International Journal of Approximate Reasoning 192 (2026) 109636

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Predicting Lockean from gradational accuracy ⋆

Igor Douven [a,*], Nikolaus Kriegeskorte [b]

[a] SND, CNRS, Sorbonne University, France
[b] Zuckerman Mind Brain Behavior Institute, Columbia University, USA

ARTICLE INFO

Keywords:
Accuracy
Calibration
Forecasting
Lockean thesis
Mediation
Scoring rules
Sharpness

ABSTRACT

The debate about which scoring rule best measures the accuracy of our credences has largely been conducted on an *a priori* basis. We pursue an empirical approach, asking which rule best predicts a practical, decision-relevant criterion: Lockean accuracy, the ability to make correct categorical judgments based on a threshold of belief. Analyzing a large dataset of probability judgments, we compare the most widely used scoring rules (Brier, logarithmic, spherical, absolute error, and power rules) and find that, among them, there is no single best one. Instead, the optimal choice is context dependent: the Spherical score is the best predictor for lower belief thresholds, while the Power^3 rule is best at higher thresholds. In particular, the widely used Brier and log scores are rarely optimal for this task. A mediation analysis reveals that while much of a rule's success is explained by its ability to reward calibration and sharpness, the Spherical and Brier rules retain significant predictive power independently of these standard virtues.

1. Introduction

According to an idea popular among formal epistemologists, the degree to which our credences (or degrees of belief) are accurate is measured via some scoring rule. There is an ongoing debate about *which* scoring rule we are to use for this purpose. The Brier score is the most widely favored, but the log score also has its proponents, as do various other scoring rules (to be detailed in due course, together with the Brier and log scores).

In this paper, we want to combine the aforementioned idea with another popular idea, to wit, that full belief does not require certainty. The best known formal implementation of this idea is the so-called Lockean thesis (LT), according to which there is a threshold $\theta \in [.5, 1)$ such that we believe a proposition φ if our credence for it exceeds θ and disbelieve φ (i.e., believe it to be false) if our credence for it is below $1 - \theta$; if we neither believe nor disbelieve φ, we are said to be *agnostic* about φ. It is known that LT creates problems (most notably the *lottery* and *preface paradoxes*) if at the same time we want to maintain that full belief is closed under conjunction, so that we fully believe $\varphi \wedge \psi$ as soon as we fully believe both φ and ψ. In response, authors have either rejected this closure principle (following, e.g., [2]), or adopted some restricted version of it (e.g., [3]), or proposed some variant of LT that makes high credence (i.e., credence above θ) only *defeasibly* sufficient for full belief (e.g., by also requiring that the credence be *stably* high, in the manner of [4]). For the study we report in this paper, it makes no difference whether we stick to LT (with or without closure) or adopt some variant of it. We thus simply assume LT in the following.

With LT on board, we can distinguish a notion of accuracy in addition to that of gradational accuracy (i.e., the accuracy of our credences), namely, what we shall call *Lockean accuracy*. For a given value of θ, Lockean accuracy measures the proportion of beliefs

⋆ Supplementary Materials, containing the Julia code ([1]) used for the analyses we report, can be downloaded from this repository: https://osf.io/43nxg/overview?view_only=abd820fcc2a7452f81618d62dc91bc9e.

* Corresponding author.
E-mail address: igor.douven@sorbonne-universite.fr (I. Douven).

https://doi.org/10.1016/j.ijar.2026.109636
Received 11 November 2025; Received in revised form 16 January 2026; Accepted 22 January 2026
Available online 30 January 2026

that are correct, that is, the sum of the things you believe to be true (given your credences and θ) that *are* true and the things you believe to be false (again given your credences and θ) that *are* false, divided by the number of things you are not agnostic about.

The notion of Lockean accuracy is theoretically interesting as a plausible quality measure for the set of things we are not agnostic about. But it is also practically important, given the ubiquity of threshold-based decisions. To give just some examples, legal systems use standards like "beyond reasonable doubt," medical guidelines specify treatment thresholds, and institutional policies often require categorical judgments even when underlying evidence is probabilistic. The same applies to artificial intelligence systems, where probabilistic classifiers must often convert continuous confidence scores into categorical predictions, a process that likewise requires selecting an appropriate threshold and raises analogous questions about which scoring rule best tracks real-world classification performance. Naturally, Bayesian decision theory can make recommendations—whether to deem a defendant guilty, whether to initiate a treatment, and so on—without resorting to any thresholds. But the theory requires as input, next to probabilities, also utilities, and, as ([5], p. 271) notes, in many contexts the latter are unavailable, unstable, or insufficiently intersubjective. In such contexts, we tend to rely on a threshold-based heuristic (e.g., we believe someone to be guilty iff we believe to a high enough degree that the person is guilty), and how well that heuristic serves us will depend on how Lockean-accurate the resulting categorical beliefs are.

This motivates our primary research question: *Which scoring rule best predicts Lockean accuracy?* While much of the debate about scoring rules has proceeded on an *a priori* basis, focusing on which mathematical properties such rules ought and ought not to have, our approach is empirical.[1] Using a large dataset of probability assignments from hundreds of participants evaluating more than 1000 general knowledge claims with known ground truth, we investigate the aforementioned question for a range of possible threshold values θ. As a further research question, we want to know *why* certain scoring rules perform better than others in distinguishing highly Lockean-accurate forecasters from ones that do less well. Specifically, we want to know to what extent the performance of scoring rules is mediated by forecaster properties that in the literature have been tied to scoring (most notably, calibration and sharpness; see [9], and below for definitions).

Our main finding is that there is no universal answer to our central question: there is no single best score; the choice depends on θ. We also find that the most popular scoring rules are surprisingly disappointing when applied to the task of predicting Lockean accuracy.

2. Theoretical background

This section defines the notions of accuracy that will be central to our empirical study, as well as the key forecaster properties that we will use to explain our main results.

2.1. Accuracy

Gradational accuracy. We distinguish between gradational and Lockean accuracy. The former is defined in terms of some scoring rule. To state the idea of a scoring rule generically, let $\mathcal{H} = \{H_1, \ldots, H_n\}$ be a hypothesis partition (i.e., a set of self-consistent, mutually exclusive, and jointly exhaustive hypotheses). Furthermore, let $\mathbf{p} = \langle p_1, \ldots, p_n \rangle$ be a probability distribution on $\mathcal{H}$ such that p_j is the probability assigned to hypothesis H_j. Then a scoring rule is a function that assigns a loss $\mathcal{S}(\mathbf{p}, i)$ to $\mathbf{p}$ when H_i is true.

Because our dataset consists of general knowledge claims that are either true or false, we state scoring rules here only for the binary case where each hypothesis partition consists of a hypothesis (according to which a given outcome occurred or will occur) and its negation (according to which the outcome did not or will not occur). Let p denote the probability assigned to the true outcome, and let $y \in \{0, 1\}$ indicate whether the outcome occurred/will occur ($y = 1$) or not ($y = 0$). Then the scoring rules to be considered in this paper are the following:

$$\begin{aligned}
\text{(Brier)} \quad & \mathsf{B}(p, y) = (p - y)^2, \\
\text{(Logarithmic)} \quad & \mathsf{L}(p, y) = -y \ln(p) - (1 - y) \ln(1 - p), \\
\text{(AbsError)} \quad & \mathsf{A}(p, y) = |p - y|, \\
\text{(Spherical)} \quad & \mathsf{S}(p, y) = -\frac{yp + (1 - y)(1 - p)}{\sqrt{p^2 + (1 - p)^2}}, \\
\text{(Power}^\alpha\text{)} \quad & \mathsf{P}^\alpha(p, y) = |p - y|^\alpha \quad (\alpha = 3, 4).
\end{aligned}$$

Note that the absolute error (AbsError) and Brier scores are, in effect, the Power$^\alpha$ rule with $\alpha = 1$ and, respectively, $\alpha = 2$.

Because scoring rules assign *losses*, lower scores are *better*. However, this is a conventional choice: we could take the negative of any given score and think of it as a reward. While we adhere to standard practice, we will often speak of a rule's ability to measure gradational accuracy, where it is understood that lower scores (losses) correspond to higher accuracy.

[1] Our paper thereby aligns with a growing body of work studying probabilistic reasoning and accuracy through experimental methods. For instance, in the domain of confirmation theory, Tentori et al. [6] empirically compared alternative measures of evidential support, examining which best captures human confirmation judgments. Similarly, Zhao et al. [7] test empirically the Bayesian orthodoxy about the relationship between conditional probability and updating. And in diagnostic reasoning, Stilgenbauer and Baratgin [8] empirically compared inference strategies for probability estimation, showing that defeasible modus ponens leads to more accurate estimates than affirming the consequent.

As mentioned in the introduction, the debate about which of these rules we ought to use for measuring gradational accuracy is ongoing. Nevertheless, there are already two clear favorites, to wit, the Brier rule, defended by, among others, Brier [10], Rosenkrantz [11], Joyce [12], Selten [13], and Leitgeb and Pettigrew [14], and the log rule, defended by, among others, Bernardo [15], Bernardo and Smith [16], and Bickel [17].[2] [3]

As also mentioned, the debate about which scoring rule is best has focused on *a priori* arguments. For example, a main argument against Power$^{\alpha}$ rules for any power other than 2 is that they are *improper*. Given a scoring rule $\mathcal{S}$ and a probability distribution $\mathbf{p}$ on $\mathcal{H}$, the $\mathbf{p}$-*expectation* of the $\mathcal{S}$-loss for a second probability distribution $\mathbf{p}^*$ also on $\mathcal{H}$ equals $\mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] := \sum_{i=1}^{n} p_i \mathcal{S}_i(\mathbf{p}^*)$. A scoring rule $\mathcal{S}$ is said to be *proper* iff for all $\mathbf{p}$, $\operatorname{argmin}_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] = \mathbf{p}$, and *strictly proper* iff each of these minima is unique. Of the above rules, all but the mentioned power rules are strictly proper. We nonetheless include those power rules (the absolute error rule and the power rules for powers 3 and 4) because our approach is *empirical*. Propriety mostly matters *ex ante* (in the terminology of [23]), specifically, for elicitation: if you want forecasters to post their true credences, you should score them (and let them know that you score them) using some strictly proper scoring rule. But our use of scores is exclusively *ex post*: we treat them as statistics for predicting Lockean accuracy from credences *already on record*. And there is no *a priori* argument to the effect that an improper rule cannot best predict Lockean accuracy, in our dataset or indeed in any dataset.

Lockean accuracy. We want to connect gradational accuracy with Lockean accuracy, which is defined in terms of the threshold θ postulated by the Lockean thesis (LT). Recall that, where Pr is a person's credence function, LT is the thesis that the person believes proposition φ if $\Pr(\varphi) > \theta$, disbelieves φ if $\Pr(\varphi) < 1 - \theta$, and otherwise is agnostic vis-à-vis φ. Thus, where $\Phi = \{\varphi_1, \ldots, \varphi_N\}$ is a given set of propositions of interest and $\Pr(\varphi_i) = p_i$ for all $\varphi_i \in \Phi$,

$$\Phi_\theta := \{\varphi_i : p_i > \theta \text{ or } p_i < 1 - \theta\}$$

is the set of propositions about which the person is *not* agnostic (i.e., which he/she either believes or disbelieves). Then, at the given value of θ, the person is Lockean-accurate to a degree equal to

$$\mathrm{ACC}_\theta = \frac{\left|\{i : p_i > \theta \wedge \varphi_i \text{ is true}\}\right| + \left|\{i : p_i < 1 - \theta \wedge \varphi_i \text{ is false}\}\right|}{|\Phi_\theta|}.$$

While everyone agrees that θ should be at least .5 (lest a proposition and its negation can both be believable), there is no further agreement on the value of θ. In the literature, a number of specific values, or at least admissible ranges, have been mentioned; for example, Shear and Fitelson [24] argue that $\theta \geqslant .618$; Kyburg [2], Kaplan [25], and Moser and Tlumak [26] all mention .9 as a plausible value; and Foley [27] even proposes .99.[4] Currently more dominant is a contextualist position vis-à-vis the value of θ, according to which the threshold should depend on what is at stake in a given context ([4]; see also [30] and [31]). We are sympathetic to this kind of contextualism and will therefore in our study survey a range of values for θ, where we will be particularly interested in how the predictiveness of scores depends on θ.

2.2. Forecaster properties

We want to know which rule or rules best predict Lockean accuracy for different θ's and therefore will look at how scores and accuracies correlate. But correlations do not tell us *why* one rule does better than another, whether, for example, the better rule rewards forecaster properties that we actually value or whether it is exploiting something else, which perhaps we do not value as much but which, by some strange coincidence, gives it the edge in our dataset. In answer to the *why* question, we analyze the performance of scoring rules in terms of fundamental forecaster properties. The mediation analysis presented in Section 3.2 will test this explanatory approach in detail.

The idea of evaluating forecasts through such properties has a long history in the statistical literature. Murphy [32] showed that the Brier score, for instance, can be mathematically decomposed into terms measuring a forecaster's *calibration* (or reliability) and their *sharpness* (or resolution). This conceptual framework, which emphasizes the goal of maximizing sharpness subject to high calibration, has become central to the modern theory of forecast evaluation ([19]). Our analysis builds on this tradition by using the aforementioned properties, along with the philosophically motivated property of *coverage*, as explanatory mediators. We define each in turn.

Calibration. Calibration measures a forecaster's fundamental, underlying trustworthiness, in that it asks whether of the propositions to which the forecaster assigns a probability of x, approximately $x \times 100$ percent are true. Formally, we say that a person with probability function Pr is *well-calibrated* iff $\mathbb{E}_{\Pr}[y \mid p] = p$ (e.g., it rains on 90 percent of the days for which a weather forecaster predicts rain with a .9 probability). Empirically, we measure *mis*calibration. Given forecasts p_i and outcomes y_i for $i = 1, \ldots, N$, choose bins $\{B_k\}_{k=1}^{K}$ partitioning $[0, 1]$, where for each bin,

$$n_k := |\{i : p_i \in B_k\}|, \quad \bar{p}_k := \frac{1}{n_k} \sum_{i : p_i \in B_k} p_i, \quad \bar{y}_k := \frac{1}{n_k} \sum_{i : p_i \in B_k} y_i.$$

[2] On the log rule, see also [18] and [19].

[3] Some authors have also argued that which scoring rule we ought to use depends on what we need the scores for; see [20], [21], and [22].

[4] Some authors (e.g., [28]; [29]) effectively identify belief with certainty. However, most think this is too demanding and entails skepticism (though [29] is an attempt to avoid a full-blown skepticism by making credences context-sensitive).

Then our measure of miscalibration is the *squared calibration error*:

$$\frac{\sum_{k:n_k>0} n_k \left(\bar{p}_k - \bar{y}_k\right)^2}{\sum_k n_k}.$$

Note that lower is better here, and that 0 means that the person is empirically perfectly calibrated (for the chosen binning).

Sharpness. While we want forecasters to be well-calibrated, a well-calibrated forecaster can still be unhelpful. To illustrate, take the dataset used for our study, which consists of probability assignments to a set of propositions with half of the set true and the other half false. Any participant who had assigned every proposition a "fifty–fifty" probability would have been perfectly calibrated: exactly half of all the propositions rated "fifty–fifty" would have been true, and there would have been no propositions assigned a different probability. However, as forecasts, these assignments would not have been of much value, given that they entirely fail to discriminate between true and false propositions. We want forecasters to be well-calibrated, but we also hope that, at least on the set of propositions that are of interest to us, they assign a high probability (ideally, a perfect probability) to truths and a low probability (ideally, a probability of 0) to falsehoods.

The notion of sharpness helps us to formalize this additional desideratum. Informally, for a set of propositions, sharpness asks: "How far from 'fifty–fifty' are the probabilities the forecaster assigns to these propositions, on average?" Formally, the sharpness of a person with probability function Pr defined on $\Phi = \{\varphi_1, \ldots, \varphi_N\}$ equals

$$\frac{1}{N}\sum_{i=1}^{N} |p_i - 0.5|,$$

where, recall, $p_i = \Pr(\varphi_i)$.

Two comments on this: First, it should be emphasized that sharpness alone is of no interest. After all, a forecaster scoring high on sharpness on a given set of propositions may be entirely misguided, having high confidence in many falsehoods and low confidence in many truths. We want trustworthy (i.e., well-calibrated) forecasters who *also* score high on sharpness on the relevant set of propositions. That is why sharpness is an *additional* desideratum, as we said.

Second, once we fix θ what matters for a forecaster's Lockean accuracy is only their sharpness on the propositions in the set we symbolized as Φ_θ. Whether, for instance, the forecaster assigns a "fifty–fifty" probability to the propositions *not* in that set or assigns them all a probability of either θ or $1-\theta$ does not change the forecaster's Lockean accuracy at the given θ, as they influence neither the denominator nor the numerator in the definition of ACC_θ. Thus, in our analysis, we will always specifically look at *sharpness at* θ (or *sharpness*$_\theta$), which is defined as

$$\frac{1}{|\Phi_\theta|}\sum_{i:\varphi_i\in\Phi_\theta} |p_i - 0.5|.$$

Coverage. The final property to be mentioned here is entirely straightforward. We previously defined Φ_θ to be the set of propositions either believed or disbelieved by a person with probability function Pr and threshold θ. Then *coverage at* θ (or coverage$_\theta$) for that person is equal to

$$\frac{|\Phi_\theta|}{N}.$$

Note that a coverage of 0 means that a person is agnostic about *everything* in Φ, while a coverage of 1 means that the person is agnostic about *no* proposition in Φ.

The idea of coverage has been discussed more in philosophy than in statistics. Specifically, epistemologists have argued that we ought to "amass a large body of beliefs with a favorable truth–falsity ratio" ([33]; also [34] and [35]). As has been recognized, the two components of this epistemic goal—having a *large* body of beliefs, and having a body of beliefs *with a favorable truth–falsity ratio*—will generally pull in different directions, and while setting θ closer to 1 will make it easier to satisfy the second desideratum, it will make it more difficult to satisfy the first, and the other way around if we set θ closer to .5.

It is obvious how coverage$_\theta$ impacts Lockean accuracy: it is the denominator in ACC_θ. As for calibration and sharpness$_\theta$, note that, *under fixed calibration*, for a believed proposition φ_i, increasing p_i increases the chance that φ_i is true, while for a disbelieved proposition φ_j, decreasing p_j increases the chance that the proposition is false.

3. Study

The foregoing theoretical discussion provides the foundation for our empirical work. To move beyond the largely *a priori* debate surrounding scoring rules, we conducted a study to investigate their performance on a practical and intuitive criterion: the prediction of Lockean accuracy. To detail our study, we begin by outlining the methods we used, including the dataset and our analytical procedure. We then present the results, addressing both our primary question of which rule or rules perform best and our secondary, explanatory question of why the rules do as well or as poorly as they do.

3.1. Methods

Data. We reanalyzed data collected for the study on judgment aggregation reported in Stinson et al. [36]. These data consist of the probability judgments from 376 participants (all US citizens) of 1200 general knowledge claims (e.g., "The first handheld calculator was built before 1975," "Mexico was the main trading partner of the US in the 1990s"), half of which are true and half false. Participants assigned probabilities to all of these claims in six online sessions, in each of which 200 unique claims were presented, in an order randomized per participant.[5]

Analysis procedure. The analysis will proceed in three stages, designed to first identify the best-predicting scoring rules and then to explain the source of their predictive power.

In the first stage, we calculate the core measures. Specifically, we compute, for each of the 376 participants in the dataset, their overall gradational accuracy according to the six scoring rules defined in Section 2.1 (i.e., AbsError, Brier, Power^3, Power^4, log, and Spherical). Next, we calculate each participant's Lockean accuracy (ACC_θ) for θ from .5 to .95 in steps of .05. Finally, for use in our explanatory analysis, we compute the three key forecaster properties for each participant, that is, the squared calibration error across all their forecasts, and for each value of θ, their $sharpness_\theta$ and $coverage_\theta$.

Then to address our primary research question, in the second stage we investigate the direct relationship between gradational and Lockean accuracy. For each threshold θ, we compute the correlation between each of the six scoring rules and the participants' Lockean accuracy scores. We use both Pearson's r to measure linear association and Spearman's ρ to assess rank-based association (we may want to use the latter whenever our goal is simply to identify the best forecasters rather than to predict the precise value of their accuracy). To get a sense of the generalizability of our correlation outcomes, we use a two-factor bootstrap procedure, which involves repeatedly resampling both participants and claims and recalculating the correlations for each new sample ([37]).

While the raw correlations answer *which* rules best predict Lockean accuracy, they do not explain *why*. A high correlation might arise because a rule effectively rewards fundamental epistemic virtues, but it could for instance also be exploiting a statistical artifact in the dataset. To open this black box and understand the source of each rule's predictive power, we finally conduct a mediation analysis.

More specifically, the goal of this analysis is to determine how much of a scoring rule's success is mediated by its ability to track the generic forecaster properties presented in Section 2.2. As pointed out there, the choice of these properties is not arbitrary. Not only are they anchored in the existing literature, they are also closely linked to the mathematical and conceptual components of Lockean accuracy itself, with (i) $coverage_\theta$ directly measuring the denominator of the Lockean accuracy formula; (ii) $sharpness_\theta$ measuring the average decisiveness of the judgments that make up the numerator; and (iii) calibration connecting these decisive judgments to the world.

Our method for this stage of the analysis is partial correlation. For each threshold θ, we recalculate the correlation between each scoring rule and Lockean accuracy, this time statistically controlling for (or "partialling out") the influence of the three forecaster properties. By comparing the raw correlations computed in the second stage with these partial correlations, we can quantify how much of a scoring rule's predictive success is explained by these fundamental virtues.

3.2. Results

Before presenting the results of our primary analyses, we first characterize the dataset to validate its quality and illustrate the key empirical patterns most directly relevant to our central research questions.

Descriptive statistics. The left panel of Fig. 1 shows density plots of four randomly chosen participants from the Stinson et al. study. These are quite representative for the kind of individual differences we find in the dataset. Some participants are cautious and stay close to the middle (like participant 4), some have a tendency to go more to the extremes of the response scale (like participants 34 and 165), and some show a mixed picture (like participant 210). This already hints at why ranking forecasters is a hard problem. If every participant had a perfect U-shaped distribution, predicting their accuracy might be trivial. But reality is messier, as the plots show, which explains the need for sophisticated tools to distinguish good forecasters from poor ones.

To verify that participant ratings are not mere noise and tend to track the truth, we fit a logistic generalized linear mixed model (GLMM) with truth values as response variable and probability rating as predictor, and with per-participant random intercepts and slopes. This showed probability rating to be a highly significant predictor of truth ($z = 50.77$, $p < .0001$). The right panel of Fig. 1 shows the fit, which also clearly shows that higher ratings are associated with higher probability of truth.

Based on the distributions shown in the left panel of Fig. 1, one expects that, as θ increases, coverage (the propositions that are either accepted or rejected) decreases. At the same time, from the plot in the right panel of that figure we know that participants tend to have higher confidence in true propositions than in false ones, and thus we expect Lockean accuracy to increase with θ. Fig. 2, which plots both mean Lockean accuracy and mean coverage (the mean in both cases over all participants) for all $\theta \in \{.5, .55, \ldots, .95\}$, confirms that both trends are present in the data.

Fig. 3 summarizes distributions of two generic main forecaster properties that we will later control for and that will help us explain our main results. The left panel shows the distribution of squared calibration error computed over all forecasts. We see that most of

[5] For a detailed description of the data and the collection procedure, see [36].

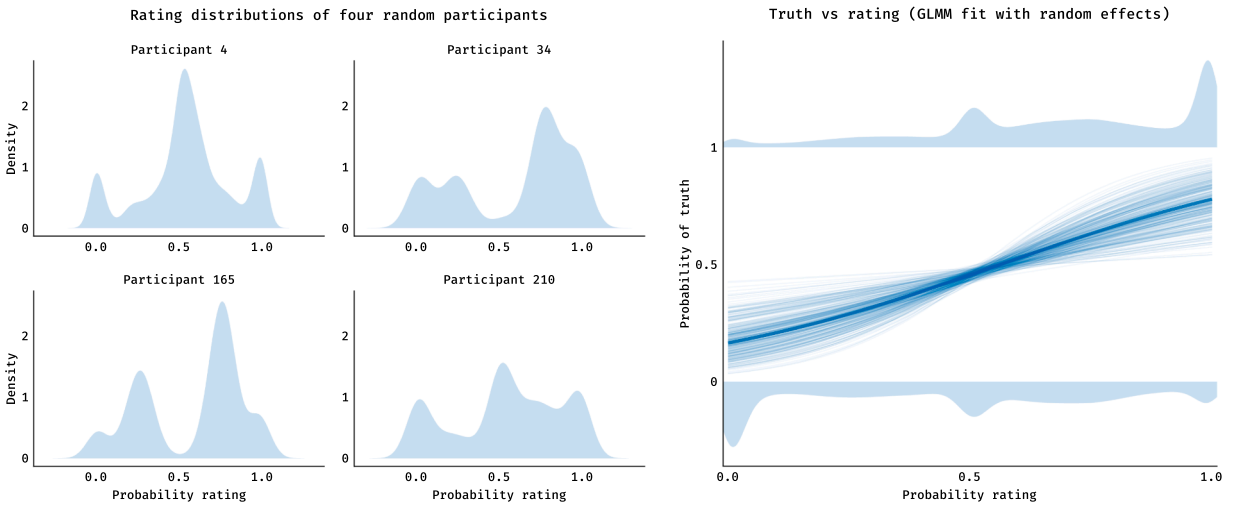

Fig. 1. Left: Density plots of rating distributions for four randomly selected participants (illustrating substantial individual variation). Right: Logistic GLMM predicting truth from rating with per-participant random intercepts and slopes; the thick curve shows the population-level fit and thin curves the participant-level fits. Marginal plots show the densities of ratings, split by truth value (top: true; bottom: false).

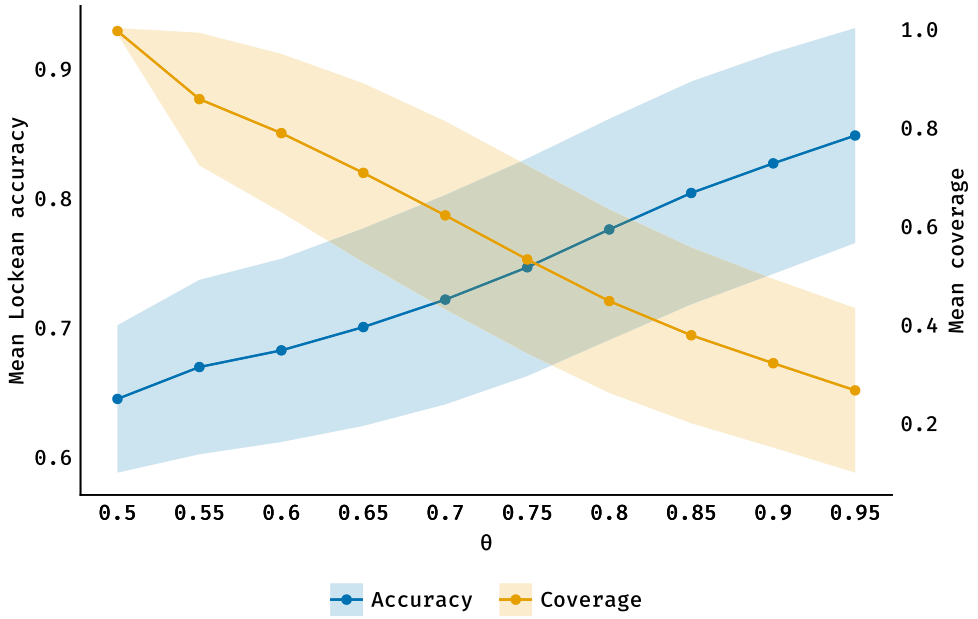

Fig. 2. Average Lockean accuracy and coverage for all thresholds θ. Shaded areas represent one standard deviation from the mean.

the mass lies near zero, which indicates that many participants are reasonably calibrated overall, although there is also evidence in the plot of substantial heterogeneity across individuals. Neither should come as a surprise, in view of the GLMM plot shown in Fig. 1, which showed that participants' probability ratings tend to track truth, though with quite a bit of variability among participants in how well their ratings track truth. The right panel is a ridge plot of sharpness at θ, rescaled and centered to make the distributions comparable across thresholds.[6] We see that, as θ increases, the distributions shift rightward, which indicates that once people are

[6] Specifically, for each θ and each participant i, we rescaled and centered using

$$\frac{\mathrm{sh}_{\theta,i} - \frac{\theta}{2}}{(1-\theta)/2},$$

where $\mathrm{sh}_{\theta,i}$ is the participant's mean $|p - 0.5|$ restricted to the items actually classified at θ.

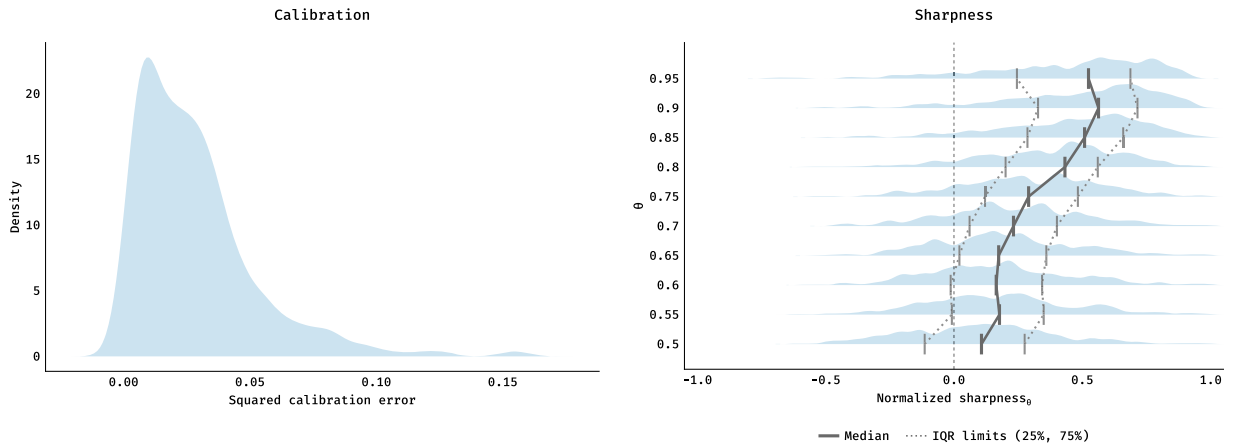

Fig. 3. Left: Distribution of squared calibration error (lower is better) across participants, computed over all forecasts. Right: Ridgeline rescaled and centered distributions of sharpness at θ on the classified items, with medians and interquartile ranges (IQR) overlaid to highlight the trend.

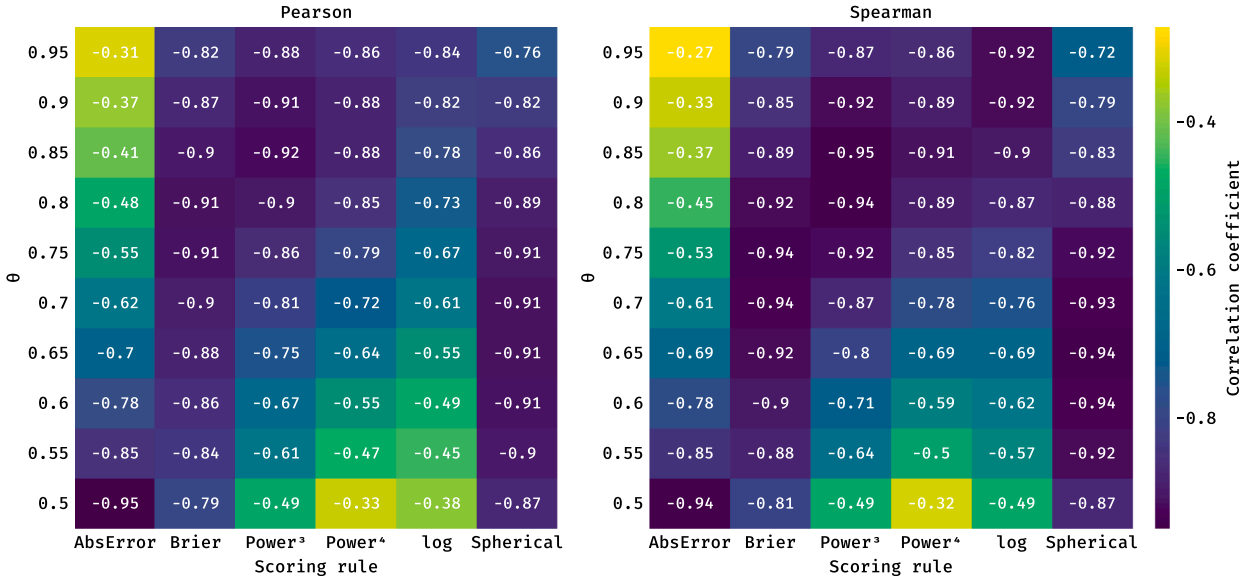

Fig. 4. Correlations between scoring rules and Lockean accuracy across threshold values. Each cell shows the correlation coefficient between a scoring rule's performance and Lockean accuracy at threshold θ. Left panel shows Pearson correlations measuring linear association; right panel shows Spearman rank correlations.

confident enough to cross a higher threshold, they tend to be *much* more confident, that, put differently, when participants classify (i.e., accept or reject) at higher θ, they do so with disproportionately extreme probabilities rather than merely meeting the threshold.

Correlations. We now turn to our primary research question: Which scoring rule best predicts Lockean accuracy? Fig. 4 presents the results of our correlational analysis. The two heatmaps show the Pearson and Spearman correlations between each of the six scoring rules (columns) and participants' Lockean accuracy at each threshold θ (rows). Because scoring rules measure loss (lower is better), a strong negative correlation is desirable, as it indicates that participants with better scores on a given rule also tend to achieve higher Lockean accuracy.

The most apparent result is that there is no single best scoring rule across all contexts. Instead, what is the best-performing rule depends strongly on the threshold θ. For lower thresholds (approximately $\theta = .5$ to .65), the Spherical score consistently shows the strongest correlations for both Pearson and Spearman measures. In contrast, for higher thresholds ($\theta \geqslant .8$), the Power^3 rule emerges as the superior predictor.

Notably, the two most popular rules in the literature, the Brier and log scores, rarely do best. While the Brier score is competitive in the mid-range of thresholds, it is outperformed by the Spherical and Power^3 rules at the lower and upper ends, respectively. The log score is only the best predictor in the single case of Spearman correlation at the highest threshold of $\theta = .95$. The overall story is

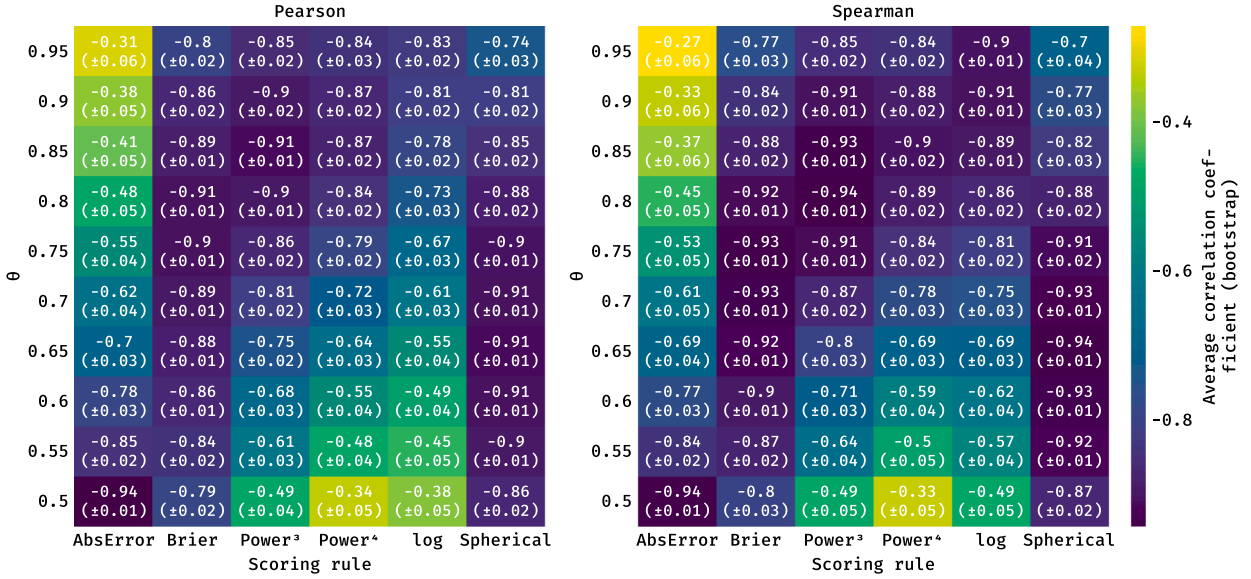

Fig. 5. Bootstrap means (±1 SD) for correlations between scoring rules and Lockean accuracy (based on 2000 bootstrap samples resampling both participants and statements). Left panel shows Pearson correlations; right panel shows Spearman rank correlations.

broadly consistent across both Pearson and Spearman correlations, suggesting that the findings are robust to whether one cares about linear predictability or simple rank ordering.[7]

To ensure that these findings are generalizable and not merely an artifact of our specific sample of participants and claims, we conducted a two-factor bootstrap analysis, which resamples both participants and statements. Fig. 5 presents the mean correlation coefficients and their standard deviations over 2000 bootstrap resamples. The bootstrap analysis confirms the robustness of our initial findings. The patterns observed in Fig. 4 are clearly replicated and the small standard deviations in all cells indicate that the results are highly stable.

Mediation analysis. The previous analysis established *which* scoring rules best predict Lockean accuracy, but it did not explain *why*. A strong correlation could mean a rule is genuinely rewarding the epistemic virtues that lead to Lockean accuracy, or it could be exploiting some other statistical feature of the data. To distinguish these possibilities and understand the source of each rule's predictive power, we conducted a mediation analysis.

The core idea of a mediation analysis is to investigate the pathway through which one variable (in our case, a scoring rule) influences another (in our case, Lockean accuracy). We hypothesize that this influence is not direct, but is *mediated* by the generic forecaster properties of calibration, sharpness, and coverage that were discussed in Section 2.2. In other words, we test the hypothesis that a good scoring rule works primarily by being an effective proxy for these more fundamental virtues.

Our descriptive analysis (see especially Fig. 3) provides the necessary foundation for this test. It showed that our participant pool exhibits considerable variation in terms of calibration (albeit with most participants being quite well calibrated) and that sharpness_θ systematically increases with the threshold θ. Thus, there is meaningful variance in both properties that could make them usable in an explanation of our correlation results.

To quantify the extent of the mediation, we use partial correlation, which allows us to recalculate the correlation between each scoring rule and Lockean accuracy while statistically controlling for the influence of calibration, sharpness_θ, and coverage_θ. The result is the *residual* association—the predictive power that a rule retains even after we have accounted for its ability to track these three core properties.

Fig. 6 presents the results of this analysis. The left panel shows the partial correlations, while the right panel shows the "Delta" (Δ), calculated as the raw correlation (from Fig. 4) minus the partial correlation. This value Δ is a direct measure of mediation: a large negative Δ indicates that most of the original predictive power of the rule was indeed mediated by the forecaster properties for which we controlled. Conversely, a Δ value close to 0 suggests that the rule has a more direct predictive link to Lockean accuracy that is independent of these generic virtues.[8]

[7] In this connection, it is also worth noting that Ahmadian et al. [38] have recently shown that the Brier and log scores can, in certain cases, assign better scores to misclassifications than to correct predictions, which led these authors to propose modified scoring rules designed to avoid this flaw. We did not include these new rules in our analysis because they explicitly penalize misclassifications, making their correlation with Lockean accuracy—itself a measure of classification success—somewhat circular. Our interest lies in understanding how and why scoring rules that do not build in such penalties can nonetheless predict Lockean accuracy. (Additionally, Ahmadian and colleagues' new rules were developed for multi-class machine learning classification tasks, a setting that differs from ours in several important respects.)

[8] The results shown in Fig. 6 are for the Pearson correlations. For the mediation analysis of the Spearman correlations, which are qualitatively identical, see the Supplementary Materials.

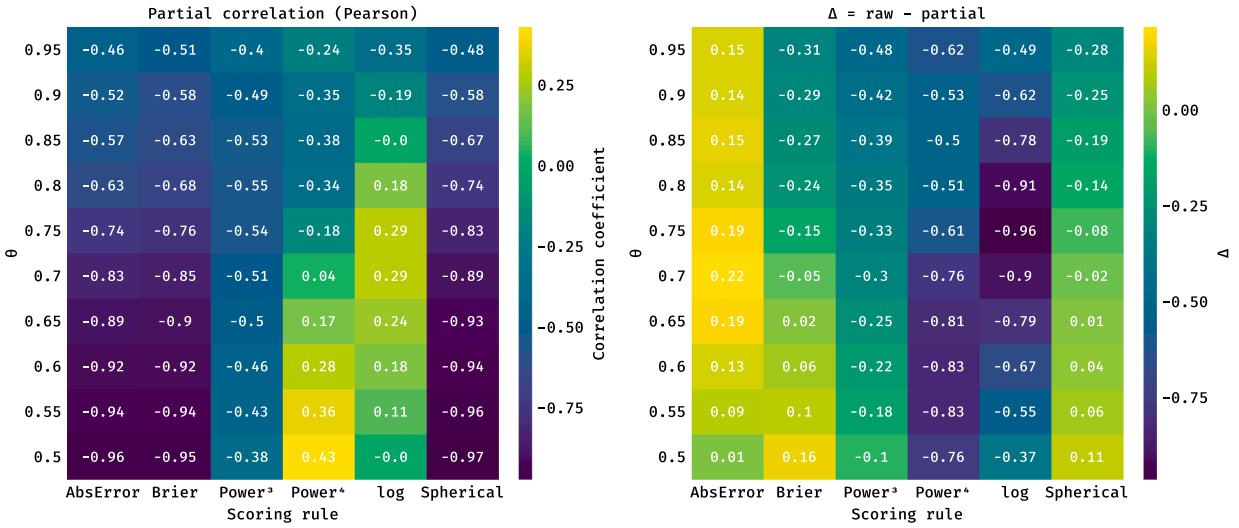

Fig. 6. Mediation analysis of scoring rule performance. The left panel shows the partial Pearson correlations between each scoring rule (columns) and Lockean accuracy at each threshold θ (rows), after statistically controlling for calibration, sharpness$_\theta$, and coverage$_\theta$. The right panel shows the Δ value, calculated as the raw correlation (from Fig. 4) minus the partial correlation. This Δ value quantifies the mediation effect: a more negative value indicates that a larger portion of a rule's predictive power is explained by its ability to track the three generic forecaster properties.

Several key patterns emerge from this analysis. First, for most rules at most thresholds, the partial correlations (left panel) are weaker than the raw correlations, and the Δ values are (often strongly) negative. This at least partially confirms our central hypothesis: for most scoring rules, a large portion of their success at predicting Lockean accuracy comes from the fact that they reward well-calibrated, sharp, and decisive forecasters.

Second, there are important differences between the rules. The log, Power^3, and Power^4 rules show the largest mediation effects (i.e., the most negative Δ values), especially at higher thresholds. This means that, once we account for their ability to reward calibration and sharpness, their independent predictive power diminishes starkly. We can thus conclude that these rules function almost exclusively as proxies for these fundamental properties.

Third, the positive Δs for the absolute error rule show a so-called suppression effect, meaning that the rule becomes a *better* predictor of Lockean accuracy if we control for calibration, sharpness$_\theta$, and coverage$_\theta$. To the extent that we value these properties in forecasters, that is a reason *not* to use the absolute error rule. Of course, Figs. 4 and 5 did not give much reason to consider the rule as a serious option anyway, since it only works best for the lowest one or two values of θ.

Finally, while their performance is also partially mediated, the Spherical and Brier rules retain substantial predictive power even after controlling for the generic properties (their partial correlations remain strong, and their Δ values are closer to zero). This is an interesting finding, although it is not immediately evident how to interpret it. The fact that these rules track an additional component beyond calibration, sharpness, and coverage does not automatically mean that this component is a *desirable* epistemic virtue in a general sense. It could, for instance, represent a cognitive bias or a response pattern that is particularly effective for the general knowledge questions that served as materials in Stinson et al.'s study (from which our data come) but that is not more broadly valuable. What we *can* confidently conclude is that, for the specific task of predicting Lockean accuracy in our present context, this additional component is *instrumentally* effective: the Spherical and Brier rules are not merely good proxies for the standard forecaster virtues but appear to track what it takes to be Lockean-accurate in a more direct or comprehensive way than the other rules. At a minimum, the finding suggests a more complex relationship between gradational and Lockean accuracy than is captured by the three forecaster properties alone.[9]

4. Conclusion

The long-standing debate over which scoring rule best measures the accuracy of our beliefs has been conducted largely on an *a priori* basis, focusing on the abstract mathematical properties of the rules. In this paper, we pursued a different, empirical approach. We asked which among the most frequently used scoring rules best predicts a practical and decision-relevant criterion: a person's

[9] Without further experimental work, one can only speculate about the nature of the relationship. But here is a speculation that is at least reasonable in view of recent work in epistemology. Molinari [39] has shown that the degree of incoherence of a credence function can be characterized as the amount of "guaranteed accuracy" the agent foregoes by adopting that function. If an analogous characterization applies in our case, the residual predictive power of the Spherical and Brier rules might reflect their sensitivity to some form of epistemic efficiency or coherence that is not fully captured by the properties considered in our analysis.

Lockean accuracy, that is, their ability to make correct categorical judgments. Our findings are clear and have both practical and theoretical implications.

Our main finding is that there is no single best scoring rule. The answer to "Which rule is best?" is context-dependent, the crucial piece of context being the belief threshold θ. For situations requiring a lower threshold for belief ($\theta \approx .5$–$.65$), the Spherical score is the superior predictor of Lockean accuracy. For situations requiring a high degree of confidence ($\theta \geqslant .8$), the Power^3 rule is best. Perhaps most surprisingly, our results show that the two most widely discussed rules in the literature, the Brier and log scores, are rarely the optimal choice for this task.[10]

This leads to a direct practical upshot for any context where probabilistic information must ultimately inform a categorical decision. The epistemically responsible choice is not to pick a scoring rule based on its abstract properties but to first determine the threshold for belief (presumably based on what is at stake in the given context) and then to select the scoring rule that is empirically best suited for that threshold.

Naturally, it may not always be possible to fix θ in advance. Then our results are still useful in that they can help one choose a "θ-robust" scoring rule. For instance, if it is wide open what value θ will have (i.e., the only constraint is that θ is in the upper half of the unit interval), then perhaps choose either the Spherical or the Brier rule, which at least on our data are the best "all-rounders." And if it is known that θ will be in a more limited range, the choice can be made specific to that (e.g., if decisions will regularly use very high θ, Power^3 would seem a good option). Or if we can specify a rough prior over θ, then a sensible recommendation would be to choose the rule that maximizes expected correlation under that prior (expected Pearson correlation if we want to predict levels of Lockean accuracy, or expected Spearman correlation if the goal is to rank forecasters).

On a more general level, our findings challenge the primacy of the *a priori* approach. We saw, for instance, that the properties that make the log score theoretically attractive do not necessarily make it the most effective tool for predicting real-world performance of the kind studied in this paper. Furthermore, our mediation analysis revealed that while, for most scoring rules, their predictive success is largely explained by their ability to track the established virtues of calibration and sharpness, these virtues do not always provide the whole story. In particular, the Spherical and Brier rules were seen to retain significant predictive power even after we controlled for the aforementioned properties, suggesting they are sensitive to some other instrumentally effective component of forecasting skill that may or may not be captured by our standard epistemic vocabulary.

Our findings are based on a dataset of general knowledge claims, and it remains to be seen whether these results generalize to other domains, such as political, economic, or medical forecasting. This suggests an obvious avenue for further research. Moreover, the nature of the "latent" property tracked by the Spherical and Brier rules is still unknown. Future work, both theoretical and empirical, could aim to isolate and identify this component of forecasting skill (and determine whether or not it is a generally desirable one). Doing so could deepen our understanding of scoring rules and might suggest a broader conception of what it means to be an accurate and reliable reasoner. Finally, we have limited ourselves to comparing the best known scoring rules and nothing we said excludes that tuned or hybrid rules could be engineered to perform well across broader θ ranges. This, too, is a issue we leave for another day.[11]

CRediT authorship contribution statement

Igor Douven: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization; **Nikolaus Kriegeskorte:** Writing – review & editing, Resources, Project administration, Conceptualization.

Data availability

No new data were collected for this paper. The dataset from Stinson et al. ([36]), which we reanalyzed, is available from Nikolaus Kriegeskorte (nk2765@columbia.edu) on reasonable request under a DUA.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Bezanson, A. Edelman, S. Karpinski, V.B. Shah, Julia: a fresh approach to numerical computing, SIAM Rev. 59 (2017) 65–98.
[2] H.E. Kyburg, Probability and the Logic of Rational Belief, Wesleyan University Press, Middletown CT, 1961.
[3] I. Douven, J. Uffink, The preface paradox revisited, Erkenntnis 59 (2003) 389–420.
[4] H. Leitgeb, The Stability of Belief: How Rational Belief Coheres with Probability, Oxford, Oxford University Press, 2017.
[5] J.K. Kruschke, Rejecting or accepting parameter values in Bayesian estimation, Adv. Methods Pract. Psychol. Sci. 1 (2018) 270–280.
[6] K. Tentori, V. Crupi, N. Bonini, D. Osherson, Comparison confirmation measures, Cognition 103 (2007) 107–119.

[10] In note 3, it was mentioned that, according to some authors, which scoring rule we should use depends on the purpose of scoring. If those authors are correct, our results show that, at least for Lockean-style decisions, there is a second layer of contextualism to be considered, which stems from the belief-threshold contextualism propagated by Leitgeb [4] and others.

[11] We are greatly indebted to Youness Ayaita, Christopher von Bülow, and two anonymous referees for valuable comments on previous versions.

[7] J. Zhao, V. Crupi, K. Tentori, B. Fitelson, D. Osherson, Updating: learning versus supposing, Cognition 124 (2012) 373–378.
[8] J.-L. Stilgenbauer, J. Baratgin, Assessing the accuracy of diagnostic probability estimation: evidence for defeasible modus ponens, Int. J. Approximate Reasoning 105 (2019) 229–240.
[9] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc. 102 (2007) 359–378.
[10] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1950) 1–3.
[11] R.D. Rosenkrantz, Foundations and Applications of Inductive Probability, Ridgeview, Atascadero CA,1981.
[12] J.M. Joyce, A nonpragmatic vindication of probabilism, Philos. Sci. 65 (1998) 575–603.
[13] R. Selten, Axiomatic characterization of the quadratic scoring rule, Exp. Econ. 1 (1998) 43–62.
[14] H. Leitgeb, R. Pettigrew, An objective justification of Bayesianism I: measuring inaccuracy, Philos. Sci. 77 (2010) 201–235.
[15] J. Bernardo, M, Expected information as expected utility, Ann. Stat. 7 (1979) 686–690.
[16] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, New York, Wiley, 2000.
[17] J.E. Bickel, Some comparisons between quadratic, spherical, and logarithmic scoring rules, Decis. Anal. 4 (2007) 49–65.
[18] D. Fallis, Attitudes toward epistemic risk and the value of experiments, Stud. Log. 86 (2007) 215–246.
[19] D. Fallis, P.J. Lewis, The brier rule is not a good measure of epistemic utility (and other useful facts about epistemic betterness, Australas. J. Philos. 94 (2016) 576–590.
[20] G. Schurz, Hume's Problem Solved: The Optimality of Meta-induction, MIT Press, Cambridge MA, 2019.
[21] I. Douven, Scoring in context, Synthese 197 (2020) 1565–1580.
[22] I. Douven, The Art of Abduction, MIT Press, Cambridge MA, 2022.
[23] R.L. Winkler, Scoring rules and the elicitation of subjective probabilities, in: D.A. Freeman, S.F.S. Freeman (Eds.), Bayesian Analysis in Statistics and Econometrics, Hoboken NJ, Wiley, 1996, pp. 213–225.
[24] T. Shear, B. Fitelson, Two approaches to belief revision, Erkenntnis 84 (2018) 487–518.
[25] M. Kaplan, A Bayesian theory of rational acceptance, J. Philos. 78 (1981) 305–330.
[26] P.K. Moser, J. Tlumak, Two paradoxes of rational acceptance, Erkenntnis 23 (1985) 127–141.
[27] R. Foley, Working without a Net: Essays in Egocentric Epistemology, Oxford University Press, New York NY, 1993.
[28] T. Williamson, Knowledge and Its Limits, Oxford University Press, Oxford, 2000.
[29] R. Clarke, Belief is credence one (in context), Philosophers' Imprint 13 (2013) 1–18.
[30] J. Hawthorne, L. Bovens, The preface, the lottery, and the logic of belief, Mind 108 (1999) 241–264.
[31] H. Lin, K.T. Kelly, A geo-logical solution to the lottery paradox, with applications to conditional logic, Synthese 186 (2012) 531–575.
[32] A.H. Murphy, A new vector partition of the probability score, J. Appl. Meteorol. 12 (1973) 595–600.
[33] W. Alston, Concepts of epistemic justification, Monist 68 (1985) 57–89.
[34] C. Sartwell, Why knowledge is merely true belief, J. Philos. 89 (1992) 167–180.
[35] I. Douven, The lottery paradox and our epistemic goal, Pac. Philos. Q. 89 (2008) 204–225.
[36] P. Stinson, J.V.D. Bosch, T. Jerde, N. Kriegeskorte, Collective inference of the truth of propositions from crowd probability judgments, https://arxiv.org/abs/2501.04983.
[37] H.H. Schütt, A.D. Kipnis, J. Diedrichsen, N. Kriegeskorte, 2023. Statistical inference on representational geometries, eLife 12 (2023) e82566.
[38] R. Ahmadian, M. Ghatee, J. Wahlström, Superior scoring rules for probabilistic evaluation of single-label multi-class classification tasks, Int. J. Approximate Reasoning 182 (2025) 109421.
[39] G. Molinari, An accuracy characterisation of approximate coherence, Synthese 203 (2024) 68.