

Article

On the Level of Measurement of Sports Probabilities: Ordinal Behavior in Rare NBA Probability

Antonio Joaquín Segura García ¹, Ziwei Shu ² and Ramón Alberto Carrasco ^{2,*}¹ Faculty of Statistics, Complutense University of Madrid, 28040 Madrid, Spain; antosegu@ucm.es² Department of Marketing, Faculty of Statistics, Complutense University of Madrid, 28040 Madrid, Spain; ziweishu@ucm.es

* Correspondence: ramoncar@ucm.es

Abstract

Probabilities in sports forecasting are shaped by betting-market mechanisms, where expectations are exchanged through prices and transformed into implied probabilities. Although probabilistic predictions are usually treated as quantitative variables, rare events may be represented with lower resolution, making some probability ranges behave closer to ordinal structures. This paper investigates this hypothesis using NBA betting odds. We directly assess ordinality by examining the loss of linearity between implied probabilities and empirical outcome frequencies while evaluating whether monotonicity is preserved across probability ranges. The results show that low-probability odds deviate from the linear behavior expected under quantitative assumptions, while largely maintaining their ordinal ordering. The ordinal hypothesis is independently evaluated through a predictive experiment based on a Gated Recurrent Unit model. The model uses team-level temporal sequences as input features to predict the probability of a target game. We compare the behavior of the ordinal-aware CORAL loss across the analyzed subsets. The results suggest that the subset exhibiting ordinal characteristics attains a lower statistical risk under this ordinal inductive bias than the subset characterized by quantitative behavior. Finally, we discuss the broader applicability of comparative training with loss-function-induced inductive biases as a practical method for identifying the measurement level of target variables.

Keywords: level of measurement; sport probabilities; machine learning; Gated Recurrent Unit; Consistent Rank Logits

MSC: 62F01; 92B26; 68T05



Academic Editors: Manuel Alberto M. Ferreira and Zengjing Chen

Received: 18 May 2026

Revised: 19 July 2026

Accepted: 28 July 2026

Published: 4 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Sports prediction has been extensively studied in the academic literature, as highlighted in the meta-research of [1]. Mathematical models developed for forecasting sporting events provide a principled basis for the formation of prior expectations. These initial estimates are subsequently refined through a process of information aggregation in sports betting markets, where beliefs are updated via price formation mechanisms. While such markets incorporate information that may not be captured by formal models, they can also introduce systematic biases arising from behavioral factors, market constraints, or institutional features.

This paper investigates the impact of such biases on expected probabilities, which are inferred from observed market prices. In this context, betting odds—commonly expressed

in Money Line (ML) format—must be transformed into implied probabilities to enable statistical analysis. The standard conversion from Money Line odds to implied probability is given by:

$$P = \begin{cases} \frac{-ML}{-ML+100}, & \text{if } ML < 0, \\ \frac{100}{ML+100}, & \text{if } ML > 0, \end{cases} \quad (1)$$

where ML denotes the quoted Money Line value. This transformation allows market prices to be interpreted as probabilistic expectations, thereby facilitating the study of their statistical properties and potential deviations from observed frequencies.

Measurement theory distinguishes different types of data, including qualitative and quantitative variables, commonly organized into nominal, ordinal, interval, and ratio scales, each with distinct mathematical properties and permissible transformations [2,3]. While these categories are typically treated as intrinsic to a variable, in practice the same measurement may exhibit different behaviors depending on the range under consideration. In particular, probabilities, which are commonly assumed to be quantitative, may lose some of their metric properties in specific regions, behaving instead more like ordinal (qualitative) variables.

Ordinal scales have long been discussed in the statistical literature as a potential source of methodological pitfalls, often referred to the first of the “sins” of statistical analysis [4]. In the context of sports prediction using machine learning, the primary assumptions regarding the nature of the target variable are typically embedded in the choice of loss function. To the best of our knowledge, existing approaches in the literature predominantly treat predicted probabilities as quantitative variables, see [1].

Recent contributions in measurement theory have questioned the adequacy of the classical taxonomy of scales when applied to empirical data. In particular, Ref. [5] highlights the limitations of the traditional nominal–ordinal–interval–ratio classification, arguing that many real-world variables do not fully satisfy the transformation invariance properties required for a strict assignment to one category. This perspective is especially relevant in cases such as the one considered in this paper, where probabilities exhibit an ordinal-like behavior in certain regions of their domain, while still lacking key properties typically associated with ordinal scales, such as invariance under arbitrary monotonic relabeling. This suggests that the level of measurement may not be an intrinsic property of the variable itself, but rather an emergent characteristic that depends on the data-generating process and the region of analysis.

Complementary evidence is provided by recent applied studies such as [6], which identify non-quantitative behavior in variables traditionally treated as quantitative, such as velocity measurements in sports performance. Other studies, such as [7], investigate the effect of mathematical properties of the input data, such as entropy, on the performance of sports prediction models. These findings reinforce the notion that metric assumptions may break down under specific conditions, leading to distortions in statistical analysis and model training. Taken together, this body of work supports the need for a more flexible and context-dependent interpretation of measurement scales, aligning with the approach adopted in this paper, where the nature of probabilities is empirically assessed rather than assumed a priori.

In this paper, two central research questions are addressed:

- Do all predicted probability intervals behave consistently with a metric quantitative interpretation, or do certain regions exhibit structural properties that are more closely aligned with ordinal frameworks?

- Is there a convergence between direct empirical measurements of ordinality and the behavior of models trained under the proposed ordinal inductive bias?

In the context of sports betting, probabilities are not directly observed but are inferred through price formation, which reflects the aggregation and exchange of information among market participants. Rare events, in particular, tend to exhibit lower resolution in their estimated probabilities. This is due to both the higher risk associated with mispricing such outcomes and to the relative scarcity of observations available for sampling. As a result, the probabilities assigned to low-likelihood events may be less precise and exhibit characteristics more consistent with ordinal rather than fully quantitative measurements.

In this work, we investigate this phenomenon using a public dataset [8] containing multiple betting odds for National Basketball Association (NBA) games collected from online bookmakers. From these odds, we derive the corresponding implicit probabilities and focus our analysis on the behavior of low-probability events relative to the rest of the distribution. Our prior hypothesis is that these probabilities, especially in the tails, exhibit characteristics closer to ordinal variables than to fully quantitative ones.

To assess this hypothesis, we directly analyze the structural properties of the dataset. In particular, we examine the loss of linearity and the preservation of monotonicity in the region hypothesized to exhibit ordinal behavior, and contrast these properties with those observed in the remainder of the dataset. This approach allows us to characterize deviations from standard probabilistic calibration and to identify whether the lower tail of the probability distribution follows a distinct regime consistent with ordinal structure.

The hypothesis is further validated through an independent modeling perspective by introducing an ordinal inductive bias via the Consistent Rank Logits (CORAL) [9] loss within a Gated Recurrent Unit (GRU) [10] architecture. This setup provides an additional layer of empirical evidence by comparing the training dynamics under ordinal and quantitative assumptions. As a result, it not only supports the hypothesis of ordinal behavior in rare NBA event predictions, but also demonstrates the effectiveness of inductive bias as a practical tool for detecting the underlying level of measurement of a target variable.

Finally, the nature of a target variable is often characterized by the invariance properties that it preserves. For instance, ordinal variables retain their empirical meaning under order-preserving monotonic transformations, since such transformations preserve the ranking among observations without requiring equal distances between adjacent values [5,11]. In our setting, however, these invariance properties cannot be directly tested, because the observed target values do not provide access to alternative admissible transformations of the same underlying measurement process. At the same time, many loss functions are constructed to be compatible with specific data types: squared losses target metric structure, while ordinal losses target ordered categories. In this paper, we therefore propose a methodology based on training the same machine learning architecture with different loss functions, each designed for a different target structure, in order to infer the most plausible nature of a dataset from its optimization behavior. This approach may also be relevant beyond the present application; for example, natural-language product reviews mapped to Likert-type ratings could be analyzed similarly, by comparing whether their predictive structure is better captured by quantitative losses or by ordinal losses.

2. Related Work

2.1. Sport-Betting Probabilities

Sport-betting price formation follow the same rule than financial derivatives, Galekwa et al. [12] proves that they can treat as another economical asset. Given a portfolio of bets managed by the accepting party (e.g., a bookmaker), the liability can be defined as the maximum potential payout that the operator is committed to under all possible outcomes.

Formally, for a given event i with a set of mutually exclusive outcomes $j \in \mathcal{J}_i$, the liability is computed as:

$$L_i = \max_{j \in \mathcal{J}_i} \left(\sum_{k \in \mathcal{B}_{i,j}} s_k \cdot o_k \right) \quad (2)$$

where s_k denotes the stake of bet k , o_k the associated odds, and $\mathcal{B}_{i,j}$ the set of bets that result in a payout if outcome j occurs. The total portfolio liability is then obtained by aggregating across independent events:

$$L = \sum_i L_i \quad (3)$$

Table 1 provides a simplified example with two matches, each with two possible selections (Home and Away). Decimal odds are complemented with their corresponding ML representation and implied probabilities.

Table 1. Example of liability computation across two matches. For Match 1, the potential payouts are 18 (Home) and 33 (Away), yielding $L_1 = 33$. For Match 2, the payouts are 21 (Home) and 0 (Away), yielding $L_2 = 21$. Hence, the total liability is $L = 33 + 21 = 54$.

Match	Selection	Stake (s)	Odds (o)	ML	Prob.
1	Home	10	1.8	−125	0.556
1	Away	15	2.2	+120	0.455
2	Home	20	1.05	−2000	0.952
2	Away	0	20.0	+1900	0.050

This quantity is particularly sensitive to unbalanced events, where one outcome is associated with significantly higher odds. In such cases, even a small relative reduction in the largest odds can produce a substantial decrease in total liability, highlighting the non-linear relationship between pricing adjustments and risk exposure. This sensitivity, in turn, induces a structural bias toward more aggressive over-rounds in selections with large Money Line values.

2.2. Machine Learning in Sports Forecasting

Machine learning techniques are now widely established in the field of sports prediction, as documented in the comprehensive review of [1]. Recent advances indicate that auto-regressive models constitute the current state of the art in this domain [13], primarily due to their ability to capture temporal dependencies and sequential patterns inherent in sports data.

Auto-regressive models are designed to predict future outcomes based on previously observed values in a time-ordered sequence. Formally, given a sequence $\{x_t\}$, these models estimate the conditional probability $P(x_t \mid x_{t-1}, x_{t-2}, \dots)$, thereby leveraging historical information to inform future predictions. In the context of sports forecasting, such models can incorporate past match results, team performance trajectories, and evolving contextual variables. Among the most widely used auto-regressive architectures are recurrent neural networks (RNNs), including Long Short-Term Memory (LSTM) networks [14] and GRUs [10], which address issues of vanishing gradients and enable the modeling of long-range dependencies.

In this work, we adopt a GRU-based architecture. This choice is not motivated by the intention to benchmark against state-of-the-art models, but rather by the need for a parsimonious and interpretable auto-regressive framework capable of capturing the behaviors of interest. The relative simplicity of GRUs, compared to more complex architectures, allows for a clearer analysis of how different assumptions regarding the

nature of the target variable influence the training dynamics and predictive performance of the model.

2.3. Loss Functions

The choice of loss function in supervised learning is intrinsically linked to the assumed nature of the target variable. In many sports prediction settings, the target—often expressed as a probability—is treated as a quantitative variable. Under this assumption, regression-based loss functions such as the Mean Absolute Error (MAE) are commonly employed; for instance, MAE is discussed by [15] as an absolute-error criterion for regression problems. However, MAE is also relevant in ordinal settings, particularly when ordered labels are embedded into a numerical scale; for example, Ref. [16] explicitly considers MAE as an evaluation metric for ordinal regression. In this sense, MAE is a versatile loss function: it can be used to penalize metric deviations when the target is interpreted quantitatively, while also preserving sensitivity to the ordering of labels when the target is treated as ordinal. The MAE is defined as:

$$\ell_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (4)$$

where y_i denotes the observed value and $\hat{y}_i$ the model prediction. When the target lies on a continuous scale with meaningful distances, this formulation linearly penalizes deviations and is therefore well suited for genuinely quantitative variables. In ordinal contexts, MAE can still be informative because larger ordinal discrepancies receive larger penalties than smaller ones. Nevertheless, its use implicitly imposes a numerical spacing between adjacent categories, which may not always be justified.

Ordinal regression has a long tradition in statistics and should not be reduced to recent neural approaches. Classical models explicitly exploit the ordered nature of the response variable without assuming a fully cardinal structure. In particular, cumulative link models define ordinal responses through cumulative probabilities and link functions, with the proportional odds model and related cumulative logit formulations being among the most widely used approaches [17,18]. Closely related latent-variable formulations, such as ordinal probit and ordinal logit models, interpret the observed ordinal label as a discretization of an underlying continuous variable through a set of thresholds [19]. These models provide a statistical foundation for treating ordinal outcomes as ordered categories rather than nominal classes or metric quantities.

In contrast, when the target variable is ordinal—i.e., it represents ordered categories without necessarily well-defined distances between them—alternative loss formulations may be more appropriate. One such approach is CORAL (COntinuous RANk Logits) [9], which models ordinal regression as a series of binary classification problems over cumulative thresholds. Within the broader ordinal-regression framework, CORAL can be understood as a deep-learning extension that preserves ordinal structure by decomposing the prediction problem into rank-consistent binary classification subtasks. The CORAL loss is defined as:

$$\ell_{\text{CORAL}} = - \sum_{i=1}^N \sum_{k=1}^{K-1} [\mathbb{I}(y_i > k) \log \sigma(v_{i,k}) + (1 - \mathbb{I}(y_i > k)) \log(1 - \sigma(v_{i,k}))], \quad (5)$$

where K is the number of ordered categories, $\sigma(\cdot)$ denotes the sigmoid function, $v_{i,k}$ are the logits associated with threshold k , and $\mathbb{I}(\cdot)$ is the indicator function. This formulation enforces a consistent ordering across categories and captures the ordinal structure of the target. CORAL has been successfully applied in domains where the response variable is

inherently ordered, such as Likert-scale survey responses (e.g., levels of agreement), credit risk ratings, medical severity scores, or quality assessment tasks.

The key innovation of this paper lies in applying an ordinal framework, specifically CORAL loss, to a range of probabilities that has traditionally been treated as purely quantitative. By reconsidering the measurement assumptions in this regime, we provide evidence that certain probability ranges—particularly those associated with extreme events—may exhibit ordinal characteristics, and that leveraging this structure can lead to improved training behavior and predictive performance.

2.4. Statistical Learning Theory

2.4.1. Bayes Optimal Predictor

In statistical learning theory, the Bayes optimal predictor is the decision rule that minimizes the expected risk with respect to the true, although usually unknown, data-generating distribution [15,20]. Given an input variable X , an output variable Y , a loss function $\ell(\cdot, \cdot)$, and a predictor f , the expected risk is defined as

$$\mathcal{R}(f) = \mathbb{E}_{(X,Y)}[\ell(f(X), Y)]. \quad (6)$$

The Bayes optimal predictor is therefore

$$f^* = \arg \min_f \mathcal{R}(f). \quad (7)$$

For example, under squared error loss, the Bayes optimal predictor is the conditional expectation $\mathbb{E}[Y|X]$, while under classification losses it is related to the conditional class probabilities $P(Y|X)$ [15,21].

In this work, we use this perspective not only as a theoretical optimum, but also as a way to reason about the functional structure that best fits a given type of data. If a model trained with a particular loss function systematically reduces the empirical loss in a specific region of the data, we interpret this reduction as evidence that the corresponding inductive assumptions are better aligned with the underlying Bayes-optimal functional form. In particular, if an ordinal loss improves performance in the low-odds region, this suggests that the Bayes-optimal representation of that region may be closer to an ordinal structure than to a purely quantitative one.

2.4.2. Inductive Bias

Inductive bias refers to the set of assumptions that a learning algorithm uses to generalize beyond the observed training data [22,23]. These assumptions may be introduced through the hypothesis class, the architecture, the regularization strategy, the optimization procedure, or the loss function. For instance, convolutional neural networks introduce locality and translation-related assumptions for image data, while recurrent architectures impose temporal structure on sequential data [24,25]. Importantly, inductive bias differs from data bias.

We deliberately introduce an ordinal inductive bias through an ordinal loss function, specifically CORAL [9], in order to test whether different regions of betting odds respond differently to ordinal assumptions. The goal is not only to improve predictive performance, but to use the model's sensitivity to this controlled bias as a diagnostic tool. If the ordinal bias consistently benefits one subset of the data, such as low NBA odds, more than another, this provides empirical evidence that the former may be better modeled as ordinal rather than cardinal.

3. Exploratory Data Analysis

3.1. Dataset Description

The dataset employed in this study [8] comprises a total of 43,815 forecasts, each associated with an implied probability derived from market odds. The distribution of these probabilities is highly imbalanced, with a markedly low representation of extreme values. In particular, only 182 observations correspond to probabilities below 0.10, and within this subset, just 18 observations fall below 0.05 (Table 2). This pronounced sparsity in the lower tail of the distribution has important implications for both statistical analysis and model training, as it limits the amount of information available to accurately characterize and learn from rare events.

Table 2. Distribution of observations in the lower tail of the probability spectrum. Out of a total of 43,815 forecasts, only a small fraction corresponds to low-probability events, highlighting the strong class imbalance inherent in the dataset.

Subset	Number of Observations
Total dataset	43,815
$\subset P \in [0.05, 0.1]$	164
$\subset P < 0.05$	18

On the other hand, liability management processes such as those described in Equation (2) provide a direct mechanism explaining the observed ordinal behavior. In particular, outcomes associated with very high odds (i.e., low probabilities) entail disproportionately large liability, which incentivizes operators to apply more aggressive over-rounds in this region. This adjustment distorts the metric structure of probabilities, compressing differences in magnitude while preserving their relative ordering. As a consequence, probabilities in the lower tail no longer behave as fully quantitative variables, but instead exhibit characteristics more consistent with an ordinal scale. The algorithms used to apply the margin are not uniquely determined, since probabilities can be normalized through several methods, each depending on specific parameters. This indeterminacy allows each online betting operator to modulate its risk exposure, and such procedures are therefore typically kept private.

We use the public dataset of NBA betting odds compiled from multiple online bookmakers [8]. As discussed in the Introduction, the available samples are inherently limited, particularly in the tails of the probability distribution. This constraint is especially relevant when analyzing rare events, where the number of observations is comparatively small.

From the quoted odds, we compute the corresponding implied probabilities and group them into bins of width 0.05. For each probability segment, we estimate the empirical frequency of victory and compare it with the associated implied probability. This aggregation allows us to study calibration effects and deviations from expected quantitative behavior across different probability ranges. See Figure 1.

3.2. Calculating Implied Probabilities

ML prices embed a margin (over-round) that results in implied probabilities summing to more than one. The magnitude and allocation of this margin are typically proprietary to each online betting operator. To mitigate this effect and obtain normalized probabilities, we rescale the implied probabilities by dividing each of them by the total sum of probabilities across all mutually exclusive outcomes. In the present setting, where each basketball match is modeled as a binary outcome (home vs. away win), the normalization is performed over the two corresponding selections.

The conversion from ML odds to implied probability ($\tilde{P}$) is given by:

$$\tilde{P}_i = \begin{cases} \frac{-ML_i}{-ML_i + 100}, & \text{if } ML_i < 0, \\ \frac{100}{ML_i + 100}, & \text{if } ML_i > 0, \end{cases} \quad (8)$$

where ML_i denotes the quoted ML value for outcome i . By normalizing the probability, we obtain the margin-free probability, also referred to as the fair probability (P), given by the following equation:

$$P_i = \frac{\tilde{P}_i}{\sum_{j \in \mathcal{J}} \tilde{P}_j}, \quad (9)$$

where $\mathcal{J}$ is the set of all possible outcomes for a given event (in this case, $|\mathcal{J}| = 2$). This transformation ensures that the adjusted probabilities sum to one and can be consistently interpreted within a probabilistic framework.

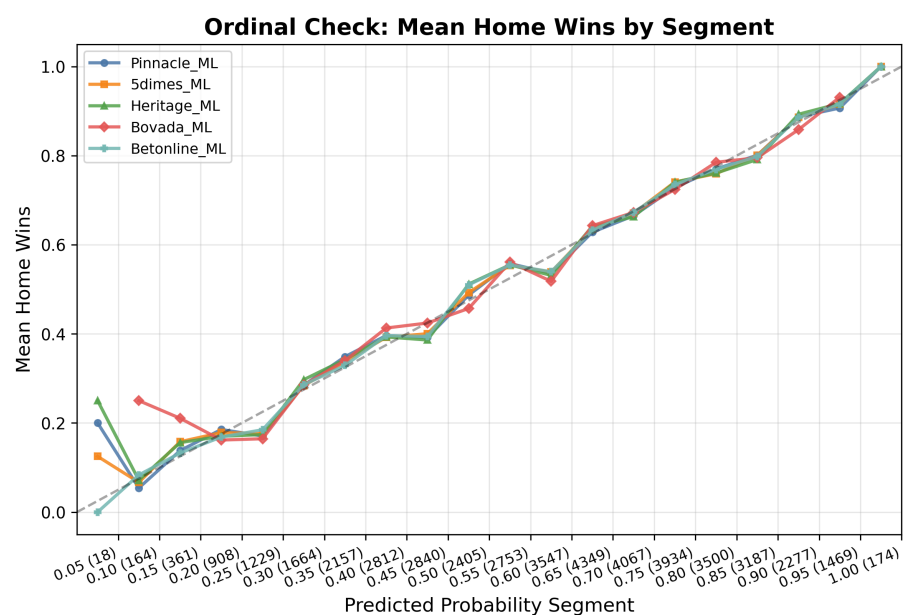


Figure 1. The figure shows the probabilities of the dataset [8] aggregated into 0.05 bins. Each color represents a different online betting place. The x-axis represents the upper limit of each bin, with the number of samples in each bin shown in parentheses.

4. Methodology: Linearity and Monotonicity Analysis

To assess the extent to which probabilities retain their quantitative nature, we analyze their linearity and monotonicity. Linearity implies perfect calibration, where the empirical frequency y scales one-to-one with the predicted fair probability p , denoted by x . This is assessed via the identity line $y = x$. Monotonicity, on the other hand, is a relaxed ordinal constraint requiring only that $y_i \leq y_j$ whenever $x_i < x_j$, regardless of the absolute distance between them. While a quantitative dataset $\mathcal{D}_Q$ requires linearity to minimize information loss, an ordinal dataset $\mathcal{D}_O$ may exhibit non-linear behavior while strictly preserving its monotonic rank order.

4.1. Linearity Analysis

We study linearity by comparing empirical outcome frequencies with the identity function (the bisector). Specifically, we compute the residuals between observed win frequencies and the corresponding implied probabilities. Deviations from the bisector are interpreted as a loss of linearity, indicating that the measurement may no longer

preserve meaningful proportional relationships and may instead behave more like an ordinal variable.

To analyze the linearity of each subset, we discretize the predictions into m bins. For each bin $j \in \{1, \dots, m\}$, we define the mean predicted probability x_j , the empirical success frequency y_j , and the sample size w_j as a weighting factor. We employ a Weighted Least Squares (WLS) approach to account for the heteroscedasticity induced by the different bin number of samples. We compute the weighted means of both variables:

$$\bar{x}_w = \frac{\sum_{j=1}^m w_j x_j}{\sum_{j=1}^m w_j}, \quad \bar{y}_w = \frac{\sum_{j=1}^m w_j y_j}{\sum_{j=1}^m w_j}, \quad (10)$$

The centered sums of squares and products are:

$$S_{xx,w} = \sum w_j (x_j - \bar{x}_w)^2, \quad S_{xy,w} = \sum w_j (x_j - \bar{x}_w)(y_j - \bar{y}_w) \quad (11)$$

The weighted slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ are calculated as:

$$\hat{\beta}_1 = \frac{S_{xy,w}}{S_{xx,w}}, \quad \hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \quad (12)$$

Statistical Significance of Linearity

We compute the statistical significance from the residual variance ($\hat{\sigma}^2$):

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^m w_j (y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)^2}{m - 2}, \quad (13)$$

Under the WLS assumptions, the standard error of the estimated slope is given by:

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx,w}}}, \quad (14)$$

To determine the significance of the metric deviation from a perfect 1:1 slope, we formulate the null hypothesis $H_0 : \beta_1 = 1$. The test statistic follows a Student's t -distribution with $m - 2$ degrees of freedom:

$$t_{\text{slope}} = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)}, \quad (15)$$

$$\rho_{\text{slope}} = 2 \cdot P(T_{m-2} \geq |t_{\text{slope}}|), \quad (16)$$

To minimize the effects of sample scarcity in the low-probability region, we apply the following bootstrap procedure, after splitting the observations into two subsets according to whether their probability is below or above 0.1. For each subset, we sample N bins. For each bin, a center c_i and a width o_i are randomly generated. The center is sampled from a truncated normal distribution centered at the midpoint of the corresponding probability range, whereas the bin width is sampled from a uniform distribution.

For the low-probability subset, defined by $p < 0.1$, we set the probability range to $[0, 0.1]$. For each bootstrap bin $i = \{1, \dots, N\}$, the bin center (c_i) is sampled from a uniform distribution between 0 and 0.1, and the bin width (o_i) is sampled from a uniform distribution between 0.002 and 0.01:

$$c_i \sim \mathcal{U}(0, 0.1), \quad (17)$$

$$o_i \sim \mathcal{U}(0.002, 0.01), \quad (18)$$

In our case, 20 samples are used for the segment between 0 and 0.1 and 200 samples for the segment between 0.1 and 1.

4.2. Monotonicity Analysis

To evaluate whether the relationship between predicted probabilities and empirical frequencies preserves an ordinal structure, we analyze the monotonicity of the empirical win-rate curve.

To assess order preservation independently of metric distance, we construct a monotonic projection of the empirical frequencies. Let y^{iso} denote the non-decreasing sequence obtained by applying the weighted Pool-Adjacent-Violators (PAV) algorithm to the match outcomes y_j , sorted by the fair probability x , as described by Dimitriadis [26]. For each group of observations, we compute the empirical frequency $z_{w,i}$ as

$$z_{w,i} = \frac{\sum_{j \in G_i} y_j}{w_i}, \quad (19)$$

where G_i denotes the set of observations in group i and w_i is the corresponding group size.

The isotonic winning frequency y^{iso} is then obtained as

$$y_i^{\text{iso}} = \begin{cases} z_{w,i}, & \text{if } y_i^{\text{iso}} \leq y_{i+1}^{\text{iso}}, \\ \frac{w_i z_{w,i} + w_{i+1} z_{w,i+1}}{w_i + w_{i+1}}, & \text{if } y_i^{\text{iso}} > y_{i+1}^{\text{iso}}. \end{cases} \quad (20)$$

In our case, $w_i = 50$. Appendix B Table A1 reports the computation of the first six groups as an illustrative example. We denote by w_i^{iso} the number of samples in each isotonic group, obtained by summing the original group sizes w_i that are merged into the same isotonic block.

We compute the isotonic Root Mean Square Error, denoted by RMSE^{iso} , as

$$\text{RMSE}^{\text{iso}} = \sqrt{\frac{\sum_i w_i^{\text{iso}} (z_i - y_i^{\text{iso}})^2}{\sum_i w_i^{\text{iso}}}}. \quad (21)$$

The monotonicity index, denoted by M , is defined as

$$M = 1 - \text{RMSE}^{\text{iso}}. \quad (22)$$

Statistical Significance

To assess the statistical significance of the results, we compute confidence intervals (CIs) through a bootstrap procedure with $B = 200$ iterations.

For each iteration b , we simulate a null replicate assuming a binomial distribution around the isotonic estimate:

$$z_{b,i} \sim \text{Binomial}(w_i, y_i^{\text{iso}}), \quad (23)$$

where w_i denotes the group size and y_i^{iso} is the isotonic win frequency estimated from the observed data. The simulated frequency is then obtained as

$$\hat{z}_{b,i} = \frac{z_{b,i}}{w_i}. \quad (24)$$

An isotonic refitting step is then performed according to Equation (20), yielding the bootstrap isotonic sequence $y_{b,i}^{\text{iso}}$. For each iteration, we compute $\text{RMSE}_b^{\text{iso}}$ following Equation (21). Repeating this procedure produces a sequence of bootstrap values

$\{\text{RMSE}_b^{\text{iso}}\}_{b=1}^B$. The bootstrap replicates are then sorted, and the 95% CI is obtained from the empirical 2.5% and 97.5% percentiles, using linear interpolation when the percentile falls between two ordered replicates. In this case, with $B = 200$, the lower 2.5% percentile lies at position 4.975, while the upper 97.5% percentile lies at position 194.025 in the ordered bootstrap sequence.

To compute the upper one-sided p -value, we count the proportion of bootstrap replicates whose isotonic error is greater than or equal to the observed value. Let $u(\cdot)$ denote the Heaviside step function. The monotone-null p -value is defined as

$$\rho_{\text{mono}} = \frac{1 + \sum_{b=1}^B u(\text{RMSE}_b^{\text{iso}} - \text{RMSE}^{\text{iso}})}{B + 1}. \quad (25)$$

5. Methodology: Train Under Inducted Bias Analysis

The methodology of training under inductive bias is based on identifying dataset characteristics through the changes induced by modifications of the inductive bias in a machine learning process. The method presented below consists of inducing an inductive bias through a loss function and training a model to compare risk differences between different biases for similar datasets.

5.1. Theoretical Framework

To establish the conditions under which the inductive bias affects the empirical risk of two similar datasets, three elements are required. First, the Bregman divergence, which restricts the scope of applicability to loss functions satisfying the Proper Scoring Rule (PSR) property. Second, a proposed definition of distance, used to determine the proximity between distributions and between different measurement natures. Third, the Empirical Risk Minimization (ERM) condition, which characterizes the training process.

Starting from the divergence ($D(P|Q)$) for a PSR proposed by Dawid [27], where $S(P|Q)$ is a PSR, P is the population distribution, and Q is the sample distribution, we have:

$$S(P|Q) = S(P|P) + D(P|Q). \quad (26)$$

In Equation (26), under a given hypothesis h , we have that $R(h) = S(P|Q)$, $S(P|P)$ is the Bayes optimum R^* , and, as Dawid shows, without loss of generality for our case, $D(P|Q)$ is the Bregman divergence [28], which we denote by $D_S(P|Q)$:

$$R(h) = R^* + D_S(P|Q). \quad (27)$$

Under the EMR condition, see Vapnik [20], we select the hypothesis that minimizes the training-sample error:

$$h_S^{\text{ERM}} = \underset{h \in \mathcal{H}}{\text{argmin}} \hat{R}_S(h). \quad (28)$$

Inserting this into Equation (27) and considering $P \neq P(h)$, we have:

$$R(h_S^{\text{ERM}}) = R^* + \min_{\mathcal{H}} D_S(P|Q(h)). \quad (29)$$

Applying this to the training of two datasets (A, B) and the set of hypotheses h_1 induced by predictor that generates Q :

$$R_A = R_A^* + \min_{h \in h_1} D_S(P_A|Q(h)), \quad R_B = R_B^* + \min_{h \in h_1} D_S(P_B|Q(h)). \quad (30)$$

The differences between the two compared training processes must be bounded by Definition 1. This condition cannot be measured directly, due to the very definition of the Bayes optimum, but it can be bounded using tools provided by Probably Approximately Correct (PAC) learning, such as Maurer's theorem [29].

Definition 1. We will consider two datasets to be similar if the risk of the Bayes optimum is similar for both datasets: $S(P_A|P_A) \approx S(P_B|P_B)$. For simplicity, we denote this as $R_A^* \approx R_B^*$.

As will be shown throughout the proofs in this section, the similarity introduced in Definition 1 can be applied when the difference between R_A^* and R_B^* is smaller than the difference between the minimum Bregman divergences of the predictor with respect to each of the corresponding Bayes optima. Taking Definition 1 and Equation (30) into account, a relationship is established between the divergence across distributions and the empirical risk:

Lemma 1. Given two similar datasets according to Definition 1, with Bayes optima P_A and P_B , if $R(P_A) < R(P_B)$, then $\min_{h \in h_1} D_S(P_A|Q(h)) < \min_{h \in h_1} D_S(P_B|Q(h))$.

Proof. Substituting (30) into $R(P_A) < R(P_B)$ and imposing $R_A^* \approx R_B^*$, we obtain:

$$\min_{h \in h_1} D_S(P_A|Q(h)) < \min_{h \in h_1} D_S(P_B|Q(h)). \quad (31)$$

□

Definition 2. We will characterize a measurement level by a manifold $\mathcal{M}$ composed of all possible distributions that constitute that measurement level. If a distribution P is considered to belong to the measurement level represented by $\mathcal{M}$, then $P \in \mathcal{M}$.

For example, for the measurement level of three categorical classes, $\mathcal{M}$ is the simplex with two degrees of freedom. To represent the ordinal measurement level, we must add a constraint to the simplex so that the classes are not interchangeable, for example: $\{p_1, p_1 + p_2, 1 - p_1 - p_2\}$.

To provide a notion of similarity between a distribution and two measurement levels, we propose the distance described in Definition 3.

Definition 3. Given a distribution A , two manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$ and $\mu(\mathcal{M})$ an arbitrary measure in a measure space with $\mathcal{M}_1$ and $\mathcal{M}_2$ in its measurable set. The distance between them is given by:

$$d(A|\mathcal{M}) = \mu(\mathcal{M}) + \min_{S \in \mathcal{M}} D_{KL}(A|S). \quad (32)$$

By construction, for a distribution that may belong to two levels of measurement, the distance is smaller with respect to the level of measurement associated with the coarser measure in the common measurable space. In other words: Given $A \in \mathcal{M}_1$ and $A \in \mathcal{M}_2$ and $\mathcal{M}_1 \subset \mathcal{M}_2$, this definition allows:

$$\begin{aligned} d(A|\mathcal{M}_1) &< d(A|\mathcal{M}_2), \\ d(A|\mathcal{M}) &> 0 \quad \forall \mathcal{M}, \\ d(A|A) &< d(A|\mathcal{M}) \quad \forall \mathcal{M}, \end{aligned} \quad (33)$$

Conversely, when the levels of measurement induce similar measures on the common measurable space, the corresponding distance is defined as the information projection of the distribution onto the manifold of distributions associated with each measurement level.

Finally, the relationship between the empirical risks and the proposed distance is given by Theorem 1. The same proof could be extended to a more general setting by considering any Bregman divergence, rather than only the Kullback–Leibler divergence, in the definition of the distance. For clarity, we retain the specific case in order to interpret the proposed distance in terms of information projection.

Theorem 1. *Given two subsets satisfying Lemma 1 and a PSR that allows $D_S = D_{KL}$, then $d(P_A|\mathcal{M}_1) < d(P_B|\mathcal{M}_1)$.*

Proof. Starting from Lemma 1 and $D_S = D_{KL}$, we have:

$$\min_{h \in h_1} D_{KL}(P_A|Q(h)) < \min_{h \in h_1} D_{KL}(P_B|Q(h)). \quad (34)$$

If the solution space of the predictor is equal to the set of distributions contained in the measurement level represented by the measurable set $\mathcal{M}_1$, that is, if the introduced inductive bias h_1 coincides with the measurable set that represents it, then we have:

$$\min_{h \in h_1} D_{KL}(P|Q(h)) = \min_{S \in \mathcal{M}_1} D_{KL}(P|S), \quad (35)$$

Therefore,

$$\min_{S \in \mathcal{M}_1} D_{KL}(P_A|S) < \min_{S \in \mathcal{M}_1} D_{KL}(P_B|S) \quad (36)$$

Adding $\mu(\mathcal{M}_1)$ to both sides:

$$\mu(\mathcal{M}_1) + \min_{S \in \mathcal{M}_1} D_{KL}(P_A|S) < \mu(\mathcal{M}_1) + \min_{S \in \mathcal{M}_1} D_{KL}(P_B|S). \quad (37)$$

Substituting using Definition 3, we obtain the conclusion:

$$d(P_A|\mathcal{M}_1) < d(P_B|\mathcal{M}_1). \quad (38)$$

□

From Definition 3, we can interpret that $d(P_A|\mathcal{M}_1) < d(P_B|\mathcal{M}_1)$ implies that the Bayes optimum of the dataset with lower loss is closer to the manifold representing ordinal numbers than the Bayes optimum of the dataset with higher loss.

This technique introduces a connection between measurement theory and the measurement level in cases of similar datasets through the study of the risk generated under different inductive biases. This criterion introduces a relative proximity between the hypothesis and the Bayes optimum; as expected, there could always exist an inductive bias that introduces a hypothesis closer to the Bayes optimum.

Application to the Proposed Experiment

This paper addresses the case of probabilities assigned to basketball games studied under the methodology of the previous Section 5.1. The similarity between populations, required by Definition 1, is produced through three techniques:

- Standardization of the variables using mean and standard deviation in both subsets.
- Stratification of the larger population, taking only elements with probabilities between 0.1 and 0.2 from the larger population.

- Subsampling of the larger population, taking samples of similar size to the smaller population (182 samples).

On the other hand, it is assumed that, due to the similar nature of the samples, both will have Bayes optima with similar risks.

The ERM condition is achieved by assuming that the training has been effective and has found the sample minimum.

In our case, the inductive bias is introduced through the CORAL loss function, described in Section 2.3. Since this loss function is based on a linear combination of binary classifiers, we can interpret the Bregman divergence D_S as the Kullback–Leibler divergence D_{KL} .

5.2. Hypotheses

In the results, we will evaluate whether ordinal risk (R_O) is lower than quantitative risk (R_Q). To this end, we define d :

$$d = R_Q - R_O. \quad (39)$$

To address the primary research question—“Do all probability ranges behave consistently with a quantitative interpretation, or do certain regions exhibit properties more aligned with ordinal structures?”—we formulate the following statistical hypothesis:

$$\begin{cases} H_0 : d \leq 0 & \text{(The ordinal region does not exhibit a better fit for ordinal inductive bias)} \\ H_1 : d > 0 & \text{(The ordinal region exhibits a significantly lower relative loss)} \end{cases} \quad (40)$$

5.3. Experiment Setting

The experimental protocol is designed to test whether the subset of observations with target values below 0.1 exhibits an ordinal structure. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the full dataset. The dataset is first partitioned into two disjoint subsets based on the original target scale:

$$\mathcal{D}_O = \{(x_i, y_i) \in \mathcal{D} : y_i < 0.1\}, \quad \mathcal{D}_Q = \{(x_i, y_i) \in \mathcal{D} : y_i \geq 0.1\}, \quad (41)$$

where $\mathcal{D}_O$ is the candidate ordinal region and $\mathcal{D}_Q$ is the candidate quantitative region. Target values are standardized independently within each split:

$$\tilde{y}_{i,S} = \frac{y_i - \mu_S}{\sigma_S}, \quad (x_i, y_i) \in \mathcal{D}_S, \quad S \in \{O, Q\}. \quad (42)$$

This independent standardization accounts for the inherent scale disparity between $\mathcal{D}_O$ and $\mathcal{D}_Q$, ensuring that loss comparisons reflect the compatibility between the loss function and the data structure rather than numerical magnitude.

To introduce an ordinal inductive bias, we train a GRU on both datasets using a CORAL loss function. For each match, the input is structured as a temporal sequence comprising the 10 preceding matches ordered chronologically for both opponents and their respective rivals. Since in a GRU the information from each cell flows only to the subsequent cell, therefore arranging the matches chronologically in each sequence prevents a possible target leak since during inference each cell can only use the information from later cells and these contain previous matches. The final recurrent cell outputs the event probability given the two contenders. Categorical features representing the teams are formalized using One-Hot Encoding (OHE).

Prior to the primary CORAL training phase, a pre-training step of 100 iterations is performed using MAE without aiming for complete convergence. This pre-training stage

is introduced to address data scarcity in one of the subsets; by leveraging this approach, we avoid resampling the validation set, thereby preserving its independent and identically distributed (i.i.d.) properties. MAE is selected as the objective function for this phase due to its robust adaptability to both ordinal and quantitative target representations.

Furthermore, we utilize a modified version of the CORAL framework to allow the same GRU architecture to adapt seamlessly to both MAE and CORAL loss objectives. This modification enables the GRU to output a scalar prediction instead of a standard vector of ordinal class logits. To achieve this, the predicted scalar is multiplied by a four-element one-dimensional tensor, facilitating a mapping that projects the scalar output into three distinct ordinal classes.

The structural and training hyperparameters of the GRU are parameterized as follows. We implement a Bayesian GRU architecture consisting of 2 stacked recurrent layers with a latent dimension of 16 hidden units per layer, incorporating a dropout rate of 0.1 to mitigate overfitting. For temporal sequencing, the model processes input sequences with a window length of 10 steps and a unit stride. Training and validation are executed symmetrically with a batch size of 16. Finally, temporal alignment is governed by the Date column, aimed at predicting the target variable calibrated through the fair odds probability method for all providers.

We repeat this experiment across 20 independent training runs, using 20% of the data for validation sampled with substitution. Each training and validation batch consists of chronologically ordered sequences sampled randomly. The loss obtained on the ordinal dataset is significantly lower than that obtained on the quantitative dataset.

5.4. Significance

As described in the previous section, positive values of d_i indicate that the ordinal subset achieves a lower validation loss than the quantitative subset. Given a sample $d_i = \{d_1, \dots, d_n\}$, we will consider deviations of d_i to be positive if its median ($\tilde{\mu}_d$) is greater than zero. Therefore, the hypothesis that the ordinal subset has a lower risk can be formulated as:

$$H_0 : \tilde{\mu}_d \leq 0, \quad H_1 : \tilde{\mu}_d > 0, \quad (43)$$

Due to the discrete and limited number of trials, we propose a non-parametric method, the Wilcoxon test [30], to test this hypothesis. This test generates two statistics (W^+ , W^-), representing the number of pairs that result in a d value greater than and less than zero, respectively. The p -value of this test is given by the frequency of all combinations that confirm the hypothesis ($c(W^+)$) out of all possible combinations (2^n), where $n = 20$ in our case.

$$\rho = P(W \geq W^+) = \frac{c(W^+)}{2^n}. \quad (44)$$

6. Results

In this section, we present two independent and complementary sets of results. The first is based on descriptive statistical analyses of dataset linearity and monotonicity, while the second examines the outcomes obtained under alternative assumptions about the level of measurement employed during model training, specifically contrasting ordinal inductive bias.

6.1. Structural Analysis

As described in the methodological section, the dataset is partitioned into two regions according to the assumed level of measurement: an ordinal regime corresponding to probabilities below 0.1, and a quantitative regime for the remaining observations. Within

each subset, we analyze the preservation of monotonicity and the loss of linearity with respect to the identity mapping. The results summarized in Table 3 provide empirical support for this distinction, showing that the low-probability subset exhibits a strong deviation from linearity while maintaining a high degree of monotonic consistency, in contrast with the near-linear behavior observed in the higher-probability regime.

Table 3. This table shows the strong deviation from linearity in the low-probability subset alongside a high monotonicity index, contrasted with near-linear behavior and slightly lower monotonicity in the higher-probability subset. Uncertainty is represented through the 95% CI of the RMSE^{iso} and β .

Subset	Linearity			Monotonicity		
	β	CI Low	CI High	M	RMSE CI Low	RMSE CI High
$p < 0.10$	−1.00	−2.75	0.75	0.97	0	0.05
$p \geq 0.10$	0.99	0.96	1.02	0.94	0.058	0.064

6.1.1. Linearity Analysis

By aggregating the data across all providers, as described in the methodology section, Figure 2 shows that this pattern persists globally for probabilities below 0.10 with the bisector representing a perfectly calibrated predictive model. Notably, the deviation is substantially larger in the aforementioned low-probability range compared to the rest of the distribution. Taken together, these results provide further evidence of a distinct linear behavioral pattern in this range of odds.

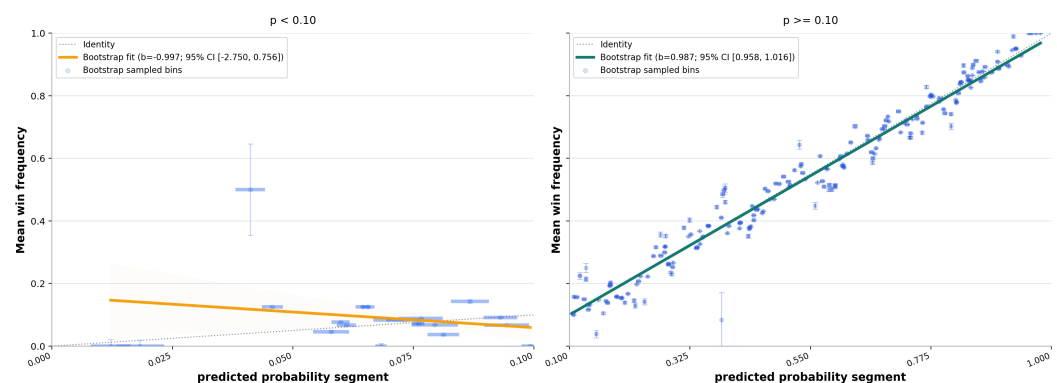


Figure 2. Representation of the bootstrap analysis of linearity. Using the method in Section 4.1, the left panel shows the 20 bins generated for probabilities less than 0.1 and 200 for probabilities greater than 0.1. The left panel corresponds to the subset with probabilities below 0.1, whereas the right panel corresponds to the subset with probabilities greater than or equal to 0.1.

$$H_0 : \beta = 1, \quad H_1 : \beta \neq 1. \quad (45)$$

Given the estimated coefficient $\hat{\beta} = -0.66$ and its standard error $SE(\hat{\beta}) = 0.001$, the test statistic is

$$t = \frac{\hat{\beta} - 1}{SE(\hat{\beta})} = \frac{-1.00 - 1}{0.84} = -2.38. \quad (46)$$

The corresponding two-sided p -value generated from the 20 points is given by Equation (16):

$$\rho = 2(1 - F_t(|-2.38|; 18)) = 0.028, \quad (47)$$

where $F_t(\cdot; \nu)$ denotes the cumulative distribution function of a Student's t distribution with ν degrees of freedom.

6.1.2. Monotonic Analysis

For validating significance of monotonicity index $= 1$ we test $H_0 : M = 1$ vs. $H_1 : M < 1$, where M denotes the monotonicity index. Given CI values from Table 3, for both probability ranges, above and below 0.1, the monotonicity index lies within the CI associated with monotonic behavior.

Both groups show similar monotonicity index M close to 1 with compatible CI s, consistent with the monotonicity conservation hypothesis in both parts.

6.1.3. Joint Hypothesis: Deviation from Linearity with Preserved Monotonicity

We evaluate the joint hypothesis that the low-probability subset ($p < 0.10$) exhibits:

- a slope significantly different from 1;
- a monotonicity index close to 1.

From the previous results, the slope estimate $\hat{\beta} = -1.0$ clearly departs from the identity line ($\beta = 1$), indicating a strong deviation from linear calibration. At the same time, the monotonicity indices of the groups of probabilities greater and less than 0.1, $M = 0.97$ and $M = 0.94$ respectively, remain close to unity, suggesting that the ordering of probabilities is largely preserved.

This combination implies that, although the mapping between predicted probabilities and observed frequencies is not linear in the low-probability regime, it retains a strong ordinal structure. In other words, predictions remain correctly ranked despite systematic distortions in scale.

6.2. Train Under Inducted Bias Analysis

For the second set of results, 20 independent pre-training and training runs were conducted in accordance with the experimental protocol described in Section 5.3. The validation losses for both subsets, $\mathcal{D}_O$ and $\mathcal{D}_Q$, are illustrated in Figure 3 and summarized in Table 4.

Substituting values from Figure 3, in significance criteria of Section 5.4:

$$W^+ = 166, \quad W^- = 44, \quad (48)$$

Finally, given the result from Equation (48), the one-sided p -value is:

$$\rho = P(W \geq 166) = \frac{11264}{2^{20}} = 0.01. \quad (49)$$

Since $\rho < 0.05$, the result provides statistically significant evidence that $\tilde{\mu}_d > 0$. Therefore, under the definition $d = R_Q - R_O$, the ordinal subset achieves a significantly lower minimum validation loss than the quantitative subset.

Table 4. Comparative validation loss analysis for the ordinal ($\mathcal{D}_O$) and quantitative ($\mathcal{D}_Q$) subsets. All runs were initialized following MAE pre-training and optimized using the CORAL loss function. Results are reported across 20 independent iterations to assess statistical significance.

Run	Ordinal Val Loss (R_O)	Quantitative Val Loss (R_Q)	Difference (d)
Repeat 01	1.5734	1.9840	0.4106
Repeat 02	1.7787	2.1163	0.3376
Repeat 03	1.9700	1.5656	−0.4044
Repeat 04	1.0314	1.8714	0.8401
Repeat 05	1.8694	1.7546	−0.1148
Repeat 06	1.6500	1.8649	0.2149

Table 4. Cont.

Run	Ordinal Val Loss (R_O)	Quantitative Val Loss (R_Q)	Difference (d)
Repeat 07	1.6338	1.9598	0.3260
Repeat 08	1.2460	1.2013	−0.0446
Repeat 09	1.8524	1.6100	−0.2424
Repeat 10	1.7842	1.8016	0.0175
Repeat 11	1.4180	2.1151	0.6971
Repeat 12	1.5520	2.0824	0.5304
Repeat 13	1.2773	2.0598	0.7825
Repeat 14	1.8150	2.1296	0.3146
Repeat 15	2.1384	1.5709	−0.5675
Repeat 16	1.7539	1.6362	−0.1177
Repeat 17	1.5053	1.6940	0.1887
Repeat 18	1.2072	1.7074	0.5002
Repeat 19	0.9708	1.9034	0.9327
Repeat 20	1.8636	2.1110	0.2474
Mean (μ)	1.5945	1.8370	0.2424
Standard Deviation (σ)	0.3178	0.2465	0.4126

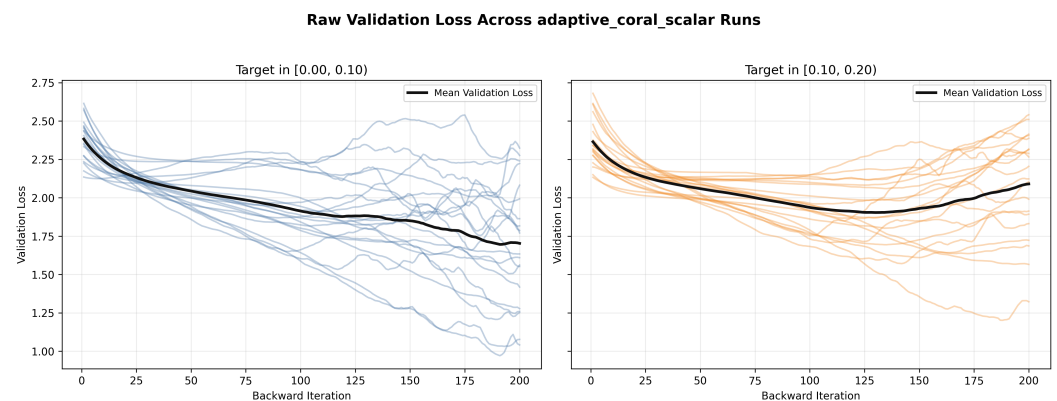


Figure 3. Validation loss during the uptraining stage across 20 independent experiments, during 200 backward iterations. The left panel (blue) shows the low-probability ($y < 0.1$) split $\mathcal{D}_O$, while the right panel (orange) shows the higher-probability ($y \in [0.1, 0.2)$) split $\mathcal{D}_Q$.

7. Discussion

Probabilities have traditionally been regarded as quantitative measures within the classification proposed by Stevens [2]. The limitations of this classification have already been discussed by other authors, such as Philippi [5]. This article studies this effect in probabilities below 0.1 for NBA games using two different methods: a direct measurement of the dataset properties and a new method based on comparing the behavior of the empirical risk during training when introducing an inductive bias associated with a given measurement level.

The direct measurements compare linearity and monotonicity between the subset of probabilities lower than 0.1 and the remaining probabilities. These measurements indicate that linearity is lost in the subset below 0.1, see Figure 2. However, monotonicity is preserved across the entire dataset, see Figure 4. These characteristics may indicate that, although this subset does not satisfy all the properties of ordinal sets, such as symmetry under monotonic relabeling, the subset below 0.1 shares more properties with ordinal structures than with quantitative ones.

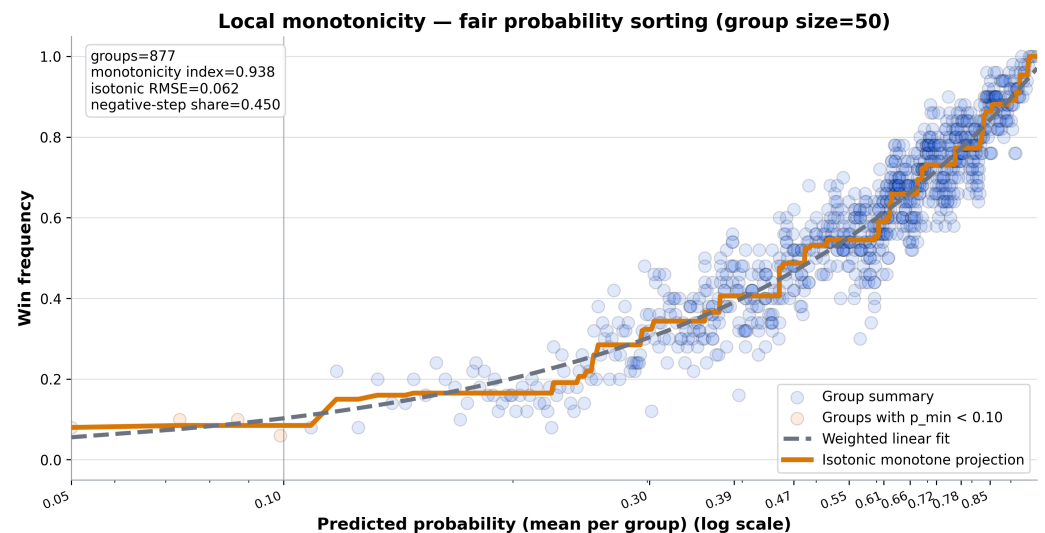


Figure 4. Dataset grouped into bins of 50 matches, ordered from lowest to highest probability. Orange line represent isotonic monotone projection, which determine monotonicity index. X-axis is represented in logarithmic scale in order to compare easier the rarer probability part.

On the other hand, we propose an independent methodology that validates the results obtained through direct measurement. This methodology consists of training a GRU with a CORAL loss function on the subset assumed to be ordinal and on the most similar subset assumed to be quantitative, in this case the probabilities between 0.1 and 0.2. This loss function restricts the hypothesis space of the predictive model to hypotheses compatible with ordinality. A significant increase in the minimum validation risk is measured in the dataset for which linearity is assumed; therefore, according to the framework presented in Section 5.1, this can be considered evidence of the ordinal character of the first subset relative to the second.

8. Conclusions and Future Work

This study has examined whether low probabilities associated with NBA game predictions can still be treated as strictly quantitative magnitudes or whether, instead, they exhibit behavior closer to that of an ordinal structure. The results support the latter interpretation for the subset of probabilities below 0.1. The direct analysis shows that, within this range, the linear relationship between the assigned probabilities and the observed outcomes deteriorates, while monotonicity is preserved. This indicates that low probabilities do not completely lose their ordering capacity, but they do exhibit a relevant loss of the properties associated with a quantitative interpretation. In this sense, the subset of probabilities below 0.1 appears to preserve an ordinal structure better than a linear metric structure. Independently, the analysis based on training with an ordinal inductive bias reinforces this conclusion. By comparing the behavior of a GRU trained with the CORAL loss function on the subset of probabilities below 0.1 and on the closest subset considered quantitative, between 0.1 and 0.2, a significant difference is observed in the minimum validation risk. Under the proposed theoretical framework, this result indicates that the first subset is better adapted to the ordinal bias introduced by the loss function, providing additional evidence of its relative ordinal character. 3 The main contribution of this article is twofold. First, it provides empirical evidence that low probabilities in sports prediction markets may not behave as fully quantitative measures, despite being commonly treated as such. Second, it proposes a complementary method for studying the measurement level of a dataset through the analysis of empirical risk under a controlled inductive bias. This approach makes it possible to connect measurement theory with the behavior observed during the

training of predictive models. Nevertheless, the study has several limitations. The limited availability of public data and the naturally low frequency of unlikely events restrict the effective size of the subset under analysis. In addition, the direct analysis has focused on fundamental properties, such as linearity and monotonicity, without exhausting all possible characterizations of ordinality. Finally, the inductive-bias-based method has been evaluated in a specific case, using a GRU architecture and the CORAL loss function, so its limits of applicability require further investigation.

As future work, a natural first extension is to apply the direct measurements to larger datasets, both within the NBA and in other sports with sufficiently dense prediction markets. This would make it possible to assess more robustly whether the loss of linearity in low probabilities is specific to the dataset studied here or a more general property of extreme probabilities in sports contexts.

A second line of work will consist of expanding the properties evaluated through direct measurement. In particular, it would be relevant to study partial monotonic relabelings that make it possible to assess the extent to which the subset of low probabilities preserves invariances characteristic of ordinal structures. This extension would directly address the current limitation of having evaluated only the most basic ordinal properties.

Finally, it will be necessary to extend the proposed theoretical framework in order to characterize more precisely the risk differences induced by different inductive biases. In particular, a relevant future direction will be to study whether it is possible to obtain a formal bound for the difference between the risks associated with different loss functions or restrictions of the hypothesis space. This extension would make it possible to move from a comparative interpretation of empirical risk toward a more robust criterion for evaluating when a given inductive bias is more compatible with the measurement structure of a dataset.

Author Contributions: Conceptualization, A.J.S.G., Z.S. and R.A.C.; Methodology, A.J.S.G., Z.S. and R.A.C.; Software, A.J.S.G.; Validation, Z.S. and R.A.C.; Formal analysis, A.J.S.G.; Investigation, A.J.S.G., Z.S. and R.A.C.; Resources, A.J.S.G., Z.S. and R.A.C.; Data curation, A.J.S.G.; Writing—original draft, A.J.S.G. and Z.S.; Writing—review and editing, A.J.S.G., Z.S. and R.A.C.; Visualization, A.J.S.G.; Supervision, Z.S. and R.A.C.; Project administration, Z.S. and R.A.C.; Funding acquisition, Z.S. and R.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Project PID2024-155289NB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Data Availability Statement: Data available at <https://www.kaggle.com/datasets/erichqiu/nba-odds-and-scores> (accessed on 18 March 2026).

Acknowledgments: This work was supported by Project PID2024-155289NB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. The authors also gratefully acknowledge the Universidad Complutense de Madrid (research group D3MOS: Data-Driven Decision Making for Organizations and Society).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GRU	Gated Recurrent Unit
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
CORAL	Consistent Rank Logits
i.i.d.	Independent and Identically Distributed
PAV	Pool-Adjacent-Violators

WLS	Weighted Least Squares
CI	Confidence Interval
OHE	One-Hot Encoding
ERM	Empirical Risk Minimization
PSR	Proper Scoring Rule
SE	Standard Error
PAC	Probably Approximately Correct
NBA	National Basketball Association

Appendix A. Extension of Theoretical Framework

In this paper, we study the effect of applying a given inductive bias to two similar populations. In Theorem A1, we instead analyze the application of two different inductive biases to the same population. We show that proximity to a level of measurement can only be guaranteed when the more specific level of measurement is also the one that yields the lower risk.

Theorem A1. *Let h_1 and h_2 be two hypotheses associated, respectively, with two levels of measurement over the same population P . If $\mu(\mathcal{M}_1) \leq \mu(\mathcal{M}_2)$ and $R_1(P) < R_2(P)$, then*

$$d(P \mid \mathcal{M}_1) < d(P \mid \mathcal{M}_2). \quad (\text{A1})$$

Proof. The proof follows analogously to the proof of Theorem 1.

$$\mu(\mathcal{M}_1) + \min_{S \in \mathcal{M}_1} D_{KL}(P|S) < \mu(\mathcal{M}_1) + \min_{S \in \mathcal{M}_2} D_{KL}(P|S), \quad (\text{A2})$$

Since $\mu(\mathcal{M}_1) \leq \mu(\mathcal{M}_2)$. Therefore,

$$\mu(\mathcal{M}_1) + \min_{S \in \mathcal{M}_2} D_{KL}(P|S) < \mu(\mathcal{M}_2) + \min_{S \in \mathcal{M}_2} D_{KL}(P|S) = d(P \mid \mathcal{M}_2), \quad (\text{A3})$$

Consequently,

$$d(P \mid \mathcal{M}_1) < d(P \mid \mathcal{M}_2). \quad (\text{A4})$$

□

Appendix B. Isotonic Winning Frequency

We provide an illustrative example of how the isotonic winning frequency (y^{iso}) is computed. In the first six groups, the first monotonicity violation occurs between groups 3 and 4. The resulting merged block then produces a violation with group 2, and the block obtained from these three groups subsequently violates monotonicity with group 5.

Table A1. Computation of z_i and y_i^{iso} for the first six groups.

Group	w_i	$\sum_j y_j$	Computation of z_i	Block $B(i)$	y_i^{iso}
1	50	4	$z_1 = 4/50 = 0.08$	1	$z_1 = 0.08$
2	50	5	$z_2 = 5/50 = 0.1$	2–5	$17/200 = 0.085$
3	50	5	$z_3 = 5/50 = 0.1$	2–5	$17/200 = 0.085$
4	50	3	$z_4 = 3/50 = 0.06$	2–5	$17/200 = 0.085$
5	50	4	$z_5 = 4/50 = 0.08$	2–5	$17/200 = 0.085$
6	50	11	$z_6 = 11/50 = 0.22$	6–7	$15/100 = 0.15$

Appendix C. Sensitivity Analysis for Smaller Subset

Appendix C.1. Linearity

The interval $[0, 0.10]$ is re-binned several times to test how stable the linearity diagnostics are when using 10, 5, 4, 3, and 2 effective bins.

Table A2. Linearity analysis for different binning strategies in the probability range $[0, 0.1]$.

Bins	Effective Width	Sample Size	Slope	Weighted R^2	p Slope ($H_0 = 1$)
20	0.005	182	−0.77	0.03	0.19
10	0.01	182	−0.77	0.06	0.2
5	0.02	182	−0.47	0.03	0.42
4	0.025	182	−0.48	0.05	0.41
3	0.033	182	−0.08	0.03	0.26
2	0.05	182	−0.77	0.002	0.14

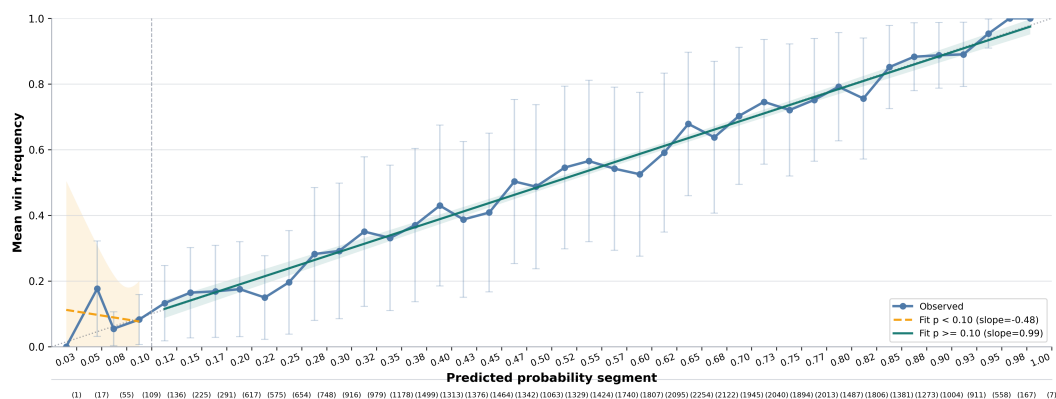


Figure A1. The figure illustrates fix-width bin analysis, with a width value set to 0.025. The nonlinearity observed in the first range (probabilities between 0 and 0.01), contrasted with the linear behavior in the remaining ranges. The dashed line represents the ideal model. The number of samples in each bin is shown in parentheses below the x-axis.

Appendix C.2. Monotonicity

Given that there are 182 samples with probability below 0.1, we consider group sizes that produce 19, 10, 4, and 3 groups, respectively, each containing at least one observation with probability below 0.1.

Table A3. Monotonicity analysis across different group sizes in the probability range $[0, 0.1]$.

Group Size	Groups	M	$RMSE^{iso}$	$\rho(H_0)$	Bootstrap Mean Index	$CI_{95\%}$
10	4382	0.86	0.14	0.7	0.844	0.84–0.85
20	2191	0.90	0.096	0.9	0.885	0.88–0.89
50	877	0.94	0.062	0.2	0.923	0.92–0.93
80	548	0.95	0.048	0.2	0.938	0.93–0.94

Appendix D. GRU Hyper-Parameters

The proposed architecture is a recurrent neural network designed to forecast match outcomes by learning from sequential, game-level historical records of the competing teams. The network is implemented using the PyTorch framework (version $\geq 2.0.0$) and is structured into three primary components: input representation, sequential temporal encoding, and the prediction head.

Appendix D.1. Input Representation

The network processes historical context over a sequence of length T (with a default of $T = 10$). At each time step $t \in \{1, 2, \dots, T\}$, the input encompasses two sets of features:

- **Token Features ($\mathbf{x}_{tok}^{(t)}$):** Four categorical variables representing the team identifiers: the home team, the home team's historical opponent, the away team, and the away team's historical opponent. Each identifier is constrained to the interval $[0, N_{teams}]$ and mapped to a one-hot encoded vector of dimension $N_{classes} = N_{teams} + 1$, where the zero index is reserved for padding or missing states.
- **Mask Features ($\mathbf{x}_{mask}^{(t)}$):** Two binary flags indicating the absence of historical opponent records for either the home or the away team.

The one-hot encoded token representations are flattened and concatenated with the mask features. This procedure yields a dense input vector $\mathbf{x}^{(t)}$ with a total dimensionality of $d_{in} = 4 \times N_{classes} + 2$.

Appendix D.2. Recurrent Sequential Encoding

To capture the temporal dynamics and historical performance trajectories of the teams, the sequence of input vectors $\mathbf{X} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(T)})$ is processed by a GRU network. The GRU formulation mitigates the vanishing gradient problem while efficiently routing historical information:

$$\mathbf{h}^{(t)} = \text{GRU}(\mathbf{x}^{(t)}, \mathbf{h}^{(t-1)}) \quad (\text{A5})$$

The GRU module consists of $L = 2$ stacked recurrent layers with a hidden state dimensionality of $d_{hidden} = 64$. To promote generalization and mitigate overfitting, dropout regularization with a probability of $p = 0.1$ is applied between the recurrent layers.

Appendix D.3. Prediction Head

The sequence of hidden states generated by the final recurrent layer, $\mathbf{H} = (\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \dots, \mathbf{h}^{(T)})$, is processed by a shared dense prediction head to yield outcome estimations. The head comprises a fully connected (linear) layer followed by a non-linear activation, applied independently at each time step:

$$\hat{\mathbf{y}}^{(t)} = \sigma(\mathbf{W}\mathbf{h}^{(t)} + \mathbf{b}) \quad (\text{A6})$$

where $\mathbf{W} \in \mathbb{R}^{2 \times 64}$ is the learnable weight matrix, $\mathbf{b} \in \mathbb{R}^2$ is the bias vector, and $\sigma(\cdot)$ denotes the element-wise Sigmoid activation function. The resulting sequence $\hat{\mathbf{y}}^{(t)} \in (0, 1)^2$ provides the estimated probabilities for the home and away team outcomes. While predictions are generated at each time step, the estimation at the final step T is typically extracted as the definitive probability forecast for the target match.

References

1. Galekwa, R.M.; Tshimula, J.M.; Tajeuna, E.G.; Kyamakya, K. A Systematic Review of Machine Learning in Sports Betting: Techniques, Challenges, and Future Directions. *arXiv* **2024**, arXiv:2410.21484. <https://arxiv.org/pdf/2410.21484>.
2. Stevens, S.S. On the Theory of Scales of Measurement. *Science* **1946**, *103*, 677–680.
3. Hand, D.J. *Measurement Theory and Practice: The World Through Quantification*; Arnold Publishers: London, UK, 2004.
4. Kuzon, W.M.; Urbanchek, M.G.; McCabe, S. The Seven Deadly Sins of Statistical Analysis. *Ann. Plast. Surg.* **1996**, *37*, 265–272.
5. Philippi, C.L. On Measurement Scales: Neither Ordinal nor Interval? *Philos. Sci.* **2021**, *88*, 929–939.
6. Bailey, C. Statistical Considerations When Measuring Absolute Reliability and Variability of Vector Data in Sport Performance. *Int. J. Strength Cond.* **2023**, *3*, 1–8 <https://doi.org/10.47206/ijsc.v3i1.163>.
7. Neuman, Y.; Israeli, N.; Vilenchik, D.; Cohen, Y. The Adaptive Behavior of a Soccer Team: An Entropy-Based Analysis. *Entropy* **2018**, *20*, 758. <https://doi.org/10.3390/e20100758>.

8. Qiu, E. NBA Odds and Scores Dataset. Kaggle. 2020. Available online: <https://www.kaggle.com/datasets/erichqiu/nba-odds-and-scores> (accessed on 18 March 2026).
9. Cao, W.; Mirjalili, V.; Raschka, S. Rank Consistent Ordinal Regression for Neural Networks with Application to Age Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 540–541.
10. Cho, K.; van Merriënboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *arXiv* **2014**, arXiv:1406.1078. <https://arxiv.org/abs/1406.1078>.
11. Hand, D.J. Statistics and the Theory of Measurement. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **1996**, *159*, 445–473.
12. Galekwa, R.M.; Tshimula, J.M.; Tajeuna, E.G.; Kyamakya, K. Towards Sports Betting as a Financial Asset: An Investigative Analysis of Risk, Investment Potential, and Future Perspectives. *IEEE Access* **2026**, *14*, 38057–38070.
13. Sun, Y.; Chu, H. The Outcome Prediction Method of Football Matches by the Quantum Neural Network Based on Deep Learning. *Sci. Rep.* **2025**, *15*, 19875.
14. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
15. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
16. Tsuchiya, T.; Charoenphakdee, N.; Sato, I.; Sugiyama, M. Semisupervised Ordinal Regression Based on Empirical Risk Minimization. *Neural Comput.* **2021**, *33*, 3361–3412. https://doi.org/10.1162/neco_a_01445.
17. McCullagh, P. Regression Models for Ordinal Data. *J. R. Stat. Soc. Ser. B (Methodol.)* **1980**, *42*, 109–142. <https://doi.org/10.1111/j.2517-6161.1980.tb01109.x>.
18. Agresti, A. *Analysis of Ordinal Categorical Data*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2010. <https://doi.org/10.1002/9780470594001>.
19. McKelvey, R.D.; Zavoina, W. A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *J. Math. Sociol.* **1975**, *4*, 103–120. <https://doi.org/10.1080/0022250X.1975.9989847>.
20. Vapnik, V.N. *Statistical Learning Theory*; Wiley: New York, NY, USA, 1998.
21. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
22. Mitchell, T.M. *The Need for Biases in Learning Generalizations*; Technical Report CBM-TR-117; Rutgers University: New Brunswick, NJ, USA, 1980.
23. Mitchell, T.M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
25. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
26. Dimitriadis, T.; Gneiting, T.; Jordan, A.I. Stable Reliability Diagrams for Probabilistic Classifiers. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016191118. <https://doi.org/10.1073/pnas.2016191118>.
27. Dawid, A.P. The Geometry of Proper Scoring Rules. *Ann. Inst. Stat. Math.* **2007**, *59*, 77–93.
28. Bregman, L.M. The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217.
29. Maurer, A. A note on the PAC Bayesian theorem. *arXiv* **2004**, arXiv:cs/0411099.
30. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biom. Bull.* **1945**, *1*, 80–83.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.