (Witkowski et al. (2021)) *The Event Lotteries Forecasting Competition Mechanism M_{ELF} is strictly incentive compatible for $m = 1$.*

The proof of this theorem follows from the linearity of expectation and strict properness of R. Each forecaster maximizes their expected winning probability by maximizing their expected score, with strict properness this is achieved by reporting truthfully (Witkowski et al., 2021). Looking at a standard two-forecaster single event setting (Lichtendahl and Winkler, 2007), the truthful incentive can be simulated as follows. In figure 3.2, the winning probability of forecaster 2 is plotted against the reported probability of that forecaster with the truthful report being $y_1 = 0.7$. Forecaster 1 reports truthfully and is therefore fixed at a certain report 0.7, which serves as a reference. The winning probability of forecaster 1 depends on the winning probability of forecaster 2 and is calculated by $\mathbb{P}(F_1 \text{ wins} \mid F_1 \text{ reports } y_1) = 1 - \mathbb{P}(F_2 \text{ wins} \mid F_1 \text{ reports } y_1)$. It can be seen that the highest point in the graph of the ELF mechanism is if forecaster 2 reports truthfully 0.7.

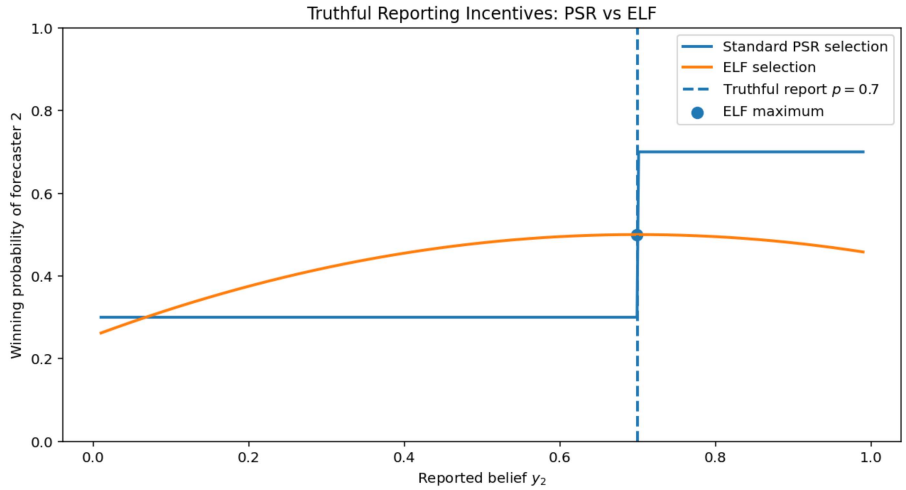


Figure 3.2: Winning probability of forecaster 2 depended on their reported probability, when forecaster 1 reports truthfully.

Extending the ELF mechanism to m events leads to the mechanism assigning each forecaster a winning probability equal to the average of their single-event ELF probabilities across all events:

$$g_i(y_1, \dots, y_n, \mathbf{x}) = \frac{1}{m} \sum_{k=1}^{m} f_{i,k},$$

where $f_{i,k}$ is the single-event ELF probability of forecaster i for event k. Strict incentive compatibility of multi-event ELF then follows directly from single-event incentive compatibility in theorem 2, since each forecaster can maximize their average winning probability event by event.

3.2. I-ELF mechanism

The choice of the winner can be quite random in the ELF mechanism. To reduce some of this randomness in choosing a winner for a prediction tournament, the I-ELF (Independent event-level forecasting) mechanism is introduced (Witkowski et al., 2021). This mechanism is an extension of ELF, where each event performs its own single event ELF competition. From each of these single event ELF competitions, a winner i is chosen for all events k and denoted as w_k. Then the winner of the whole tournament is chosen as follows:

$$\arg\max_i \sum_{k=1}^{m} \mathbb{1}_{(w_k=i)}$$

Where $\mathbb{1}_{(w_k=i)}$ is the indicator function giving 1 if $w_k = i$ and 0 otherwise. The purpose of I-ELF is to reduce randomness in the choice of a winner as the number of events grows, compared to the ELF mechanism.

4

Modelling prediction errors

After establishing the prediction tournament paradox and addressing it by tournament mechanism involving randomness, it is important to look at the strength of different types of mechanisms to see how to build the best working prediction tournament. This can be done by making simulations. To make simulations, a noise model is needed to mimic errors made in real life forecasting competitions. Each event k is associated with a true probability θ_k. Forecasters do not observe this true probability directly, but instead receive a noisy estimate

$$p_{i,k} = \theta_k + \varepsilon_{i,k},$$

where $\varepsilon_{i,k}$ represents a random error term.

More accurate forecasters have predictions closer to the true probabilities on average and are therefore modeled by having a smaller variance in the noise term. By analyzing the prediction tournament it is found that even though some forecasters are more accurate on average, the presence of randomness in both the forecasts and the outcomes implies that the best forecaster does not necessarily have the highest probability of winning the tournament. With the help of this error term, the prediction tournament paradox and the working of different kind of mechanisms can be modeled. In figure 4.1, the prediction tournament paradox is visualized by Aldous (2019). 300 contestants are ranked based on their accuracy with the help of an error parameter. After running simulations it can be seen that the most amount of wins lie around the 100th most accurate forecaster and the most accurate forecaster never wins.

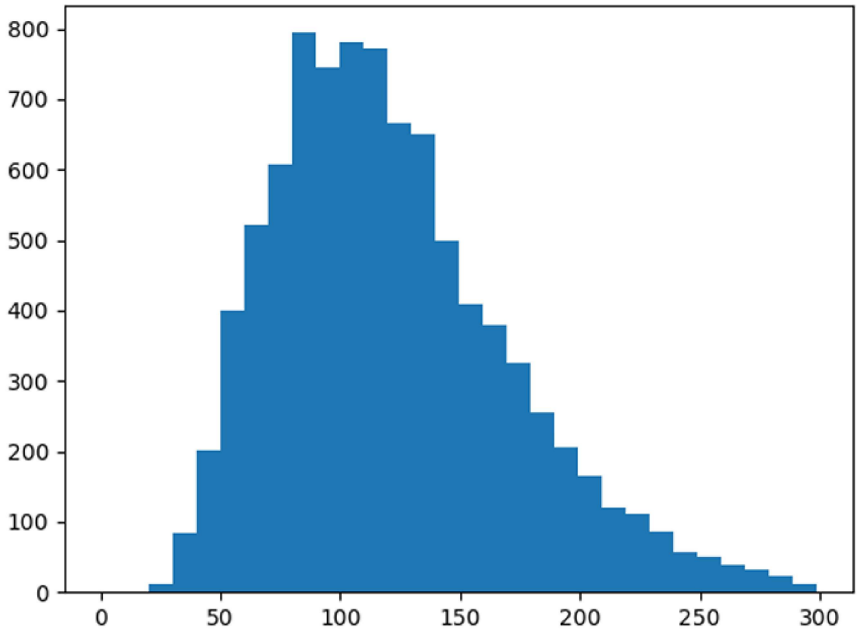


Figure 4.1: Rank of tournament winner, 300 contestants, error parameters $0 < \sigma < 0.3$ (Aldous, 2019)

4.1. Different kinds of noise

For simulating with noise, different noise models can be used. The choice of noise model plays a crucial role in determining the behavior of prediction tournaments. The choice of noise model does not determine the design of the prediction tournament, but rather the environment in which the tournament operates. When simulating a prediction tournament it is important to know how accurate the forecasters are for mimicking their behavior. To choose a noise model, they first have to be evaluated.

4.1.1. Normal (Gaussian) noise

A noise term that is used often is the normal distributed noise model with error term given by:

$$\varepsilon_{i,k} \sim \mathcal{N}(0, \sigma_i^2).$$

This term assumes symmetric errors, where small deviations are more likely than large ones.

The advantages of this noise model are that it is simple and widely used in literature (Gneiting and Raftery, 2017). And through the variance, a clear interpretation of the forecasters accuracy can be found. However, it can produce values outside $[0, 1]$, requiring truncation and it does not capture asymmetry or overconfidence effects.

4.1.2. Uniform noise

Alternatively, the noise term can be uniformly distributed:

$$\varepsilon_{i,k} \sim \mathcal{U}(-a_i, a_i).$$

This implies that all errors within a fixed range are equally likely.

Again, this noise term is simple and easy to implement. It ensures the error $\varepsilon_{i,k}$ is guaranteed to remain within the interval $[-a_i, a_i]$. In other words, the forecast $\theta_k + \varepsilon_{i,k}$ cannot exceed the range $[\theta_k - a_i, \theta_k + a_i]$ before any clipping. However, the uniform noise is less realistic as all errors are equally likely and it does not reflect the fact that small errors are more common than large ones as done in the Gaussian noise. Uniform noise produces less extreme forecasts, which may reduce the impact of the paradox compared to other models.

4.1.3. Beta-distributed forecasts.

Instead of adding noise, one can directly model forecasts as random variables centered around the true probability θ_k:

$$p_{i,k} \sim \text{Beta}(\alpha_i, \beta_i),$$

This approach ensures that the predictions are within the interval $[0, 1]$ and allows for modeling over- or under-confidence. However, it can be more complex to calibrate and parameter selection is of high importance. The α_i and β_i are crucial because they control the mean, variance, and shape of forecasters predictions, which impacts accuracy, randomness, and the incentive structure of a tournament.

4.1.4. Logit-normal noise.

A more refined model applies noise using logit:

$$\text{logit}(p_{i,k}) = \text{logit}(\theta_k) + \varepsilon_{i,k}.$$

This avoids boundary issues near 0 and 1 and provides a more realistic representation of probabilistic forecasts. Again, as a disadvantage it can be more complex mathematically and computationally, harder to interpret directly and it introduces a bias in the prediction, because the expected value of $p_{i,k}$ is not equal to θ_k: $\mathbb{E}[p_{i,k}] \neq \theta_k$. This is found after applying the inverse logit and it being non-linear.

$$p_{i,k} = \frac{1}{1 + e^{-\left(\text{logit}(\theta_k) + \epsilon_{i,k}\right)}}$$

4.1.5. Point mass noise

The last looked at error term is given in the form of the point mass model. The prediction deviates from the true probability by a fixed term σ. The forecast is given by:

$$p_{i,k} = \begin{cases} \theta_k + \sigma & \text{with probability } \frac{1}{2}, \\ \theta_k - \sigma & \text{with probability } \frac{1}{2}. \end{cases}$$

Point mass is like the Gaussian noise used most in practice (Witkowski et al., 2021) (Aldous, 2019). This is because it is easy to implement, it has no large or unrealistic deviations because it is controllable. This simpleness can also be seen as a disadvantage of this noise model, because it can cause unrealistic errors. Real errors are continuous and can vary in size. Point mass is mainly used in binary outcomes, because it is less natural for continuous outcomes.

In figure 4.2, the different kinds of noise models that are introduced are compared in terms of winning probability of the forecaster with the highest score in the standard mechanism. The point-mass and Gaussian noise model seem to have the highest probabilities that the best forecaster wins.

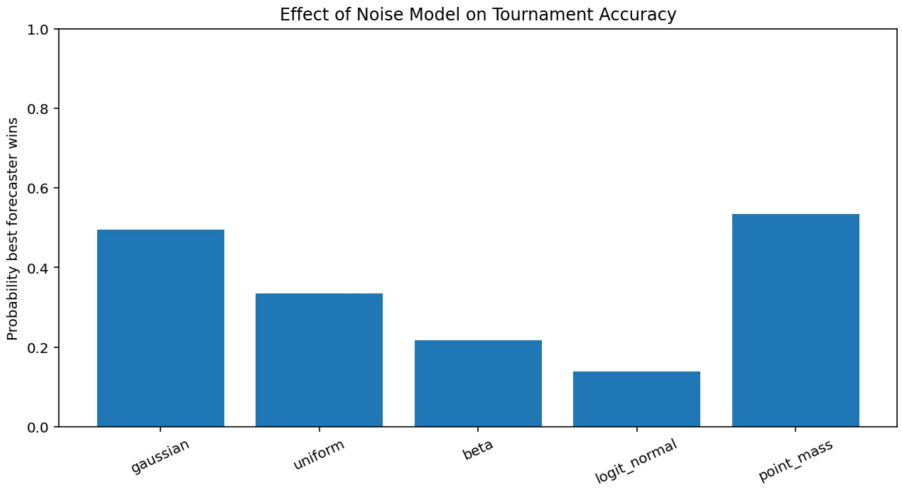


Figure 4.2: Effect of noise model on tournament accuracy

4.2. Noise Model Comparison

The Gaussian and point mass noise models are selected for further analysis due to their commonness in the forecasting literature, their ability to closely mimic real-world unpredictability, and their relative mathematical simplicity.

- **Point mass noise**: The noise term is presented by $+\sigma$ or $-\sigma$, each with probability $\frac{1}{2}$:

$$\varepsilon_{i,k} = \sigma \in (0, 1)$$

- **Gaussian noise**: The noise term is drawn from a normal distribution with mean 0 and variance σ^2:

$$\varepsilon_{i,k} \sim \mathcal{N}(0, \sigma^2)$$

For comparing these noise terms, first the variance has to be calculated. With the variance of a random variable X being $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

for point mass the expectation and variance are:

$$\mathbb{E}[\varepsilon_{i,k}] = 0.5 \cdot \sigma + 0.5 \cdot (-\sigma) = 0$$

$$\mathrm{Var}(\varepsilon_{i,k}) = \mathbb{E}\left[\left(\varepsilon_{i,k} - \mathbb{E}[\varepsilon_{i,k}]\right)^2\right] = \frac{1}{2}(\sigma - 0)^2 + \frac{1}{2}(-\sigma - 0)^2 = \sigma^2$$

For a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the variance is simply:

$$\mathrm{Var}(\varepsilon_{i,k}) = \sigma^2.$$

By setting the parameter σ to be identical in both models, we ensure that the expected magnitude of forecasting error is the same:

$$\mathrm{Var}_{\text{Point Mass}} = \mathrm{Var}_{\text{Gaussian}} = \sigma^2.$$

With this normalization a controlled setting for comparing the two noise models is set. In figure 4.3, the difference in frequencies in Brier scores can be found. The two noise models are applied to a set of realized outcomes and the Brier scores are calculated over the noisy outcomes of $n = 5$ forecasters. To make sure the forecasts are inside the interval $[0, 1]$ a clip function is used: $clip(\theta_k + \sigma, 0, 1)$, this function ensures that every forecast outside the interval will be denoted as 0 or 1, depending on which one was closest. The graph shows no large deviations. The Gaussian noise seems to have more frequency in scores around the average of scores than the point mass, where the point mass seems to have slightly more extreme scores. The graph is made for choosing a noise model and to detect any unrealistic outcomes in Brier scores. Because there are no sufficiently large deviations in the two models, it is not of great influence which noise model is chosen for the upcoming simulations. The point mass model is therefore used throughout this paper.

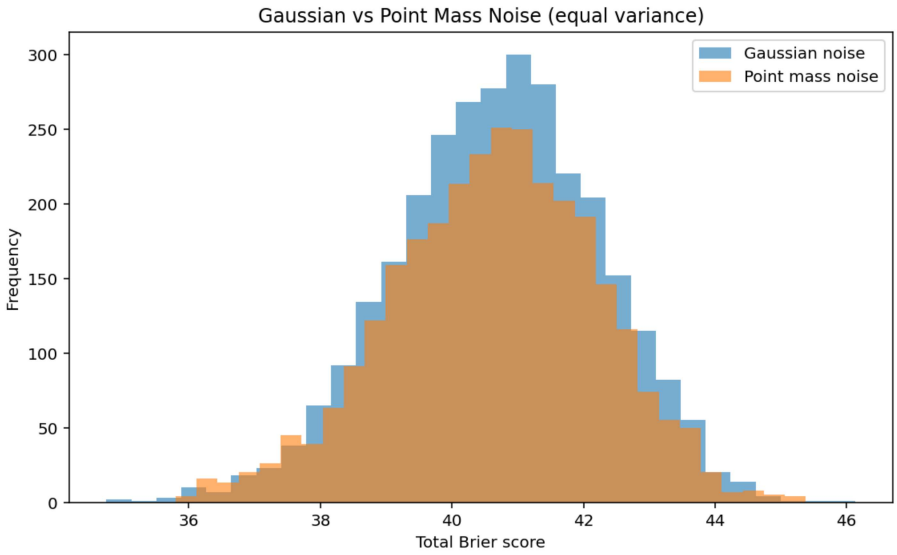


Figure 4.3: Brier score frequencies with point mass and gaussian noise models

5

The WOMAC mechanism

Srinivasan et al. (2025) introduced another mechanism, for setting up a prediction tournament, called WOMAC (Wisdom of the Most Accurate Crowd). Instead of just comparing each forecaster directly to the realized outcome, which can be noisy, WOMAC does something else. For each forecaster and each event, it ignores that forecaster's prediction and looks at the other forecasters' predictions. It creates a reference prediction from the remaining forecasters, asking what the group of other forecasters would predict for this event. The forecaster is then scored by how close their prediction is to this reference prediction. This reference prediction is designed to incentivize truthful and accurate reporting by forecasters (Kilgour, 2003).

Let $W \in \mathbb{R}^{m \times n}$ denote the matrix of predictions, where m is the number of events and n is the number of forecasters. Let $\mathbf{y} \in \mathbb{R}^m$ denote the vector of realized outcomes. For each event $k \in \{1, \ldots, m\}$ and each forecaster $i \in \{1, \ldots, n\}$, we remove both event k and forecaster i and learn an aggregation function using the remaining data. Formally, we define

$$\beta_{ki} = \arg\min_{\beta'} \left\| \mathbf{y}_{-k} - f_{\beta'}(W_{-ki}) \right\|^2,$$

where:

- $\mathbf{y}_{-k}$ denotes the outcome vector with event k removed,
- W_{-ki} denotes the prediction matrix with event k and forecaster i removed,
- $f_{\beta'}$ is a parametric aggregation function

A parametric aggregation function is a function that combines multiple predictions into a single aggregated prediction using a fixed number of parameters β'.This aggregation function f should be constructed specifically for the event k in question, rather than applied uniformly across all events. Linear regression can be seen as a common choice, which works fine for $m > n$. However, prediction tournaments can have more forecasters than predictions ($n > m$). Then marginal feature screening can be used to select a smaller effective n. (Srinivasan et al., 2025)

After receiving the learned parameter β_{ki}, a reference prediction is computed for forecaster i on event k based on the other forecasters:

$$t_{ki} = f_{\beta_{ij}}(\mathbf{w}_{k,-i}), \tag{5.1}$$

where $\mathbf{w}_{k,-i}$ is the vector of predictions for event k excluding forecaster i. Next, each forecaster i is assigned a WOMAC score defined as the squared deviation from the corresponding reference predictions:

$$S_i = \sum_{k=1}^{m} (w_{ki} - t_{ki})^2. \tag{5.2}$$

As the WOMAC mechanism does not use the Brier score that is used in the simulations of other mechanisms described in this paper, it is necessary to look at the properness and boundedness of this S_i.

Lastly, the WOMAC mechanism selects the expert with the lowest score:

$$i^* = \arg\min_i S_i.$$

The approach of WOMAC aims to reduce the impact of outcome noise and to better capture the relative predictive accuracy of forecaster. This reduces the effect of randomness in the outcome, meaning a forecaster will not just get "lucky" because the real outcome was unusual. It rewards forecasters for being closer to the consensus of accurate forecasters, which is usually closer to the true probabilities. However, this might not always be the case resulting in some randomness still.

5.1. Incentive-compatible

To use this found mechanism, it is firstly important to know if it is incentive-compatible. The following definitions from Srinivasan et al. (2025) therefore need to be presented. It is used that forecasters do not necessarily have access to the true probabilities, but receive a signal z_{ki} (for forecaster i and event k) that informs their estimate, and can therefore make a strategy to win the competition.

Definition 3 (Radially Symmetric) *We say a function $f_{\vec{c}} : \mathbb{R}^{d\times k} \to \mathbb{R}$ is* radially symmetric at $\vec{c} \in \mathbb{R}^d$ *if there exists a measurable function $h : \mathbb{R}^k \to \mathbb{R}$ such that*

$$f_{\vec{c}}(\vec{v}_1, \dots, \vec{v}_k) = h\Big(\|\vec{v}_1 - \vec{c}\|^2, \dots, \|\vec{v}_k - \vec{c}\|^2\Big)$$

For any $\vec{v}_1, \dots, \vec{v}_k \in \mathbb{R}^d$. That is, $f_{\vec{c}}$ is rotation-invariant in each argument around $\vec{c}$.

The function $f_{\vec{c}}$ treats all directions equally, it only cares about how far each vector is from the center $\vec{c}$. Rotating or reflecting vectors around $\vec{c}$ does not change the function's value.

Definition 4 (Strictly Radially Decreasing Function) *A function $f_{\vec{c}} : \mathbb{R}^d \to \mathbb{R}$ is said to be* strictly radially decreasing *from a center $\vec{c}$ if, for any two vectors $\vec{x}, \vec{y} \in \mathbb{R}^d$ such that*

$$\|\vec{x} - \vec{c}\|^2 < \|\vec{y} - \vec{c}\|^2,$$

it holds that

$$f_{\vec{c}}(\vec{x}) > f_{\vec{c}}(\vec{y}).$$

The closer a point is to the center $\vec{c}$, the higher the function value. This makes a smooth hill that peaks at $\vec{c}$, where the function decreases as you move away from the peak.

To understand definition 5 a bit of game theory needs to be explained and mostly Bayesian games, meaning a strategic game with incomplete information. Strategies, being the set of probability reports submitted for the events, are made by forecasters, because they desire to win. These strategies are based on signals forecasters receive about the true probability of the outcomes. Their utility measures how successful they are, typically the probability of winning the tournament.

Definition 5 (Bayes-Nash Equilibrium) *A strategy profile* $\mathbf{s} = (s_1, \dots, s_n)$ *is a* Bayes-Nash equilibrium *if for every forecaster $i \in \{1, \dots, n\}$ and every signal z_{ki} received by forecaster i, given the beliefs about the other forecasters' types and strategies, the expected utility of playing s_i is at least as large as the expected utility of playing any alternative strategy s_i':*

$$\mathbb{E}\Big[U_i(s_i, s_{-i} \mid z_{ki})\Big] \geq \mathbb{E}\Big[U_i(s_i', s_{-i} \mid z_{ki})\Big], \quad \forall s_i' \in \mathcal{S}_i,$$

where s_{-i} denotes the strategy profile of all forecasters except i, and $\mathcal{S}_i$ is the set of feasible strategies for forecaster i. In other words, no forecaster can increase their expected utility by unilaterally deviating from s_i, conditional on their observed signal z_{ki}.

A forecaster's signal z_i is drawn from a distribution conditional on the true outcome θ:

$$z_i \sim Q_i(\cdot \mid \theta),$$

where Q_i is the signal distribution for forecaster i. This Bayes-Nash equilibrium conditions ensure that, given beliefs about other forecasters' signals, no forecaster can increase expected utility by unilaterally deviating from truthful reporting. This condition is therefore wanted for an incentive compatible mechanism. The following theorem demonstrates that for the WOMAC mechanism truthful reporting is a Bayes-Nash equilibrium:

Theorem 3 (Srinivasan et al., 2025) *Let there be n forecasters and m independent events. Denote each forecaster i's vector of predictions across events as $\mathbf{w}_i = (w_{1i}, \dots, w_{mi})$ and the realized outcomes as $\mathbf{x} = (x_1, \dots, x_m)$. Let $\mathbf{t}_i$ denote the reference (crowd-based) prediction for forecaster i. Then, under the WOMAC mechanism, truthful reporting*

$$\mathbf{w}_i = \mathbf{t}_i \quad \forall i$$

is a Bayes-Nash equilibrium if the following conditions hold:

1. *Each forecaster's posterior belief over the outcomes is **strictly radially decreasing** from the reference vector $\mathbf{t}_i$.*
2. *Each forecaster's expected posterior prediction is **radially symmetric** with respect to the ground truth $\mathbf{x}$.*
3. *The conditional probability densities of the forecasters' signals are from a location family.*

Where a location family describes a set of probability distributions that are identical in shape but may differ by a shift in location. For a mechanism that scores forecasters against the ground truth such as WOMAC, this theorem holds (Srinivasan et al., 2025). This makes the WOMAC mechanism Bayes-Nash incentive compatible. This is because the three conditions of theorem 3 hold. Let $\mathbf{w}_i \in \mathbb{R}^m$ denote the forecast vector of forecaster i over m events, and let $\mathbf{t}_i \in \mathbb{R}^m$ denote the reference prediction computed via leave-one-out aggregation. In the WOMAC mechanism, forecaster i is scored by formula 5.2, where t_{ki} is the reference prediction for event k.

For the first condition, to formally capture the average performance across all events, the expectation is taken of the negative score. The expected score is strictly decreasing with the Euclidean distance from the reference:

$$\mathbb{E}[-S_i(\mathbf{w}_i)] \text{ strictly decreases as } \|\mathbf{w}_i - \mathbf{t}_i\|_2 \text{ increases.}$$

Hence, the posterior belief of forecaster i is strictly radially decreasing from $\mathbf{t}_i$.

For the second condition, since the WOMAC score depends only on the squared Euclidean distance:

$$S_j(\mathbf{w}_i) = \|\mathbf{w}_i - \mathbf{t}_i\|_2^2,$$

it is rotation-invariant around $\mathbf{t}_i$. Formally, for any $\mathbf{w}_i, \mathbf{w}_i'$ with

$$\|\mathbf{w}_i - \mathbf{t}_i\|_2 = \|\mathbf{w}_i' - \mathbf{t}_i\|_2,$$

we have

$$\mathbb{E}[-S_i(\mathbf{w}_i)] = \mathbb{E}[-S_i(\mathbf{w}_i')],$$

establishing radial symmetry.

The last condition holds because, assume forecaster signals are generated as:

$$\mathbf{w}_i = \mathbf{x} + \epsilon_i, \quad \epsilon_i \sim Q_i,$$

Where x is the true outcome vector for all events. ε_i is a random error drawn from a location-family distribution Q_i. The leave-one-out reference $\mathbf{t}_i = f_{\beta_i}(\mathbf{w}_{-i})$ is a linear function of other forecasters, which preserves the location-family property. Thus, the conditional distributions satisfy the location-family assumption:

$$\mathbf{w}_i \mid \mathbf{w}_{-i} \sim \text{location-family centered at } \mathbf{t}_i.$$

5.2. Comparison with the ELF mechanism

The WOMAC mechanism does not use the randomness in the found mechanisms of ELF and I-ELF; it takes a fundamentally different approach to the problem of incentive compatibility. WOMAC challenges the assumption that deterministic mechanisms must fail. Srinivasan et al. (2025) proves that deterministic competitions can be Bayes-Nash incentive-compatible as shown in section 5.1. When experts are scored against a noiseless estimate of the ground truth, it becomes clear that outcome noise, rather than determinism itself, is the main barrier to incentive compatibility in standard competitions. This is a notable departure from Theorem 1 Witkowski et al. (2021), which showed that no deterministic mechanism can be incentive compatible. WOMAC achieves the weaker Bayes-Nash variant instead.

WOMAC does not achieve dominant strategy incentive compatibility in the sense of Witkowski et al. (2021), meaning that truthful reporting is optimal only given beliefs about other forecasters' behavior, not regardless of what others report. Nevertheless, empirical results on real life forecasting datasets show that WOMAC is a more reliable predictor of forecasters performance than the standard mechanism (Srinivasan et al.,2025), suggesting that it may be valuable in practice even where its theoretical incentive property is slightly weaker.

6

Minimising Randomness in Incentive-Compatible Mechanisms

In the previous sections, different kinds of mechanisms have been evaluated and their incentive compatibility has been established. However, by the making of these mechanisms, some randomness is added in choosing a winner for their tournament, or in case of the WOMAC mechanism, the forecasters are scored against a reference prediction made from predictions of other forecasts adding some amount of randomness in not being scored against the realized outcomes. Scoring against the realized outcomes will possibly result in a different outcome in winner. Consequently, the highest scored forecaster (against the realized outcome) does not have the highest probability of winning, which can be described as randomness in the WOMAC mechanism. Since one of the goals of this thesis is to study tournaments with as little randomness as possible, it is useful to quantify the amount of randomness used by a mechanism.

6.1. Formalizing randomness

To be able to quantify the amount of randomness involved in the different kinds of mechanisms, it first needs to be established how randomness should be measured. This can be done using different methods, and these will be explained in the following subsections.

6.1.1. Expectation

The first way to measure randomness is as follows. Form a list where forecasters are ranked based on their scores, together with one of the mechanisms described in this thesis (deterministic, ELF, I-ELF or WOMAC), a winner of a prediction tournament is chosen. Let X be a random variable equal to the rank number. Then $\mathbb{E}[X]$ can be taken as a measurement of the expected rank that is chosen as the winner, since the winner of a tournament is not always the first ranked forecaster. The expected rank of the tournament winner is denoted by:

$$\mathbb{E}[X] = \sum_{i=1}^{n} i \cdot \mathbb{P}(\text{winner has rank } i),$$

where n is the number of forecasters and $\mathbb{P}(\text{winner has rank } i)$ is the probability that the forecaster with rank i (based on total scores) is selected as the winner. If $\mathbb{E}[X] = 1$, the mechanism always selects the first ranked forecaster. In this case, there is no randomness in the mechanism. This is true for a deterministic mechanism that always selects the first ranked forecaster. If $\mathbb{E}[X] > 1$, the mechanism sometimes selects lower ranked forecasters. The higher the $\mathbb{E}[X]$, the more randomness in the tournament outcome.

In tables 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6, an example is proposed using this measurement of randomness with the ELF mechanism. The simulated noise for this example is chosen from $\sigma \in [0.05, 0.10, 0.15, 0.20, 0.25]$.

Six simulations are produced with $n = 5$ forecasters and $m = 50$ events. After the simulations, the expected rank is calculated as $\mathbb{E}[X] = 2.5$

Rank	Forecaster	Score	Probability
1	A	42.722	0.211
2	B	41.321	0.204
3	C	39.949	0.198
4	E	39.164	0.194
5	D	38.973	0.193

Table 6.1: Winner rank 4: E

Rank	Forecaster	Score	Probability
1	A	41.892	0.212
2	B	40.900	0.207
3	C	39.405	0.199
4	D	38.719	0.196
5	E	36.920	0.187

Table 6.2: Winner rank 3: B

Rank	Forecaster	Score	Probability
1	C	42.926	0.208
2	B	41.828	0.203
3	A	41.687	0.202
4	E	40.659	0.197
5	D	39.537	0.191

Table 6.3: Winner rank 3: A

Rank	Forecaster	Score	Probability
1	B	42.358	0.203
2	D	42.111	0.202
3	E	42.034	0.202
4	A	41.774	0.200
5	C	40.349	0.193

Table 6.4: Winner rank 1: B

Rank	Forecaster	Score	Probability
1	C	42.218	0.204
2	A	41.912	0.203
3	B	41.245	0.199
4	D	40.864	0.197
5	E	40.719	0.197

Table 6.5: Winner rank 1: C

Rank	Forecaster	Score	Probability
1	E	40.937	0.208
2	A	40.239	0.205
3	B	39.579	0.201
4	C	38.689	0.197
5	D	36.994	0.189

Table 6.6: Winner rank 3: B

The I-ELF mechanism selects its winner differently, though it uses an identical ranking system. The WOMAC mechanism can also be evaluated by this measure of randomness. While the Brier score is not applied within this mechanism, it can be calculated separately to produce a ranking.

6.1.2. Shannon entropy

The following way of measuring randomness is by using the Shannon entropy (Shannon, 1984). Suppose a mechanism selects a winner according to the probability vector

$$p = (f_1, \dots, f_n)$$

where f_i is the probability that forecaster i wins. The Shannon entropy of this distribution is defined as

$$H(p) = -\sum_{i=1}^{n} f_i \log_2(f_i).$$

The term $\log_2(f_i)$ is not defined for $f_i = 0$. this can be resolved by defining:

$$0 \cdot \log_2 0 := 0,$$

This can be done because the contribution of a forecaster with zero probability of winning should not affect the total entropy. If one forecaster wins with probability one, then the mechanism is deterministic and

$$H(p) = 0.$$

If all forecasters are equally likely to win, so that $f_i = \frac{1}{n}$ for all i, then the entropy is maximal:

$$H(p) = \log_2(n).$$

To make entropy comparable across tournaments with different numbers of forecasters, normalized entropy is used:

$$H_{\text{norm}}(p) = \frac{H(p)}{\log_2(n)}.$$

Thus,

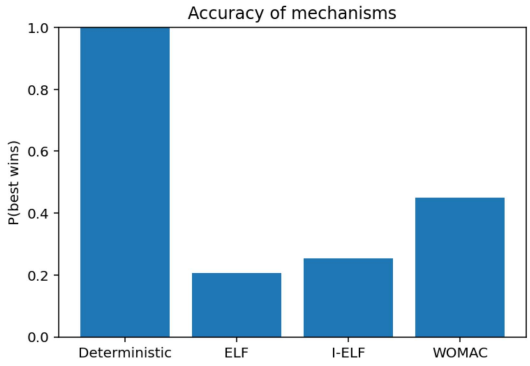
$$H_{\text{norm}}(p) \in [0, 1].$$

A value of 0 means that no randomness is used, while a value of 1 means that the mechanism is maximally random. This measure allows for studying the trade-off between selecting the best forecaster and minimizing the amount of randomness used in the tournament.

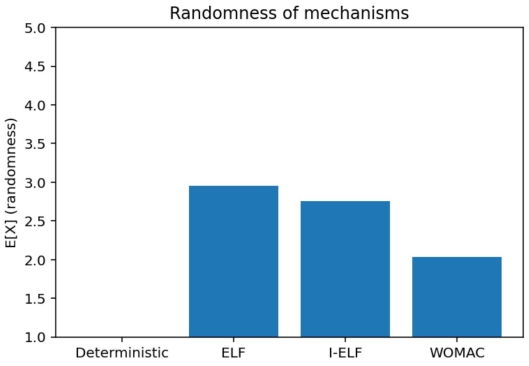
The Shannon entropy is based on a probability vector p only to be found in the ELF mechanism. For the I-ELF and WOMAC mechanisms, this entropy measure cannot be applied directly. WOMAC does not assign winning probabilities to forecasters, and while I-ELF does so for individual events, the tournament winner is ultimately determined by which forecaster wins the most events overall, rather than by a single probability distribution over winners. An alternative way of applying Shannon entropy to these mechanisms may exist, but it is not immediately apparent. This is the reason that, for the upcoming simulations, expectation is used as a measure of randomness.

6.2. Simulating Randomness

After establishing how to measure randomness, simulations of forecasters' performance under various mechanisms can be made. The prediction tournaments will be simulated with n forecasters and m binary events. For each event $k = 1, \dots, m$, a true probability $\theta_k \in [0, 1]$ is randomly drawn from a uniform distribution. The outcome is then made according to $X_k \sim \text{Bernoulli}(\theta_k)$. Each forecasters prediction is generated by adding a noise term $\varepsilon_{i,k} = \sigma$ to the true event probability θ_k, where the point mass noise model is used for these simulations. Because only the forecasts need to be in the interval $[0, 1]$, a clip function is used, denoted as $clip(\theta_k + \sigma, 0, 1)$. This function ensures that if a forecast is generated outside the interval, it is changed to the nearest value inside the interval, being 0 or 1. In section 2.1, different types of scoring rules are presented. These are necessary for the scoring of the forecasters. For the simulations the Brier score is chosen, because it is strictly proper, bounded and relatively easy. The winner is then chosen under one of the discussed mechanisms: *Deterministic*, *ELF*, *I-ELF* or *WOMAC*. For the WOMAC mechanism, an aggregation function needs to be chosen for the simulations. Linear regression is used to estimate a reference prediction for each event based on the predictions of all other forecasters. It fits a line through the remaining forecasters predictions to best predict the outcome, minimizing the squared difference between predicted and actual values. Linear regression is chosen because it is simple, it naturally aligns with the squared-error scoring rule and allows efficient leave-one-out computation for each forecaster and event, ensuring that truthful reporting remains the best strategy. The simulation for these four mechanisms is repeated 1000 times to first estimate the probability that the most accurate forecaster wins $\mathbb{P}(\text{best forecaster wins})$ and then the randomness, given in figure 6.1 with $n = 5$ forecasters and $m = 50$ events. The accuracy and randomness in the different mechanisms are found, where the WOMAC mechanism seems the most accurate and contains the least amount of randomness of the researched incentive-compatible mechanisms.



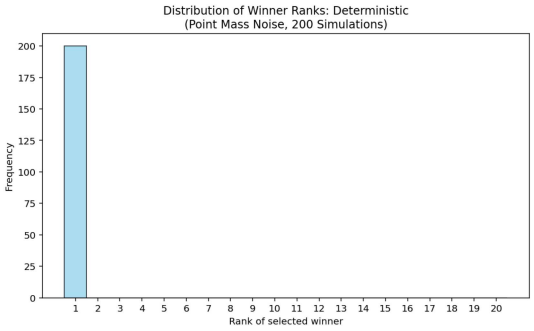
(a) Accuracy



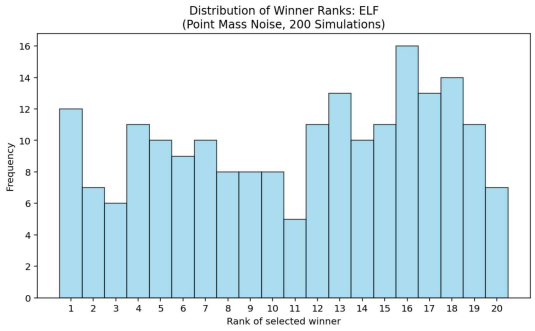
(b) Randomness

Figure 6.1: The accuracy and randomness of the different mechanisms shown by $\mathbb{P}$(best forecaster wins) and $\mathbb{E}$[rank of tournament winner]

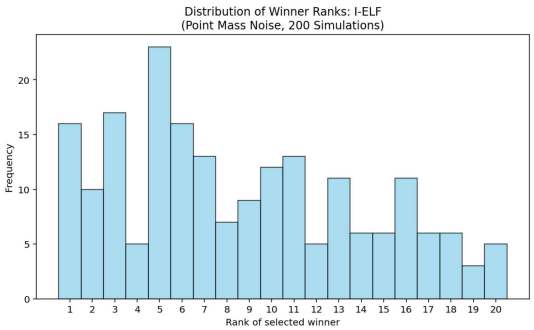
To see what really happens in these mechanism, it is useful to simulate the outcomes in a different way as well. In figure 6.2, the simulations are presented with error term $\sigma \in [0.05, 0.10, 0.15, 0.20, 0.25]$, $n = 20$ forecasters, $m = 400$ events. In 200 simulations, it can be seen how many times each forecaster in a certain rank, based on their total scores, is chosen as the tournament winner.



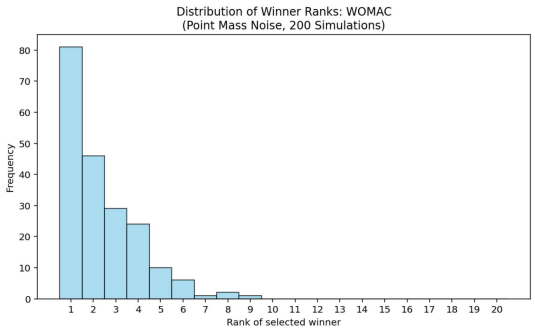
(a) Standard mechanism



(b) ELF mechanism



(c) I-ELF mechanism



(d) WOMAC mechanism

Figure 6.2: 200 simulations for Randomness measured over $n = 20$ and $m = 400$

The randomness in the standard mechanisms is shown to be zero, the forecaster with the highest Brier score always wins the tournament, this is how the standard mechanism works. However, in figures 6.2b and 6.2c, the randomness is shown quite clearly. The forecaster that has the highest Brier score does not win the most of the tournaments, even the forecaster with the lowest score wins the tournament multiple times. The tables 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6 show that if the error σ of the forecasters is relatively low, that the probabilities of winning will be very close to each other, meaning all forecasters almost have the same probability of winning. This results in a very high randomness factor and can explain why in the ELF mechanisms it seems like the lower ranked forecasters are cho-

sen as winners more often. Figure 6.2b comes close to a uniform distribution, ELF is therefore almost completely random.

Looking at WOMAC, nonetheless, the highest scored forecaster wins the most of the tournaments, decreasing by rank, showing significantly less randomness in this mechanism. Meaning that the reference predictions made in formula (5.1) seem to be close to the true probabilities of the simulated prediction tournaments.

7

Event quantity Influence on the randomness in I-ELF

After evaluating the randomness in the different mechanisms, it is interesting to look at the importance of the number of events. In the ELF mechanism, Frongillo (2021) established the need for $O(n \log n/\varepsilon^2)$ events to decrease the amount of randomness, which is significantly large. Considering the I-ELF mechanism, which is specifically designed to reduce randomness as the number of events increases, so that $\lim_{m\to\infty} \mathbb{P}(\text{best forecaster wins}) = 1$ (Witkowski et al., 2021), it is worth investigating whether this mechanism can achieve high reliability with fewer events. In this chapter the number of events are studied with a lower bound and with simulations.

7.1. Hoeffding-based lower bound for event number

By applying the Hoeffding inequality (Theorem 4 below) to establish a limit on accuracy (Hoeffding, 1963), a lower bound on the number of events required for the I-ELF mechanism is derived. This lower bound is informative, as it indicates how many events may be needed to reduce the randomness in the mechanism to a level where it becomes negligible. Hoeffding's inequality states that the probability of the actual sum S_m deviating from its expected value by more than t decreases exponentially as the number of variables m increases.

Theorem 4 (Hoeffding's inequality) *Let $X_1, \dots, X_m$ be independent random variables bounded by the interval $[0, 1]$. Define*

$$S_m = X_1 + X_2 + \cdots + X_m.$$

Then, for any $t > 0$,

$$\mathbb{P}\big(S_m - \mathbb{E}[S_m] \geq t\big) \leq \exp\Big(-\frac{2t^2}{m}\Big),$$

and

$$\mathbb{P}\big(\mathbb{E}[S_m] - S_m \geq t\big) \leq \exp\Big(-\frac{2t^2}{m}\Big).$$

The lower bound is given by:

$$m \geq \frac{2(n-1)^2}{\varepsilon^2} \log\left(\frac{4(n-1)}{\delta}\right) \tag{7.1}$$

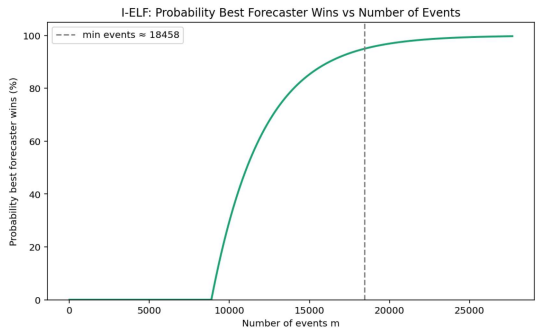
ε gives the accuracy gap between the best and second-best forecaster, this accuracy gap is given by the difference between the expected scores of the most accurate and the second-most accurate forecaster. The following notation is for the expected score of report y_i given scoring rule R and the joint probability θ with m events.

$$R(y_i, \theta) := \mathbb{E}_{X\sim\theta}\left[\frac{1}{m}\sum_{k=1}^{m} R(y_{i,k}, X_k)\right]$$

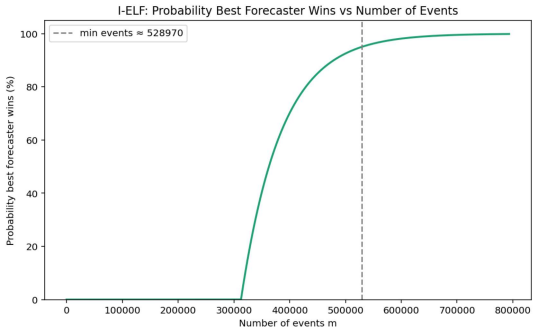
$R(y_i, \theta)$ gives the expected average score that forecaster i would receive, because we want a measure of how well a forecaster is expected to perform on average in all events. The accuracy gap is then measured by:

$$\varepsilon := \min_{j \neq i} \left(\max_{i} R(y_i, \theta) - R(y_j, \theta) \right)$$

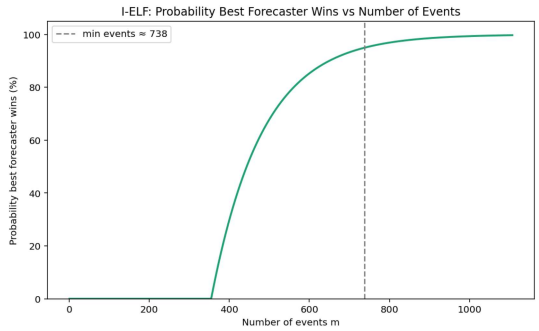
Furthermore, n is given by the number of forecasters and δ by the desired failure probability. this failure probability defines how confident we want to be that the top forecaster actually wins in the tournament. $1 - \delta$ is measured by the probability that the highest-scoring forecaster wins a tournament, making δ the failure probability (probability that the best forecaster is not chosen as the winner). This δ can be adjusted to the desired amount of randomness in prediction tournaments. To see the results of different n, ε and δ some simulations of equation 7.1 are shown in figure 7.1.



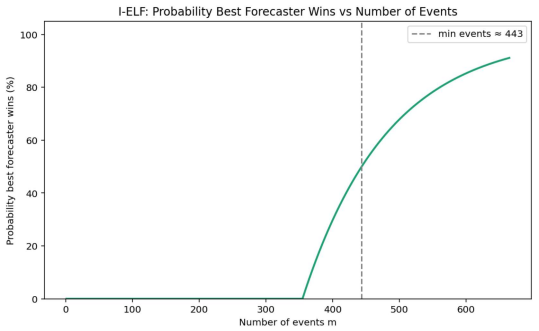
(a) $n = 5, \varepsilon = 0.1, \delta = 0.05$



(b) $n = 20, \varepsilon = 0.1, \delta = 0.05$



(c) $n = 5, \varepsilon = 0.5, \delta = 0.05$



(d) $n = 5, \varepsilon = 0.5, \delta = 0.5$

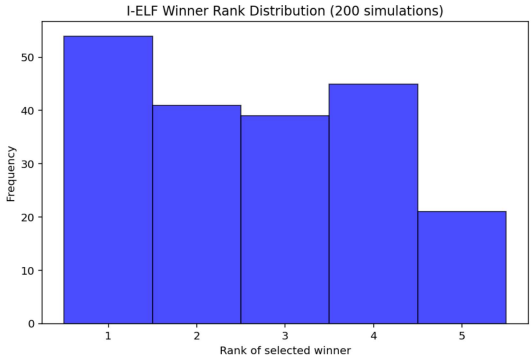
Figure 7.1: simulations of equation 7.1 given by different n,ε and δ

With small values of ε and δ (figure 7.1a), representing near-zero desired randomness, the derived lower bound on the number of required events is substantially large, reaching approximately $m = 18458$. This number grows even further as the number of participants increases seen in figure 7.1b. Even under more lenient conditions, such as a $\delta = 50\%$ failure rate with an accuracy gap of $\varepsilon = 0.5$ (figure 7.1d), the required number of events only reduces to $m = 443$, which remains considerably large. It can be concluded that the lower bound derived using the Hoeffding inequality is not yet practically applicable in prediction tournaments. Therefore, the I-ELF mechanism can not yet be seen as a mechanism with the least amount of randomness for a realistic amount of events.

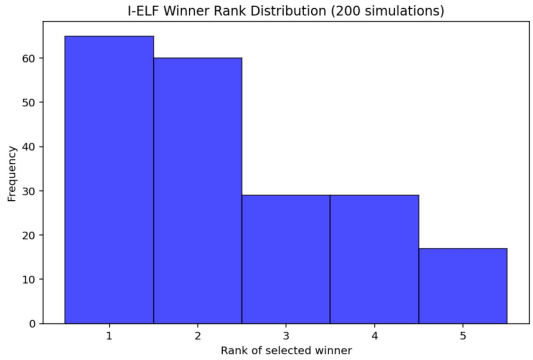
7.2. Simulations for event number lower bound

To know if the lower bound found in equation 7.1 is somewhat realistic, simulations can be done with different number of events in the I-ELF mechanism. To compare this to the Hoeffding lower bound in figure 7.1 it is important to look at the same number of forecasters n and look at the percentage of randomness involved δ. Since the noise of each forecaster is simulated randomly, the accuracy gap between the best and second best forecaster ε cannot be predetermined. The simulations for searching for this lower bound on the events of I-ELF are done in the same way as the simulations in

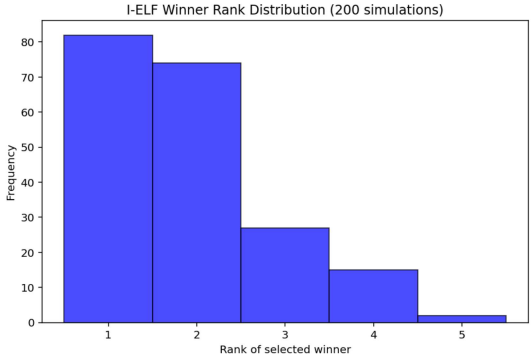
section 6.2 by looking at the rank of the chosen winner in each simulation. The number of forecasters for the simulations will be $n = 5$ to compare with the plots in figures 7.1a, 7.1c and 7.1d.



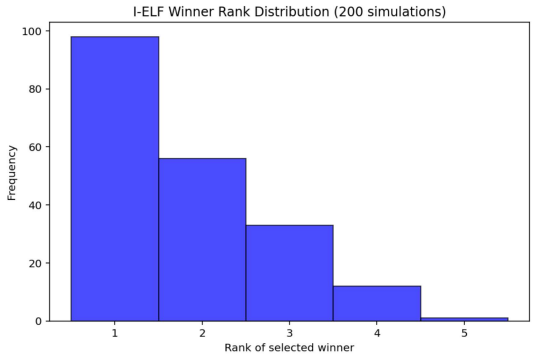
(a) $n = 5, m = 400$



(b) $n = 5, m = 1000$



(c) $n = 5, m = 5000$



(d) $n = 5, m = 10000$

Figure 7.2: 200 Simulations of events with $n = 5$ forecasters based on the rank of the chosen winner in each simulation.

These simulations result in the plots in figure 7.2 and calculate the probability that the first ranked forecaster will win for both I-ELF and WOMAC, to have a good comparison. The different plots in figure 7.2 show event numbers of size $400, 1000, 5000$ and 10000, the plots clearly show a decrease in randomness when the number of events increase. However, since 400 events is already a large number of events, figure 7.2a displays already more randomness than desired. The fact that there is more randomness than desired can also be found in the comparison of winning probabilities of I-ELF and WOMAC. In the simulation with 10000 events, the I-ELF mechanism gets a percentage of 44.50% for the first ranked forecaster to win, which gives $\delta = 0.555$. For the WOMAC mechanism, the percentage is 92.5%, which gives $\delta = 0.075$. This is a significantly large difference, suggesting that increasing the number of events does not meaningfully reduce I-ELF's randomness to a level comparable with that of WOMAC.

Since figure 7.1b showed that increasing the number of forecasters results in a higher bound, representing these simulations with $n = 5$ forecasters should be good enough to conclude that with increased number of events, WOMAC still outperforms I-ELF.

8

Conclusion

In this thesis, the design of prediction tournaments has been examined with the objectives being incentive compatibility and minimal randomness in winner selection. The reason for this research was the established prediction tournament paradox, where in a standard winner-take-all competition, the most accurate forecaster does not always have the highest probability of winning, because more extreme predictions can result in higher scores despite being less accurate on average. To get a mechanism for prediction tournaments, simulations had to be done to find an optimal mechanism. Although different kind of mechanisms have been established and studied in prior literature, their behavior has rarely been visualized or compared empirically. Since visual representation is essential for better comparison between mechanisms, this thesis introduces simulations to address that gap.

Firstly scoring rules were evaluated and the Brier score (Brier, 1950) was chosen because of its strict properness, boundedness and relative easiness. For the simulations, a noise model was also needed. This is because the simulations need to mimic real life prediction tournaments as well as possible. After evaluating different kinds of noise models and comparing them through equal variance, the point mass noise model was chosen. There were relatively few differences with the Gaussian model, however point mass was easy to implement and used in past literature and therefore chosen.

Because prediction tournaments are build from a mechanism, this is the most important part where incentive compatibility and randomness is decided. The following mechanisms are researched; Deterministic, ELF, I-ELF and WOMAC. The research showed ELF and I-ELF to be incentive compatible mechanisms, however they seemed to be the most random of the mechanisms. This can be seen in figure 6.2b and 6.2c where 200 simulations are done with error $\sigma \in [0.05, 0.10, 0.15, 0.20, 0.25]$ over $m = 400$ events. In the plot for the ELF mechanism it can be found, after ranking 20 forecasters based on their scores, that all the forecasters seem to have almost the same probability of winning, even the last ranked forecaster wins the competition a couple of times. The plot for I-ELF looks almost the same, with a peak at forecaster with rank 5. However, the higher ranked forecasters won slightly more then in the ELF mechanism, resulting in a slight preference to I-ELF. Next to ELF and I-ELF a third incentive-compatible mechanism is researched. First, Bayes-Nash incentive-compatibility is found for this mechanism, which can be seen as slightly weaker, however still incentivizes truthful reportings. In Figure 6.2d, a positive outcome is observed: forecasters are again ranked, and the first ranked forecaster wins the majority of the tournament, with the probability of winning decreasing for lower-ranked forecasters. Comparing the WOMAC with the ELF and I-ELF mechanisms, WOMAC seems to win in terms of least randomness involved in the mechanism. WOMAC has a slightly weaker version of incentive compatibility, however this difference is not as big as the randomness difference shown in the presented plots. Therefore, choosing the WOMAC mechanism for building a prediction tournament will give the highest chance the most accurate and highest scored forecaster will win.

For the I-ELF mechanism some more simulations were conducted. This mechanism was made to reduce randomness in choosing a winner as the number of events grow. If a prediction tournament would consist of a high amount of events, maybe this mechanism could out perform WOMAC. With the

help of the Hoeffding inequality, a lower bound for the number of events needed to only have a certain amount of randomness left, is build in equation 7.1. With the help of this lower bound simulations were made in figure 7.1. This lower bound however resulted in an unrealistic amount of events needed for I-ELF to be a better choice than WOMAC. To make sure that this lower bound gives the right outcomes, extra simulations were done with different amounts of events. However, these simulations have similar outcomes, again resulting in an unrealistic amount of events needed to have less randomness than WOMAC.

To summarize, for organizations seeking to design a prediction tournament that reliably identifies the most accurate forecaster, the WOMAC mechanism is recommended under typical tournament conditions. It offers the best empirical performance in terms of selecting the best forecaster, introduces the least randomness among the incentive-compatible mechanisms studied, and its slightly weaker Bayes-Nash incentive guarantee remains sufficient to incentivize truthful reporting in practice. Future research could extend simulations to a broader range of settings. For instance, varying the number of forecasters and events and experimenting with different noise models for forecasts. This would provide a more comprehensive understanding of how mechanisms perform across different tournament scenarios, resulting in more robust simulations. It can also be good to focus on deriving tighter bounds on the number of events required for I-ELF to outperform WOMAC, exploring whether a mechanism exists that achieves both dominant strategy incentive compatibility and low randomness simultaneously, and extending the analysis to prediction tournaments involving correlated events or continuous outcome spaces, where the current framework may need to be adapted.

9 Discussion

This thesis provides a systematic comparison of prediction tournament mechanisms and offers recommendations for prediction tournament organizers. However, several limitations and directions for future research need some careful consideration. An assumption made in the beginning of the thesis is that all events in the prediction tournament are independent of each other. This assumption was made to simplify mathematically and to allow for clean comparisons between mechanisms, but it does not always hold in practice. In many real-world forecasting settings, events are correlated. As an example, you can take the outcome of a geopolitical event that may influence the likelihood of related economic or political developments. When events are correlated, the scoring and aggregation procedures used in the ELF, I-ELF, and WOMAC mechanisms may no longer behave as intended. The independence assumption is also used in the proof of incentive compatibility for both ELF and I-ELF (Witkowski et al., 2021). With correlated events, this might not even hold. Frongillo et al. (2023) studied forecasting competitions with correlated events and show that correlation between events introduces additional complexity into the mechanism design problem. Correlated events in prediction tournaments therefore require further research.

Another limitation concerns the lower bound on the number of events derived in Chapter 7 using Hoeffding's inequality. This inequality is known to be loose in many settings, it does not exploit the specific distributional properties of the forecasting errors or the score differences between forecasters. As a result, the derived bound can be unrealistically large. A tighter bound could potentially be obtained by applying Bernstein's inequality instead, which accounts for the variance of the random variables in addition to their range (Frongillo et al., 2021). Applying this to the score differences in the I-ELF mechanism may yield a smaller lower bound on the number of events required to have a small amount of randomness, potentially making I-ELF a more practically viable alternative to WOMAC than the current analysis suggests. With the help of more simulations done in figure 7.2 with different amounts of events, the lower bound for the events was not improved in simulation setting. Because the forecasters noise σ is chosen randomly in these simulations, the accuracy gap between the best and second best forecaster ε could not be predetermined. Figure 7.1c suggested that when increasing ε, the number of events will decrease significantly. In further research, more simulations can be done where the ε can be predetermined, which can lead to interesting results.

Not only the I-ELF mechanism has a potential of being better with further research, other mechanisms can also be found that might perform better in prediction tournaments. For example, Frongillo et al (2021) did some further research finding the Follow the Regularized Leader (FTRL) mechanism. FTRL is an online learning framework in which forecasters update their predictions sequentially across events, and the mechanism selects a winner based on a regularized version of the cumulative score. Compared to I-ELF, FTRL may require fewer events to achieve a comparable level of accuracy in winner selection, to such a point that it may come close to the WOMAC mechanism in the sense of randomness. FTRLs online nature makes it particularly well suited for settings where forecasts are updated dynamically as new information becomes available. This mechanism is not simulated yet so it still needs to be compared to the mechanisms researched in this paper.

Finally, it is important to note that the evaluation of the WOMAC mechanism in this thesis is based entirely on simulated data generated under controlled noise models. While the simulations consistently show WOMAC to outperform ELF and I-ELF in terms of selecting the best forecaster with the least randomness, it is not guaranteed that this advantage will be observed identically compared to real-world prediction tournaments. In real life, forecasters might be more diverse in their predictions, resulting in a completely different reference prediction, which is very important for this mechanism. To know if WOMAC mechanism will perform well, it needs to be tested by several prediction tournaments, to see how the simulation results will translate to practice.

The performance of WOMAC relies heavily on the quality of the reference predictions constructed from the remaining forecasters, which in turn depends on the number of participants, the diversity of their predictions, and the degree to which their signals are informative about the true outcome. In settings with few forecasters, highly homogeneous predictions, or strong correlations between forecasters' signals, the reference prediction may be a poor proxy for the true probability, potentially undermining the mechanism's accuracy and incentive properties. Empirical validation of WOMAC on real forecasting datasets, beyond the results reported by Srinivasan et al. (2025), would therefore be an important step before recommending its adoption in practice.

Bibliography

[1] D. J. Aldous (2019). *A prediction tournament paradox*. The American Statistician, vol. 75, no. 3, pp. 243–248. General, Taylor & Francis.

[2] G. W. Brier (1950). *Verification of forecasts expressed in terms of probability*. Monthly Weather Review, vol. 78, no. 1, pp. 1–3. War Department, Office of the Chief Signal Officer.

[3] A. P. Dawid and M. Musio (2014). *Theory and applications of proper scoring rules*. Metron, vol. 72, no. 2, pp. 169–183. Springer.

[4] V. M. van der Eng (2025) *Analysis of the Prediction Tournament Paradox*, bachelor's thesis, TU delft.

[5] R. Frongillo et al. (2021). *Efficient competitions and online learning with strategic forecasters*. Proceedings of the 22nd ACM Conference on Economics and Computation, pp. 479–496.

[6] R. Frongillo et al. (2025). *Forecasting competitions with correlated events*. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 13, pp. 13873–13880.

[7] T. Gneiting and A. E. Raftery (2007). *Strictly proper scoring rules, prediction, and estimation*. Journal of the American Statistical Association, vol. 102, no. 477, pp. 359–378. Taylor & Francis.

[8] I. J. Good (1952). *Rational decisions*. Journal of the Royal Statistical Society: Series B (Methodological), vol. 14, no. 1, pp. 107–114. Wiley.

[9] Good Judgment Project (2026). *Website*. Available at: `https://goodjudgment.com/` (accessed May 1, 2026).

[10] W. Hoeffding (1963). *Probability inequalities for sums of bounded random variables*. Journal of the American Statistical Association, vol. 58, no. 301, pp. 13–30. Taylor & Francis.

[11] Kaggle (2026). *Competitions platform*. Available at: `https://www.kaggle.com/competitions` (accessed May 5, 2026).

[12] F. G. Kilgour (2004). *An experiment using coordinate title word searches*. Journal of the American Society for Information Science and Technology, vol. 55, no. 1, pp. 74–80. Wiley.

[13] K. C. Lichtendahl Jr. and R. L. Winkler (2007). *Probability elicitation, scoring rules, and competition among forecasters*. Management Science, vol. 53, no. 11, pp. 1745–1755. INFORMS.

[14] J. E. Matheson and R. L. Winkler (1976). *Scoring rules for continuous probability distributions*. Management Science, vol. 22, no. 10, pp. 1087–1096. INFORMS.

[15] N. Metropolis and S. Ulam (1949). *The Monte Carlo Method*. Journal of the American Statistical Association, vol. 44, no. 247, pp. 335–341. Taylor & Francis.

[16] T. R. Palfrey and S. W. Wang (2009). *On eliciting beliefs in strategic games*. Journal of Economic Behavior & Organization, vol. 71, no. 2, pp. 98–109. Elsevier.

[17] L. J. Savage (1971). *Elicitation of personal probabilities and expectations*. Journal of the American Statistical Association, vol. 66, no. 336, pp. 783–801. Taylor & Francis.

[18] C. E. Shannon (1948). *A mathematical theory of communications*. Bell System Technical Journal, vol. 27, pp. 379–423.

[19] S. Srinivasan et al. (2025). *WOMAC: A Mechanism For Prediction Competitions*. arXiv preprint arXiv:2508.17907.

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

[21] B. Waggoner (2017). *Lecture notes: Algorithmic Game Theory: Proper Scoring Rules and Prediction Markets*. Not published.

[22] J. Witkowski et al. (2023). *Incentive-compatible forecasting competitions*. Management Science, vol. 69, no. 3, pp. 1354–1374. INFORMS.