Scoring, truthlikeness, and value

Igor Douven

Scoring, Truthlikeness, and Value*

Igor Douven
IHPST/CNRS/Panthéon–Sorbonne
igor.douven@univ-paris1.fr

Abstract

There is an ongoing debate about which rule we ought to use for scoring probability estimates. Much of this debate has been premised on scoring-rule monism, according to which there is exactly one best scoring rule. In previous work, I have argued against this position. The argument given there was based on purely a priori considerations, notably the intuition that scoring rules should be sensitive to truthlikeness relations if, and only if, such relations are present among whichever hypotheses are at issue. The present paper offers a new, quasi-empirical argument against scoring-rule monism. This argument uses computational simulations to show that different scoring rules can have different economical consequences, depending on the context of use.

1. Introduction. Weather forecasts commonly refer to probabilities. There is going to be rain tomorrow with a probability of x, snow with a probability of y, but it may also stay dry, with a probability of $1 - x - y$. The media may sometimes report such forecasts in strictly qualitative terms, by stating that there is going to be rain tomorrow (if the probability for rain is over 95 percent, say), or that they expect tomorrow to be windy (perhaps if the probability for strong winds is over 75 percent). But that is only because such qualitative reports are easier to interpret for the general public.

There is an ongoing debate about how weather forecasts, and probabilistic forecasts generally, are to be assessed. Participants to this debate share the conviction that probabilistic forecasts can be accurate to different degrees. The debate is about how to *measure* such degrees. Most authors agree on the "extreme" kind of case, in which one possible outcome is forecasted with certainty. Such a forecast is usually taken to be maximally accurate or maximally inaccurate, depending on whether the predicted outcome occurred or not. What to say about the intermediate cases is more contentious. For instance, there has been much discussion about whether the accuracy of a forecast should depend on anything other than what it predicted about the event that materialized. If tomorrow is rainy, then—some authors hold—there is no need to look at what the forecast said about the probability of snow.

Much of the debate has proceeded on the assumption that there is exactly one best measure of accuracy for probabilistic forecasts, or one best *scoring rule*, to use the by now standard name for such measures. It has been said, plausibly, that scoring rules should reward the

*The Supplementary Materials for this paper, containing the R code used for the simulations to be reported, can be retrieved from this repository: https://osf.io/n8e2g/?view_only=72d468e2eabe4100b9409985c4a10950.

features that we would like to see in forecasts (Cooke 1991, p. 121). But there is no unanimity about which those features are. As a result, there is no unanimity about which of the many scoring rules in the literature is the right one.

In previous work (Douven 2020, 2021, Ch. 5), I have argued against the near-consensus of scoring-rule monism.[1] My argument started from the observation that forecasts may concern events, or equivalently hypotheses, which exhibit some kind of order. In particular, if one hypothesis turns out to be true, the others need not all be "equally false": some false hypotheses may still be closer to the truth than others. Such relations of "truthlikeness" *need* not obtain, and—I argued—depending on whether they do, we should use different scoring rules.

While much of the debate about scoring rules has centered around the notion of accuracy, with proponents of different scoring rules emphasizing different intuitions about that notion, Murphy (1993) has been one of the few authors to draw attention to other aspects of a probabilistic forecast's goodness. For one, such a forecast should—in Murphy's view—be "based on the forecaster's rational distillation of the information contained in her knowledge base" (p. 282). The argument given in Douven (2020) was that, by ignoring information about truthlikeness relations among the hypotheses to which the to-be-scored probabilities are assigned, the standard scoring rules (see below) fail precisely on this count, at least in some contexts: in those contexts, they leave out relevant information about the hypotheses at issue even though this information is readily available to the forecaster.

Another aspect of a forecast's goodness highlighted by Murphy is that of *value*, by which he means that users who base their decisions on the forecast should benefit from it, economically or otherwise. As Murphy (1993, p. 286) notes, forecasts are not intrinsically valuable: "They acquire value through their ability to influence the decisions made by users of the forecasts." *How* scoring and value are connected is, as he also notes (p. 291), a still relatively unexplored question.

The present paper focuses on specific cases in which the scoring–value connection is easy to grasp, and uses them to argue again for the claim that in different contexts different scoring rules may be called for. Depending on the context, one scoring rule may be better at identifying valuable forecasts than another, meaning that using the former may lead to economically wiser choices than using the latter. This is illustrated by showing that, in contexts in which they apply, scoring rules sensitive to truthlikeness often—though not always—have a greater tendency to assign better (i.e., lower; see below) scores to more valuable forecasts than scoring rules not of that kind. My previous argument against scoring-rule monism was intuition-based: When we are scoring probabilities assigned to hypotheses that stand in certain relations of truthlikeness to each other, it makes *pre-theoretically good sense* to use a rule that is able to take such relations into account, while doing so makes *no* sense when truthlikeness relations are absent. By contrast, the new argument to be presented is quasi-empirical: I use computational simulations to show that the same scoring rules can be of different value in different contexts, in Murphy's sense. Before laying out the argument, I provide some background on scoring rules and the debate they have given rise to.

[1] For other dissenters, see Joyce (2009), Levinstein (2017), and Schurz (2019).

2. Theoretical background. This section states the main scoring rules and provides some detail about my previous reason for believing scoring-rule monism to be false.

The scoring of probabilistic forecasts is nothing but a generalization of the simple scoring of true/false-questions that we have been familiar with since we were first-graders. In the case of those questions, there was no way to do better than to assert, categorically, of something true that it was true and of something false that it was false; and there was no way to do worse than to assert, categorically, of something true that it was false and of something false that it was true. When we assign probabilities, we can still assign probability 1 to one particular hypothesis and probability 0 to any rival hypotheses. If we do and the hypothesis that we expressed full confidence about is true indeed, then again there is no way we could have done better. Scoring rules are usually taken to assign *penalties*, so in this case one would want to say that *zero* penalty is incurred. If the hypothesis that receives probability 1 is in fact false, then some would say there is no way we could have done worse, and so we should be maximally penalized. As will be seen below, in the latter case one could also argue that how badly one does by assigning probability 1 to a false hypothesis depends on how far from the truth that hypothesis is, so that a non-maximal penalty might be more appropriate. Setting that aside for now, the intermediate cases, in which we assign some positive probability to more than one hypothesis, are the trickier ones when it comes to scoring.

One main question has been whether it should matter at all what probability is assigned to any hypothesis other than the truth. Proponents of the so-called log rule answer this question in the negative.[2] Let $\{H_i\}_{i=1}^n$ be a hypothesis partition, that is, a set of mutually exclusive and jointly exhaustive hypotheses. Then, according to these authors, someone assigning probability p_i to hypothesis H_i, for each $i \in \{1, \ldots, n\}$, incurs a penalty of

$$\mathcal{L}\big(\{p_i\}_{i=1}^n, j\big) \;=\; -\ln(p_j),$$

on the assumption that H_j is the truth. Note that this means that the probabilities the person assigns to the false hypotheses can be ignored.

The log rule is one of the two most popular scoring rules, the other being the Brier score, according to which the same person would incur a penalty of

$$\mathcal{B}\big(\{p_i\}_{i=1}^n, j\big) \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(\delta_{ij} - p_i\big)^2,$$

with δ_{ij} being 1 if $i = j$ and 0 otherwise. For proponents of this rule, it clearly not only matters what probability is assigned to the truth but also what probabilities are assigned to the rest of the hypotheses.[3] It can for instance be shown that, given that a person assigns probability p to the truth, that person minimizes her Brier score by assigning one and the same probability of $(1 - p)/(n - 1)$ to each of the false hypotheses.

But on some occasions ignoring the probabilities assigned to hypotheses beyond the truth—as the log rule does—and rewarding a flat assignment to those probabilities—as the Brier score does—*both* seem wrong. Consider three hypotheses about the outcome of a

[2] See, for instance, Good (1952), Bernardo (1979), Bernardo and Smith (2000), Bickel (2007), and McCutcheon (2019). See also Fallis (2007) and Fallis and Lewis (2016).

[3] See, for instance, Brier (1950), Rosenkrantz (1981), and Selten (1998). Joyce (1998) also advocates the Brier score as the one true scoring rule, but he abandons scoring-rule monism in his (2009).

football match: the match ends in a home win (H), it ends in a draw (D), or it ends in an away win (A). Suppose you believe the home team to be the stronger one and therefore deem H most likely. A friend of yours thinks the other team is stronger and therefore deems A most likely. Another friend, finally, thinks the two teams are about equally strong and therefore deems D most likely. Suppose the home team wins. Then, it would seem, any scoring rule that does not have you come out as receiving the lowest score (i.e., the lowest penalty) should be rejected out of hand. But it equally seems that the friend who deemed a draw most likely should receive a lower score than the friend who deemed an away win most likely, given that hypothesis D is in an intuitively clear sense closer to what turned out to be the truth than hypothesis A. If necessary, the intuitively clear sense can even be formally explicated using work on truthlikeness by authors such as Kuipers (2000, 2001, 2019), Niiniluoto (1984, 1998, 1999), and Schurz (1987, 1991, 2011).

While the log and Brier score do not allow us to take such truthlikeness relations into account when scoring a probabilistic forecasts, there are scoring rules that do. In Douven (2020), I introduced a family of (what I called) verisimilitude-sensitive scoring rules (VS rules, for short), which assign weights to hypotheses based on their distance from the truth and then let those weights determine how much the probabilities assigned to those hypotheses contribute to the overall penalty. Where H_j is the true member of hypothesis partition $\{H_i\}_{i=1}^n$, a VS rule assigns a penalty of

$$\mathcal{V}\left(\{p_i\}_{i=1}^n, j\right) \quad = \quad \sum_{i=1}^{n} \omega_{ij} \left(\delta_{ij} - p_i\right)^2$$

to someone whose probability for H_i is p_i, for all $i \in \{1, \ldots, n\}$, and with δ_{ij} as before. The weight ω_{ij} measures how far from the truth H_i is, for $i \in \{1, \ldots, n\}$. The only constraints on the weights are that they must be positive, sum to 1, and reflect the truthlikeness relations among the hypotheses at least to the extent that hypotheses further from the truth should be weighted more heavily than hypotheses closer to the truth. Needless to say, that leaves a lot of freedom, meaning that there is a broad range of VS rules.

There is, however, a seeming problem with all of these rules, related to a property that many regard as a desideratum for scoring rules, to wit, that of *propriety*. A scoring rule R is said to be *proper* exactly if the expected R-score of any given probability assignment is minimal relative to that same assignment. To illustrate, supposing I assign probabilities 2/3, 1/3, and 0 to the aforementioned hypotheses H, D, and A, then the R-score I *expect* to receive is 2/3 of the R-score I receive if the home team wins plus 1/3 of the R-score I receive in case of a draw. If, and only if, R is proper, my expectation is that I cannot do better than this; relative to my current probabilities, by adopting *other* probabilities I can only do worse. If, for any probability assignment, the expected R-score is minimal *only* relative to that assignment—by adopting other probabilities I *will* do worse, relative to my current probabilities—then R is said to be *strictly proper*.

As proven in Douven (2020), VS rules are, without exception, *im*proper. But while some might find that reason to immediately reject these rules, in the same paper I argued that that would be rash. Propriety is an important property of scoring rules when such rules are used to elicit probabilities, given that an improper scoring rule could incentivize a person to lie about her probabilities. But scoring rules can serve other purposes as well. Suppose, for instance,

a broadcasting company intends to hire a new weather forecaster. There are a number of candidates, and to assess them, the company scores all the forecasts these candidates made in the past year. The company hopes to be using a scoring rule which helps them make a good hire. Whether that rule is proper, or strictly proper, is immaterial. After all, the forecasts are already public and can no longer be altered. And there is no argument to the effect that the, or a, scoring rule that helps the company make the best hire must be proper.[4]

More importantly still, as also noted in Douven (2020), not all scoring rules that are sensitive to truthlikeness are improper. In particular, the ranked probability score (RPS) is not (Epstein 1969; Murphy 1969). According to this rule, if my probability for H_i is p_i, for all $i \in \{1, \ldots, n\}$, with $\{H_i\}_{i=1}^n$ again a hypothesis partition and H_j being the truth, my ranked probability score equals

$$\mathcal{R}\big(\{p_i\}_{i=1}^n, j\big) \;=\; \frac{\sum_{k=1}^n \Big(\gamma_{kj} - \sum_{i=1}^k p_i\Big)^2}{n-1},$$

where $\gamma_{kj} = 1$ if $k \geqslant j$, and 0 otherwise.

If a scoring rule is available that is both proper and sensitive to truthlikeness relations, should we care at all about the VS family, or other scoring rules that are sensitive to truthlikeness but at the expense of propriety? We might be inclined to say *no*, but that would not necessarily be the right response. For VS rules have a seemingly desirable feature which the RPS rule lacks.[5]

To see this, note that for the RPS rule there is no more to truthlikeness relations among the members of a hypothesis partition than their *ordering* in that partition. Now consider this example: Two football teams that are about equally strong, and whose past five encounters have all been draws, just played another match, in which neither team managed to score. Then the prediction that the match would end in a 0 : 4 away win appears to be quite a bit further from the truth than the prediction that the match would end in a 0 : 1 win. After all, the former much more than the latter would suggest a strongly dominant away team, which would have been a surprise in view of the relative strengths of the teams and the outcomes of their past encounters. On the other hand, a 0 : 24 away win prediction would hardly be further from the truth than a 0 : 21 away win prediction: we would find these end results about equally stunning and might say that both are "about as far from the truth as can be."[6]

[4] It is entirely consistent with everything said in the present paper, or in Douven (2020), that there are *still* other purposes that scoring rules can serve, and that some of those may again require propriety. For instance, Roche and Shogenji (2018) argue that we should measure informativeness in terms of inaccuracy reduction, and that inaccuracy should then be measured by a proper scoring rule. There is no conflict here with the claim made in Douven (2020), which after all is merely that scoring rules can *also* serve purposes which do *not* require propriety.

[5] Thanks to Ilkka Niiniluoto for bringing this to my attention.

[6] An anonymous referee disagreed at this point, maintaining that we are to measure distance from the truth here in terms of the difference in goals scored. In my opinion, it is more reasonable to look at how different the various mentioned non-actual worlds (the world in which the match ends in a 0 : 1 win, the world in which the match ends in a 0 : 4 win, and so on) are from the actual world. And given what we know about the teams, our world would, as mentioned, have to be rather different from the actual world for the match to end 0 : 4 while it would not have to be very different for the match to end 0 : 1. By contrast, for the match to have ended 0 : 21, something entirely out of the ordinary would have had to occur, and whatever that would have been, it would have been about equally compatible with a 0 : 24 end result. For instance, if all players who normally play for the home team had been suspended, and the coach of that team had to line up their most inexperienced players, then

Or consider any value representable in a similarity space (in the manner of Gärdenfors 2000), say, a color shade. We might want to order different hypotheses about this value on the basis of distances in color space (CIELAB space or CIELUV space; see Fairchild 2013 or Douven et al. 2017), but while such distances correlate well with human judgments at a short range, they stop doing so at a longer range (Shepard 1987). Thus, if the actual value of the color we are looking for is some shade of red, then the hypothesis that it is a particular shade of orange may be closer to the truth than the hypothesis that it is a particular shade of yellow, but the hypothesis that it is a particular shade of blue would probably appear as far from the truth as the hypothesis that it is particular shade of green, even if it turned out that, say, the particular shade of blue is closer in color space to the actual shade than the particular shade of green.

It is hard to see how one could accommodate the intuitions at play in these examples just in terms of an ordering of hypotheses. However, one can find refined measures of truthlikeness in the literature that do allow one to go beyond merely ordering hypotheses in terms of truthlikeness and to express truthlikeness relations among the aforementioned hypotheses in a way which does justice to the aforementioned intuitions (see, e.g., Niiniluoto 1984, Ch. 7, 1987, Ch. 12; Kuipers 2000, Ch. 12). And while the RPS rule cannot take such refined measurements into account, VS rules *can*.

As a further comment, I would like to point to what I regard as an open question concerning VS rules. Oddie (2019) argues that a condition he terms "Proximity" is a desideratum for any accuracy measure. According to this condition, your accuracy should not decrease if you go from being certain that H to being certain that a particular H-world is actual, provided the latter is among the H-worlds closest to the actual world. As an anonymous referee observed, VS rules fail to satisfy this condition, on the supposition that the relevant weights reflect truthlikeness relations.[7,8] That is only a problem for VS rules insofar as we are committed

---

a devastating loss would be explainable—but the explanation would be about as good in the case of a 0 : 21 end result as it would be in the case of a 0 : 24 end result. (Thanks to Theo Kuipers and Ilkka Niiniluoto for helpful discussion here.)

[7]The referee helpfully provided a proof: Suppose we have worlds $\{w_1, w_2, w_3\}$, where w_2 and w_3 are H-worlds and w_1, the only $\neg H$-world, is actual. Now compare probability assignments p and p^* to these worlds: $p(w_1) = p(w_3) = 0$ and $p(w_2) = 1$; $p^*(w_1) = 0$, $p^*(w_2) = 1 - x$, and $p^*(w_3) = x$. Furthermore, let the distances among the worlds be given simply by their ordering in the set. Then if your current degrees of belief are given by p, you incur a VS score of $\omega_{11} + \omega_{21}$, while if they are given by p^*, your VS score equals $\omega_{11} + \omega_{21}(1 - x)^2 + \omega_{31}x^2$. And $\left(\omega_{11} + \omega_{21}(1-x)^2 + \omega_{31}x^2\right) - \left(\omega_{11} + \omega_{21}\right) = -\omega_{21} + \omega_{21}(1-x)^2 + \omega_{31}x^2$, which is negative for small values of x. Hence, your VS score can go up by becoming certain of the closest H-world, while previously you were only certain that H.

[8]Incidentally, this is not a reason to think the RPS rule is more attractive after all, because that rule fails to satisfy Oddie's Proximity condition as well. To see this, consider a set of worlds $\{w_1, w_2, w_3, w_4\}$, where w_1, w_3, w_4 are H-worlds and w_2 is the only $\neg H$-world and is actual. Again, the distances among the four worlds are given by their order in the set, meaning that w_1 and w_3 are the H-worlds closest to the actual world. Now let p and p^* be such that $p(w_1) = p(w_3) = .5$ and $p(w_2) = p(w_4) = 0$, while $p^*(w_1) = 1$ and $p^*(w_i) = 0$ for $i \in \{2, 3, 4\}$. Suppose your current degrees of belief are given by p. Then if a scoring rule is to satisfy Proximity, it should not make you come out less accurate if you replace those degrees of belief by ones given by p^*. But according to the RPS rule, doing so *would* make you less accurate. Given your current degrees of belief, the rule assigns you a penalty of 1/6. But if you switch to p^*, your RPS penalty doubles, becoming 1/3. Oddie (2019) proves that, given certain mild restrictions on the semantics, any additive scoring rule that satisfies his condition is improper, where a scoring rule is additive if it can be written as the sum of local inaccuracies, which look only at what probability is assigned to a world and whether or not that world is actual. (For a similar result, see Levinstein 2019.) That the RPS rule does not satisfy Proximity, as just shown, might be thought to follow already from Oddie's formal result. That is not so, however.

to Proximity, and Schoenfield (2021) and McCutcheon (2021) have recently argued that the principle is to be rejected for being too strong. I must confess to have no clear intuitions either way. I will proceed on the assumption that McCutcheon and Schoenfield are right, but—as said—consider the issue to be open.[9]

Note that nothing in the foregoing constitutes a general argument against either the log or the Brier score. If truthlikeness relations among the hypotheses at issue are absent, then we may do best to use one of the standard scoring rules. But if such relations *are* present, then, at least intuitively, a scoring rule should take them into account. Hence, scoring is context-sensitive.

A weakness of this argument is that it relies heavily on our intuitions about truthlikeness and its relevance to scoring. Enthusiasm for intuition-based approaches has considerably waned over the past decade or so, as philosophers came to digest findings like those reported in Machery et al. (2004) and Weinberg et al. (2010), which call into question the reliability of philosophers' intuitions. In the following, I aim to go beyond an appeal to intuition by looking at the practical consequences of scoring. In doing so, I am taking on board Murphy's (1993) suggestion that, in assessing and comparing scoring rules, we should also look at the potential economic benefits of using a particular rule. If, by using one scoring rule rather than another, I am generally able to make better business deals, hire better people, figure out more profitable investment strategies, or whatever, then, ceteris paribus, that is good reason to go with the first rule and not with the second. Again, the argument will not be that there is one scoring rule that does better than all the others in this respect. Rather, the claim to be argued for is that, in some contexts, relying on a scoring rule that is sensitive to distance from the truth leads to economically better decisions than, for instance, the Brier or log scores, while in other contexts, the former type of scoring rules offer no such benefit.

3. Scoring and value: Illustrating the connection. Suppose a hospital considers hiring a new engineer who is going to be responsible for the purchase of instruments used in the hospital laboratories. After a first selection, three candidates remain. As a final test, the head of the human resources department, who is going to decide about the hire, wants the candidates to determine as best they can the reliability of a kind of device for measuring certain blood values. All three candidates are given an instrument of this sort. To figure out how reliable it is, they can work with it for a while: make measurements, take the thing apart—do anything that might help them figure out how reliable the instrument is. They are asked to report their estimate in the form of a probability assignment to the members of $\{H_i\}_{i=1}^{11}$, where H_i is the hypothesis that the device has a reliability of $(i-1)/10$.

For although the RPS rule is proper, it is not additive: it is not enough to know of each world whether it is actual and what probability it gets assigned; its place in the ordering, and the probabilities assigned to the other worlds, matter, too.

[9] Schoenfield (2021) proposes a number of principles weaker than Proximity but still strong enough—according to her—to capture truthcloseness intuitions. McCutcheon (2021) argues for a stronger replacement for Proximity that he calls "Proxvexity." The referee who brought to my attention that VS rules fail to satisfy Proximity also pointed out to me that they do satisfy McCutcheon's Proxvexity condition. It is worth noting that McCutcheon (2021) proposes a set of scoring rules that satisfy the same condition but that in addition are proper. However, McCutcheon's rules build on the log score and shares with that the unboundedness problem (Carvalho 2016, p. 226), which some may find serious enough to reject those rules.

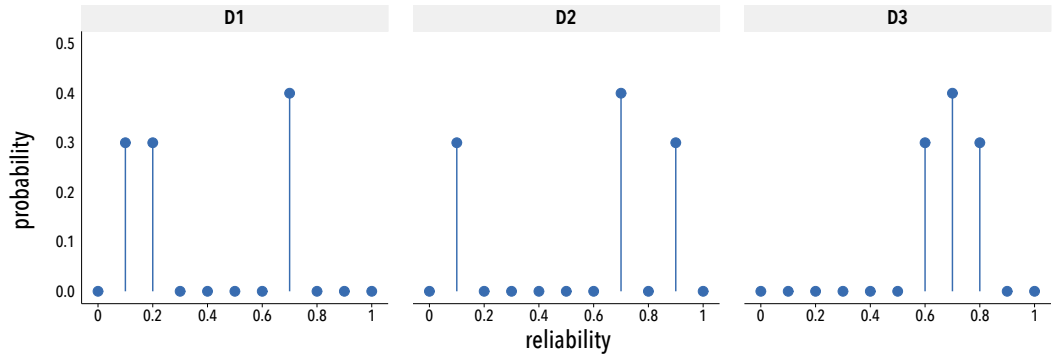

Figure 1: Three probability distributions on $\{H_i\}_{i=1}^{11}$.

The candidates come up with the somewhat different probability assignments given in Figure 1, with D1 the probabilities of the first candidate, and so on. We see that while all three give the same probability of .4 to the hypothesis that the device has a reliability of .7 and also all assign a probability of .3 to two other hypotheses, these other two hypotheses are not the same for the three of them.

The type of device they were given in the test has long been used in the hospital. It is known to the head of human resources that it outputs the correct blood values in 70 percent of the applications, meaning that H_8 is in fact the true hypothesis. So she can score the probability assignments D1–D3. Which rule should she use for that? Does it even matter?

Before turning to these questions, it is to be noted that, in the following, we will, for purposes of illustration, pick one VS rule and stick to it. The particular instance we will assume measures distance from the truth simply as "distance" in the ordering of whichever hypothesis partition is at stake. Specifically, where the truth receives a weight of 0, we first assign the hypothesis or hypotheses whose index is equal to the truth's plus or minus k a weight of k and then normalize the weights to make them sum to 1. That appears to yield a reasonable weighting function for the cases considered in this paper. But for the main line of argument in this paper, the exact form of the weighting function is immaterial.

To see, then, why it may matter which scoring rule the head of human resources uses, assume—plausibly—that how valuable for the hospital a device of the sort at issue is depends on how reliable it is. For concreteness, suppose the value of this type of device is given by the function $r(x) = x^4$, where $x \in [0, 1]$ is the reliability of the device. Then consider Table 1, which gives the Brier, log, ranked probability, and VS scores for D1–D3 on the supposition that H_8 is true. It also shows the absolute difference or divergence (denoted by Δ) between the *actual* value of the device (given its reliability of .7 and supposing that its value is represented by the function r) and the *expected* value conditional on each of D1–D3. For instance, for D1 this divergence equals $|r(.7) - \mathbb{E}_{D1}[r]|$, where $\mathbb{E}_{D1}[r] := \sum_{i=1}^{11} D1(H_i) r\big((i-1)/10\big)$ is the expected value given D1 (with $D1(H_i)$ the probability of H_i in D1).

The truth—H_8, we are assuming—receives the same probability in all three of D1–D3 while the remaining probability is divided equally over two false hypotheses. In D1, however, the false hypotheses that receive positive probability are almost at maximum distance from the truth, while in D3 they are as close to the truth as can be; D2 presents an intermediate

Table 1: Brier, log, ranked probability, and VS scores, as well as Δ-values, for the distributions shown in Figure 1, on the supposition that H_8 is true.

	Brier	log	RPS	VS	Δ
D1	0.049	0.916	0.189	0.029	0.144
D2	0.049	0.916	0.072	0.021	0.053
D3	0.049	0.916	0.018	0.005	0.018

case. It is thus no surprise that the two truthlikeness-sensitive rules penalize D1 more heavily than D2, and D2 more heavily than D3. By contrast, the Brier and log scores, which are *in*sensitive to distance from the truth, penalize all three distributions equally. More surprising, perhaps, is that the RPS and VS rankings correspond to how the Δ-values rank the probability assignments: D1 leads to the greatest absolute difference between real and expected value and D3 to the smallest such difference. This suggests that probability assignments which score better on either the RPS or the VS rule will yield more accurate estimates of the value of a kind of device the hospital may well consider purchasing in the future.

Naturally, it is in the hospital's interest that the person they put in charge of buying measuring devices is able to make as accurate as possible estimates of the value of such equipment. The hospital does not want to overpay—as might well have happened with the device for measuring blood values if the second candidate had been in charge of making the deal—or decide *not* to buy a device because it is deemed too expensive, where in fact it offers good value for the money, as might well have happened with the present device if the deal had been up to the first candidate. Thus, they are hoping to hire someone who will be able to come up with an assessment of the probabilities of the relevant reliability hypotheses which makes the expectation about the value of whichever device they consider buying match that device's real value within reasonable bounds. As the current example suggests, the head of human resources might then do best to use one of the scoring rules sensitive to truthlikeness rather than the log or Brier score (which might earn the third candidate a job offer).

But this is only a suggestion. The example itself gives no reason to hold that ranked probability scores, or VS scores, track divergences between expected and actual value, let alone that a lower ranked probability score, or a lower VS score, *guarantees* a smaller divergence between expected and actual value. The correlation exhibited in Table 1 might be attributable to the fact that we are making a particular assumption about which reliability hypothesis is true and are considering only three probability distributions. Indeed, if we calculate the

Table 2: Brier, log, ranked probability, and VS scores, as well as Δ-values, for the distributions shown in Figure 1, on the supposition that H_7 is true.

	Brier	log	RPS	VS	Δ
D1	0.122	∞	0.169	0.031	0.033
D2	0.122	∞	0.112	0.028	0.163
D3	0.067	1.204	0.058	0.011	0.128

scores and divergences for the same probability assignments, but now assuming that the reliability of the device is .6 and so H_7 is true, we obtain the values shown in Table 2. It now appears that *all* rules, including the two rules sensitive to truthlikeness, penalize most heavily the distribution which in fact gives rise to the smallest divergence.

Thus, we shall have to investigate more systematically whether rankings of probability assignments by score reflect accuracy rankings of value estimates better if the scores come from the RPS or VS rule than if they come from the Brier or log rule. We will do so in the next section, where we report the outcomes of running computer simulations with many different probability distributions, systematically working through the various locations where the truth may be found in the space of possibilities.

4. Scoring and value: Computer simulations. To make the investigation more systematic still, and to avoid the impression that we are capitalizing on what may just be a peculiarity of the value function r from the previous section, we will consider three further, rather "differently-shaped" value functions, all of which have, like r, a natural interpretation.

For the first additional function, suppose we are considering paying someone for advice on stock trading. For any given stock on any given day, the price is about as likely to go up as it is to go down. So, a stock advisor becomes valuable only if she does better than guessing—but then she quickly becomes *very* valuable (though from some point onwards, the increase in value may taper off somewhat, because of the decreasing marginal utility of money). Specifically, we suppose that her value as stock advisor is a function of her success rate, as follows:

$$s(x) \;=\; \begin{cases} 0 & \text{if } 0 \leqslant x \leqslant .5, \\ 8(x - .5)^2 & \text{if } .5 < x \leqslant .75, \\ 1 - 8(x-1)^2 & \text{if } .75 < x \leqslant 1, \end{cases}$$

with $x \in [0, 1]$ indicating that her recommendation is profitable $100 \cdot x$ percent of the time.

The second additional value function concerns a software package for detecting Trojan horses that you consider purchasing for the mainframe computer of your company. The value of this package depends entirely on how likely it is to miss Trojans that have infected the device, and it is specified by the function $t(x) = \big(\cos(\pi x) + 1\big)/2$, with $x \in [0, 1]$ the probability of missing a Trojan.

For our final function, we may imagine that a casino owner is about to acquire a new roulette table. The table should be fair, or balanced, in that it yields equal chances for a ball to land on red and to land on black. A roulette table that is 100 percent fair—that has a bias of .5, as we might say—may be hard to manufacture; even the most carefully crafted roulette tables are likely to have a slight bias one direction or the other. Nevertheless, we can say that roulette tables are most valuable when they have a bias of .5 and that their value quickly diminishes the further the bias deviates from .5. Where $\mathrm{pdf}(x \mid \mu, \sigma)$ is the probability density function of the normal distribution with mean μ and standard deviation σ, we suppose that the value of a roulette table for the casino owner is given by $u(x) = \mathrm{pdf}(x \mid 0.5, 0.075)/\mathrm{pdf}(0.5 \mid 0.5, 0.075)$,

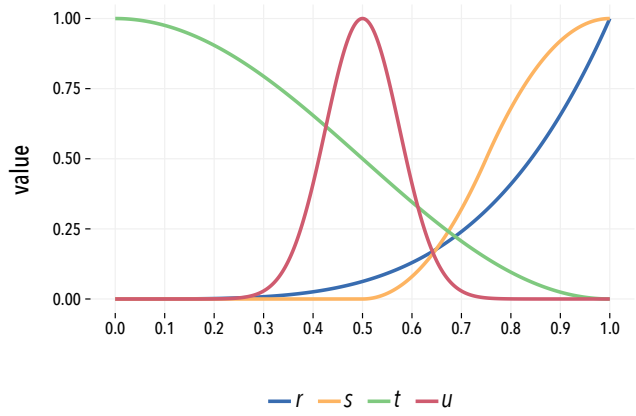

Figure 2: Graphs of value functions.

with x the bias or (as we might say for mnemonic reasons) "unbalance" of the table.[10] To get a better impression of this and the previous value functions, see their graphs in Figure 2.

In our medical device example, we assumed that there were eleven mutually exclusive and jointly exhaustive reliability hypotheses. But when truthlikeness relations are involved, we should reckon with the possibility that scoring rules which are sensitive to such relations may outperform other rules the more strongly, the more finely we can measure distances from the truth. For this reason, we considered in our simulations not only the case where the possibility space was divided evenly into 11 hypotheses but also the cases where the space was divided evenly into 21, 51, and 101 hypotheses, respectively. In other words, we assumed hypothesis partitions $\{H_i^n\}_{i=1}^n$ for $n \in \{11, 21, 51, 101\}$, where H_i^n is the hypothesis that the reliability of the medical device/the success rate of the stock advisor/the probability of missing a Trojan horse/the bias of the roulette table equals $(i-1)/(n-1)$.

In the simulations, we sampled, for each $n \in \{11, 21, 51, 101\}$, 1000 distributions from a uniform Dirichlet distribution[11]; calculated for each of those distributions the Brier score, log score, ranked probability score, and VS score for every possible location of the truth; calculated for each combination of distribution, possible location of the truth, and value function—one of r, s, t, and u—the divergence between actual and expected value; and finally calculated the correlations between the scores and the divergences.

Let $\mathbf{p}^n = (p_1, \ldots, p_n)$ be a probability assignment to $\{H_i^n\}_{i=1}^n$, with p_i being the probability assigned to H_i, for all i. Then, more exactly, the simulations proceeded as follows:

(1) For all $n \in \{11, 21, 51, 101\}$, sample 1000 probability distributions, $\mathbf{p}^{n,1}, \ldots, \mathbf{p}^{n,1000}$, with, for all $i \leqslant 1000$, $\mathbf{p}^{n,i} \sim \mathrm{Dir}(\mathbf{1})$.

[10] We are dividing by pdf(0.5 | 0.5, 0.075) to make 1 the maximum value of the function, so as to make it more easily comparable to the value functions in the other examples. Most probably, to obtain the monetary value of the commodities figuring in our examples, each of r, s, t, and u would have to be multiplied by a different constant. Doing so would be immaterial to the results of the simulations, however, given that these concern correlations, and given that correlations are unaffected by linear transformations of the variables.

[11] To sample a probability distribution on an n-element hypothesis partition from a uniform Dirichlet distribution essentially means that each point in the $(n-1)$-dimensional probability simplex has the same chance of being selected.

Table 3: Coefficients for correlations between Δ-values and scores, averaged over all possible locations of the truth; standard deviations appear in brackets.

		number of hypotheses			
		11	21	51	101
r	Brier	.27 (.17)	.18 (.13)	.11 (.10)	.08 (.07)
	log	.21 (.14)	.15 (.11)	.09 (.08)	.06 (.06)
	RPS	.65 (.33)	.64 (.37)	.64 (.39)	.64 (.40)
	VS	.54 (.23)	.51 (.20)	.50 (.23)	.49 (.23)
s	Brier	.29 (.14)	.20 (.11)	.12 (.08)	.09 (.06)
	log	.23 (.13)	.16 (.09)	.09 (.07)	.07 (.05)
	RPS	.67 (.33)	.66 (.37)	.64 (.39)	.65 (.41)
	VS	.55 (.24)	.52 (.22)	.50 (.25)	.49 (.24)
t	Brier	.31 (.10)	.21 (.08)	.13 (.06)	.09 (.05)
	log	.26 (.09)	.17 (.08)	.10 (.05)	.07 (.05)
	RPS	.89 (.12)	.88 (.16)	.87 (.18)	.88 (.19)
	VS	.67 (.14)	.61 (.17)	.61 (.18)	.58 (.20)
u	Brier	.24 (.19)	.19 (.13)	.11 (.09)	.08 (.07)
	log	.20 (.16)	.14 (.12)	.08 (.07)	.06 (.06)
	RPS	.18 (.26)	.19 (.28)	.20 (.28)	.21 (.27)
	VS	.07 (.29)	.05 (.28)	.06 (.27)	.05 (.25)

(2) For all $n \in \{11, 21, 51, 101\}$, $i \leqslant 1000$, $j \leqslant n$, and $\mathcal{S} \in \{\mathcal{B}, \mathcal{L}, \mathcal{R}, \mathcal{V}\}$, calculate $\mathcal{S}_j(\mathbf{p}^{n,i})$, that is, the score, according to the given rule, on the assumption that H_j is the true hypothesis.

(3) For all $n \in \{11, 21, 51, 101\}$, $i \leqslant 1000$, $j \leqslant n$, and $f \in \{r, s, t, u\}$, calculate $\Delta(n, i, j, f) := |f(j/n) - \mathbb{E}_{\mathbf{p}^{n,i}}[f]|$.

(4) For all $n \in \{11, 21, 51, 101\}$, $j \leqslant n$, $\mathcal{S} \in \{\mathcal{B}, \mathcal{L}, \mathcal{R}, \mathcal{V}\}$, and $f \in \{r, s, t, u\}$, calculate the Pearson product–moment correlation coefficient between $\{\mathcal{S}_j(\mathbf{p}^{n,i})\}_{i=1}^{1000}$ and $\{\Delta(n, i, j, f)\}_{i=1}^{1000}$.

Table 3 displays the outcomes of the simulations, averaged over all values of $j \leqslant n$ (i.e., over all possible locations of the truth).

The most general observation to be made is that in the first three examples—that is, for value functions r, s, and t—there are vast differences in the correlation coefficients obtained for the two truthlikeness-sensitive scoring rules and those obtained for the Brier and log scores, where further the RPS rule tops the VS rule. We see that for those examples there are, on average, and for all values of n, strong correlations between Δ-values and ranked probability scores, moderately strong correlations between Δ-values and VS scores, but only weak (for $n = 11$) to very weak (for all other values of n) correlations between Δ-values and Brier or log scores. The fourth example shows a different picture. Here, we see weak to very

weak correlations for *all* scores, although with increasing n the results improve somewhat for ranked probability scores, get worse for Brier and log scores, and stay at about 0 for the VS rule.[12]

It merits emphasis that while in the initial medical device example, involving only distributions D1–D3, we had a one-to-one correspondence between the *rankings* of the distributions by their Δ-values and those by their ranked probability scores as well as their VS scores, the coefficients in Table 3 show that, for that example as well as for the second and third example, there is actually a close (for the RPS rule) to moderately close (for the VS rule) *linear* relationship between Δ-values and both of those scores. This means that the *extent* to which one ranked probability score is higher than another tightly co-varies with the *extent* to which the Δ-value corresponding to the first score is higher than that corresponding to the second score, and that almost the same conclusion holds for the VS scores.

Practically speaking, the results so far confirm the earlier impression that one should score the candidates' assessments of the reliability of the medical device, in the example from Section 3, via the RPS rule. The recommendation is the same if you head an investment company and some of your staff members have the special task of hiring new stock advisors, or if some of your IT staffers are to decide which anti-malware to buy for your mainframe computer. By contrast, if you are a casino owner and are looking for someone to whom you can delegate the purchase of new roulette tables, the results do not seem to suggest any recommendation as to which scoring rule you should use in the selection process: none of them appears to offer much help in this kind of case.

But there is a twist here. We have been considering *average* correlation coefficients. In reality, however, we will typically have good reason to expect the bias of any given roulette table to deviate only minimally from .5 (that is, from being perfectly fair). Analogous things might hold with respect to medical devices, stock advisors, and anti-malware. We might thus want to look more closely at correlation coefficients for particular hypotheses. For the roulette table example (and only for this example), doing so reveals a remarkable pattern, which is exhibited by the graphs in Figure 3. Whereas the averaged results gave no reason to hope that any of the scoring rules at issue might be useful for selecting whom to rely on to assess the bias of a roulette table, the graphs show that, supposing that the bias of roulette tables tends to be close to .5, all four scoring rules may be useful—as long as we are relying on a relatively coarse-grained hypothesis partition. The top panel of Figure 3 presents the results for the $n = 11$ case, which shows that in the case of fair tables, all four scoring rules yield outputs that correlate moderately strongly to very strongly with the relevant Δ-values. The bottom panel shows the results for the $n = 101$ case, and there we see that, while the results for the RPS rule are still excellent, and those for the VS rule still acceptable, correlations between Δ-values

[12]The procedure described in Meng, Rosenthal, and Rubin (1992) allows one to test for differences among so-called correlated correlations, which are correlations between pairs of variables where one of the variables is shared by the pairs. For instance, we can test whether the correlation of the Brier scores for a given $n \in \{11, 21, 51, 101\}$ with the Δ-values for that n and a given value function differs significantly from the correlation of the ranked probability scores for that n with the same Δ-values. Using this procedure, it was found that, for all n and each of r, s, and t, the correlations for the ranked probability scores as well as for the VS scores are significantly higher (at $\alpha = .0001$) than those for the Brier and log scores, and the correlations for the ranked probability scores are also significantly higher (at the same α level) than the VS scores. In the case of value function u, the differences among the correlations are not significant for the cases $n = 11$ and $n = 21$, but for the remaining cases the correlations for the ranked probability scores are significantly higher (at $\alpha = .001$) than those for the other rules.

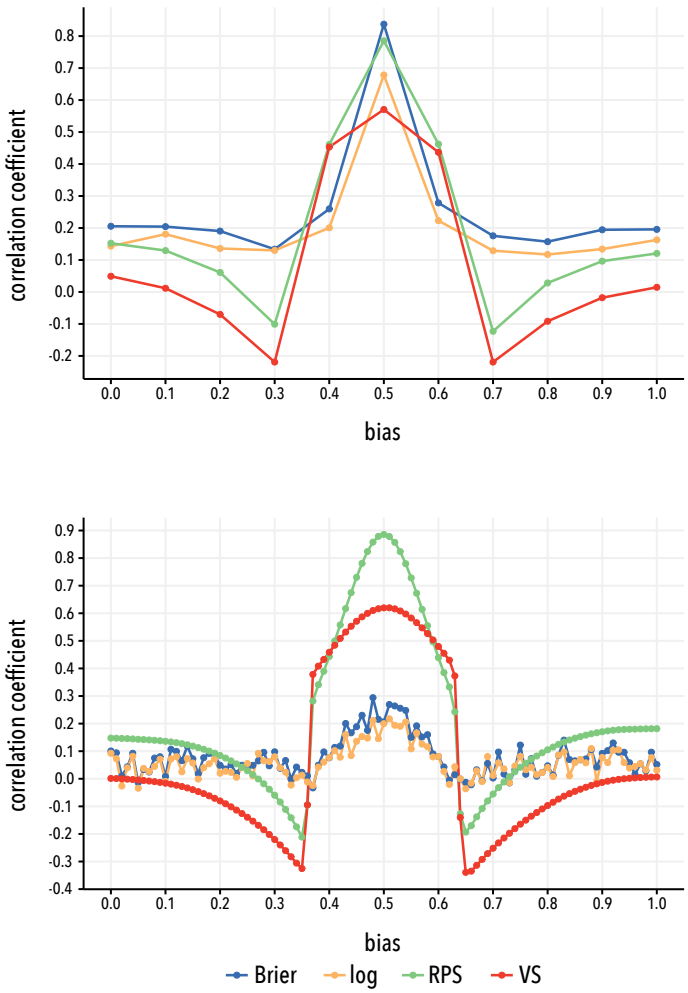

Figure 3: Coefficients for correlations between Δ-values based on value function u and scores for all bias hypotheses, the top panel showing the results for the $n = 11$ case and the bottom panel those for the $n = 101$ case.

and either Brier or log scores become weak to very weak. The results for the intermediate cases $n = 21$ and $n = 51$ (not displayed) show "intermediate" patterns: the correlations for the RPS rule, and to some extent also those for the VS rule, have virtually the same shape when plotted, while those for the other two scoring rules gradually flatten out as we go from $n = 11$ to $n = 101$ via the intermediate steps.

It is worth pointing out why these results are entirely unmysterious and have everything to do with the shape of the value function u (see again Figure 2). Compare two probability assignments to the hypotheses $\{H_i\}_{i=1}^{11}$ in the roulette example, both of which assign a probability of .45 to H_4, with one assigning the same probability to H_3 and the other assigning the same probability to H_5 and both dividing the remaining probability equally over the other hypotheses. Now assume that H_4 is true (meaning that the table has a bias of .3). Then both probability assignments put almost all probability at or close to the truth, the only difference

being that the first puts a lot of probability "to the left" of the truth while the other puts that same probability "to the right" of the truth. Accordingly, they receive very similar ranked probability scores (0.0241 and 0.0280, respectively) and identical VS scores (.006). However, the two assignments lead to very different divergences between expected and actual value (0.005 and 0.185, respectively), due to the fact that while under the supposition that H_3 is true, the table is basically worthless, it is at about half its maximum value under the supposition that H_5 is true. The same holds for the symmetrical situation in which H_8 is true and two probability assignments differ only in that one assigns a lot of probability mass to H_7 while the other assigns the same mass to H_9. That is why we find the lowest correlations for biases .3 and .7, as seen in Figure 3. By contrast, suppose that two probability assignments differ only in that one assigns a lot of probability to H_5 while the other assigns that same probability to H_7. Then if the table is well balanced (so H_6 is true), the divergence between expected and actual value will be the same under the two probability assignments, and the RPS and VS scores will be close to each other or identical.[13]

The Supplementary Materials for this paper contain the code that was used for running the simulations described above. Interested readers will have no difficulty re-using the code for other value functions they might want to experiment with. Also, the code lets readers experiment with value functions that associate random values with the members of a hypothesis partition. As can easily be verified, such functions yield correlations between scores and divergences not significantly different from 0, for all scoring rules. Again, there is no mystery here. What led to the moderately high to high correlations for the truthlikeness-sensitive scoring rules in our simulations was that hypotheses that were similar to each other in terms of truthlikeness were also similar to each other with respect to associated value. In such cases, concentrations of probability mass around the true hypothesis lead *both* to low RPS and VS scores (which reward assigning high probabilities to truthlike hypotheses, all else being equal) *and* to small divergences between expected and true values (given that weighted sums of the values under the various hypotheses will be closer to the value under the true hypothesis the higher the weights given to hypotheses under which the value is similar to the true value, all else being equal). In the case of hypotheses that do not stand in any truthlikeness relations to each other, there is nothing similar that could assure a close connection between scores—given *any* scoring rule—and expected values.

5. Conclusion. Over the past fifty years or so, a welter of scoring rules have been proposed in the literature. Much ink has been spilled over the question of which of those rules is the true measure of inaccuracy. I have previously argued against the implicit assumption underlying that debate, to wit, that there is exactly one true scoring rule. In particular, my claim was that different contexts may call for different scoring rules.

My argument for this claim turned on the observation that, whenever truthlikeness relations are present among the hypotheses of interest, we prefer forecasts that assign higher probabilities to hypotheses close to the truth, all else being equal. Of better known scoring rules, only the RPS rule is able to reflect that preference. The VS rules succeed in this respect

[13] Looking at the shapes of the other value functions, we can also understand the rest of the results reported in Table 3. In particular, it is easy to see why the correlations between ranked probability scores and divergences are particularly high for t, which has a more or less steady slope of about −1 across its entire domain.

as well.[14] That, often enough, we want to score probabilities assigned to hypotheses that do *not* stand in any truthlikeness relation to each other—in which case it makes no sense to use either of the two aforementioned rules—just went to underscore the case for scoring-rule pluralism.

The findings in this paper provide further and independent support for that position. It is not just that, in the kind of cases in which they apply, the RPS and VS rules offer scores that better accord with intuition than do the scores we get from the more standard scoring rules. Using either of these rules, and in particular the RPS rule, is likely to lead to better decision-making. For instance, scoring the candidates in the medical device example by the RPS rule is more likely to result in a hire from which the hospital will benefit economically. Such economic considerations are relevant to comparing scoring rules as well, as we said in the introduction.

The same findings also show that scoring is even more context-sensitive than followed from the earlier work. Not only the presence or absence of truthlikeness relations may matter. In the examples we looked at, the shape of the value function mattered, and in one example even granularity—how finely we partition logical space—mattered.

Finally, it is to be noted that all our examples took some value function to be objectively given. As Murphy (1993, p. 286) points out, however, value functions may not just vary from one context to another, but also from one individual to another. But then, in light of the foregoing, even different individuals in the same context could be advised to use different scoring rules. That conclusion is about as far removed as can be from the scoring-rule monism that dominates the current literature.[15]

References

Bernardo, J. M. (1979) "Expected information as expected utility." *Annals of Statistics* 7:686–690.

Bernardo, J. M. & Smith, A. F. M. (2000) *Bayesian Theory*. New York: Wiley.

Bickel, J. E. (2007) "Some comparisons between quadratic, spherical, and logarithmic scoring rules." *Decision Analysis* 4:49–65.

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

Carvalho, A. (2016) "An overview of applications of proper scoring rules." *Decision Analysis* 13:223–242.

Cooke, R. M. (1991) *Experts in Uncertainty*. Oxford: Oxford University Press.

Douven, I. (2020) "Scoring in context." *Synthese* 197:1565–1580.

Douven, I. (2021) *The Art of Abduction*. Cambridge MA: MIT Press, in press.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017) "Measuring graded membership: The case of color." *Cognitive Science* 41:686–722.

[14] As do the rules proposed in McCutcheon (2021). And there also exist weighted versions of the Brier score that deliver at least some aspects of the said desideratum. See Greaves and Wallace (2006), Dunn (2018), and Schoenfield (2021).

[15] I am greatly indebted to Christopher von Bülow and to two anonymous referees for valuable comments on previous versions.

Dunn, J. (2018) "Accuracy, verisimilitude and scoring rules." *Australasian Journal of Philosophy* 97:151–166.

Epstein, E. S. (1969) "A scoring system for probability forecasts of ranked categories." *Journal of Applied Meteorology* 8:985–987.

Fairchild, M. D. (2013) *Color Appearance Models*. Chichester UK: Wiley.

Fallis, D. (2007) "Attitudes toward epistemic risk and the value of experiments." *Studia Logica* 86:215–246.

Fallis, D. & Lewis, P. J. (2016) "The Brier rule is not a good measure of epistemic utility (and other useful facts about epistemic betterness)." *Australasian Journal of Philosophy* 94:576–590.

Gärdenfors, P. (2000) *Conceptual Spaces*. Cambridge MA: MIT Press.

Good, I. J. (1952) "Rational decisions." *Journal of the Royal Statistical Society* B14:107–114.

Greaves, H. & Wallace, D. (2006) "Justifying conditionalization: Conditionalization maximizes expected epistemic utility." *Mind* 115:607–632.

Joyce, J. M. (1998) "A nonpragmatic vindication of probabilism." *Philosophy of Science* 65:575–603.

Joyce, J. M. (2009) "Accuracy and coherence: Prospects for an alethic epistemology of partial belief." In F. Huber & C. Schmidt-Petri (eds.) *Degrees of Belief* (pp. 263–297). Dordrecht: Springer.

Kuipers, T. A. F. (2000) *From Instrumentalism to Constructive Realism*. Dordrecht: Kluwer.

Kuipers, T. A. F. (2001) *Structures in Science*. Dordrecht: Kluwer.

Kuipers, T. A. F. (2019) *Nomic Truth Approximation Revisited*. Basel: Springer.

Levinstein, B. A. (2017) "A pragmatist's guide to epistemic utility." *Philosophy of Science* 84:613–638.

Levinstein, B. A. (2019) "An objection of varying importance to epistemic utility theory." *Philosophical Studies* 176:2919–2931.

Machery, E., Mallon, R., Nichols, S., & Stich, S. P. (2004) "Semantics, cross-cultural style." *Cognition* 92:B1–12.

McCutcheon, R. G. (2019) "In favor of logarithmic scoring." *Philosophy of Science* 86:286–303.

McCutcheon, R. G. (2021) "A note on verisimilitude and accuracy." *British Journal for the Philosophy of Science*, in press.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992) "Comparing correlated correlation coefficients." *Psychological Bulletin* 111:172–175.

Murphy, A. H. (1969) "On the 'ranked probability score'." *Journal of Applied Meteorology* 8:988–989.

Murphy, A. H. (1993) "What is a good forecast? An essay on the nature of goodness in weather forecasting." *Weather Forecasting* 8:281–293.

Niiniluoto, I. (1984) *Is Science Progressive?* Dordrecht: Reidel.

Niiniluoto, I. (1987) *Truthlikeness*. Dordrecht: Reidel.

Niiniluoto, I. (1998) "Verisimilitude: The third period." *British Journal for the Philosophy of Science* 49:1–29.

Niiniluoto, I. (1999) *Critical Scientific Realism*. Oxford: Oxford University Press.

Oddie, G. (2019) "What accuracy could not be." *British Journal for the Philosophy of Science* 70:551–580.

Roche, W. & Shogenji, T. (2018) "Information and inaccuracy." *British Journal for the Philosophy of Science* 69:577–604.

Rosenkrantz, R. D. (1981) *Foundations and Applications of Inductive Probability*. Atascadero CA: Ridgeview.

Schoenfield, M. (2021) "Accuracy and verisimilitude: The good, the bad and the ugly." *British Journal for the Philosophy of Science*, in press.

Schurz, G. (1987) "A new definition of verisimilitude and its applications." In P. Weingartner & G. Schurz (eds.) *Logic, Philosophy of Science and Epistemology* (pp. 177–184). Vienna: Hölder-Pichler-Tempsky.

Schurz, G. (1991) "Relevant deduction." *Erkenntnis* 35:391–437.

Schurz, G. (2011) "Verisimilitude and belief revision: With a focus on the relevant element account." *Erkenntnis* 75:203–221.

Schurz, G. (2019) *Hume's Problem Solved: The Optimality of Meta-induction*. Cambridge MA: MIT Press.

Selten, R. (1998) "Axiomatic characterization of the quadratic scoring rule." *Experimental Economics* 1:43–62.

Shepard, R. N. (1987) "Toward a universal law for psychological science." *Science* 237:1317–1323.

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). "Are philosophers expert intuiters?" *Philosophical Psychology* 23:331–355.