Scoring in Context*

Igor Douven
SND/CNRS/Sorbonne University
igor.douven@sorbonne-universite.fr

Abstract

A number of authors have recently put forward arguments pro or contra various rules for scoring probability estimates. In doing so, they have skipped over a potentially important consideration in making such assessments, to wit, that the hypotheses whose probabilities are estimated can approximate the truth to different degrees. Once this is recognized, it becomes apparent that the question of how to assess probability estimates depends heavily on context.

Keywords: Bayesian epistemology; probability; scoring rules; truthlikeness; uncertain reasoning.

1. Introduction. Suppose you are giving an exam containing 10 questions, where each question has a straightforward yes/no answer, and each question is to contribute equally to the final score. Then if a student gets 9 questions right, the reasonable grade for this student appears to be 90%, or whatever this translates to in your grading system (e.g., an A in the American public school system, or a 3.6 in a 4.0-based system).

When questions do not have straightforward yes/no answers, grading can be more complicated. A student may get none of the questions completely right and yet show a good understanding of the relevant subject matter. For example, a student may turn in an exam that gives clear evidence of her mathematical acumen, on the one hand, and of a certain inattentiveness when it comes to carrying out simple algebraic operations, on the other. Most mathematics teachers would not want to fail such a student, but would rather encourage her to address her sloppiness. What mark to give in such a case is not so easy to say, however, and may be a decision that is to some degree subjective.

Grading can be complicated as well when we are dealing with questions that ask for a probability estimate of some future event. Such questions may not occur so frequently on a school exam (though see Bickel 2010), but they are the bread and butter for a variety of professionals, ranging from financial analyst to physician, from football coach to engineer, and from foreign policy advisor to weather forecaster. Suppose one weather forecaster predicts rain for tomorrow with a probability of .25, while a second predicts rain for tomorrow with a probability of .75. If tomorrow stays dry, neither forecaster can be said to have been wrong, for neither predicted rain with certainty. Yet from a

*This paper is dedicated to Gerhard Schurz, on the occasion of his 60th birthday.

pre-theoretic perspective, it appears that the former forecaster did a better job than the latter; that forecaster was—we would like to say—the more accurate of the two.

How to turn this intuition into a difference in grades (or scores) for the two forecasters—supposing our only concern is with their predictions on the issue of rain tomorrow—might also be deemed a partly subjective matter. But researchers from various quarters think otherwise and have sought to answer the foregoing question by providing fully objective formal rules. More specifically, proposals for how to assess the forecasters, and how to assess the quality of probability estimates generally, have come in the form of so-called scoring rules. The first of these date back to the 1950s (Brier 1950; Good 1952; McCarthy 1956), but by now a bewildering variety of scoring rules exists (Rosenkrantz 1981, Ch. 2; Cooke 1991, Chs. 8 and 9). Given that these rules may even lead to different qualitative verdicts (e.g., on which weather forecaster did best), and given that, as said, this kind of scoring is supposed to be objective, the further question has arisen of which of the many scoring rules available today we are to rely on in practice.

Theorists have almost invariably tried to answer this question by arguing that this or that scoring rule uniquely satisfies certain standards of "goodness" (Winkler and Murphy 1968). There is debate about this issue because there is no generally shared conception of goodness in the relevant sense and hence no agreement on which criteria to impose on scoring rules. Nonetheless, it is fair to say that most theorists are, or have been, in favor of one of two scoring rules, namely, either the Brier rule[1] (e.g., Rosenkrantz 1981; Joyce 1998; Selten 1998; Greaves and Wallace 2006; Leitgeb and Pettigrew 2010) or the logarithmic (or "log") rule (e.g., Good 1952; Bernardo 1979; Bernardo and Smith 2000; Bickel 2007, 2010; Levinstein 2012).

In this essay I argue that in the debate about scoring rules, not enough attention has been given to the fact that probability estimates may concern hypotheses that differ in their distance from the truth. In scoring, disregarding truthlikeness relations among the hypotheses of interest (when present) can lead to pre-theoretically unsatisfactory scores. Once this is recognized, however, it becomes apparent that the question of how to assess probability estimates depends heavily on context.

Section 2 states the aforementioned scoring rules in formal detail, shows their insensitivity to truthlikeness relations, and also presents a family of scoring rules that *are* thus sensitive. Section 3 addresses the concern that, even if truthlikeness relations may matter to scoring pre-theoretically, scoring rules that attend to such relations may come at a prohibitively high cost.

2. Standard scoring rules and truthlikeness. Let $\{H_i\}_{i=1}^n$ be a *hypothesis partition*, that is, a set of mutually exclusive and collectively exhaustive hypotheses, and let δ_{ij} be the Kronecker delta, which equals 1 if $i = j$ and 0 otherwise. Suppose a person assigns probabilities $\mathbf{p} = (p_1, \ldots, p_n)$ to the elements of $\{H_i\}_{i=1}^n$, with p_i her probability that H_i is true. Then, assuming that H_j is in fact true, the Brier score for this person equals $\mathscr{B}_j(\mathbf{p}) \coloneqq 1/n \sum_{i=1}^n (\delta_{ij} - p_i)^2$, whereas her log score equals $\mathscr{L}_j(\mathbf{p}) \coloneqq -\ln(p_j)$. It is customary to conceive of scoring rules as assigning *penalties*, so that lower scores are better.

[1] Or the quadratic scoring rule, which is a generalization of the Brier score; see below.

But this is a convention: if we wish, we can take the negative of any given scoring rule and think of its scores as *rewards*.

Given a scoring rule $\mathcal{S}$ and a probability distribution $\mathbf{p}$ on a hypothesis partition, the $\mathbf{p}$-*expectation* of the $\mathcal{S}$-score (or $\mathbf{p}$-expected $\mathcal{S}$-score, for short) for a second probability distribution $\mathbf{p}^*$ on the same partition equals $\mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] \coloneqq \sum_{i=1}^{n} p_i \mathcal{S}_i(\mathbf{p}^*)$. A scoring rule $\mathcal{S}$ is said to be *proper* iff, for all $\mathbf{p}$, $\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathcal{S}(\mathbf{p}^*)] = \mathbf{p}$, and *strictly proper* iff each of these minima is unique. Scoring rules were originally proposed for eliciting probabilities (Cooke 1991, p. 121), and when used for that purpose, they should not give a person an incentive to announce probabilities which she does not actually hold. This is why many theorists regard strict propriety as an important requirement for scoring rules.

Both the Brier and the log score are known to be strictly proper. They also both achieve their minimum of 0 when probability 1 is assigned to the truth, which is generally regarded as another important desideratum. Nevertheless, the following example brings out that the rules can lead to very different verdicts.

Suppose there will soon be an election for a new president of your university. There are three candidates in the running: Ashley, Bertrand, and Charlotte. Your colleagues David, Emma, and Frank hold different views on which of the candidates is most likely to become the new president. Specifically, their relevant probabilities are as given in Table 1. Suppose Charlotte wins the election. Then David and Emma have the same Brier score: $(.1^2+.5^2+.6^2)/3 \approx .21$; Frank's Brier score is lower and thus better: $(.3^2+.3^2+.6^2)/3 = .18$. On the other hand, all of them have the same log score, to wit, $-\ln(.4) \approx 0.92$.

That David and Emma do equally well on either scoring rule seems as it should be: David and Emma both assign a probability of .1 to one of the false hypotheses and a probability of .5 to the other, and it is difficult to see how it might matter that they do not assign these probabilities to the same hypotheses, at least given the information provided here.

That Frank's Brier score is lower than David's and Emma's illustrates the general fact that, for any hypothesis partition, given a particular probability assigned to the true hypothesis, one minimizes one's Brier score by assigning equal probabilities to the remaining hypotheses. How reasonable is this?

According to advocates of the log rule, this is not reasonable at all. Why should we care—they ask—what probabilities a person assigns to any hypothesis other than the truth? In their view, given that David's, Emma's, and Frank's probabilities for C are the same, they should be assigned the same score (Winkler 1969; Bernardo and Smith 2000,

Table 1: Probability assignments to hypotheses A, B, and C.

	David	Emma	Frank
Ashley wins (A)	.1	.5	.3
Bertrand wins (B)	.5	.1	.3
Charlotte wins (C)	.4	.4	.4

p. 72; Bickel 2010, pp. 347–348). The log rule is in fact known to be the only strictly proper scoring rule that guarantees this outcome (McCarthy 1956).

In the debate about scoring, various other intuitions in favor of or against either of the above rules have been called upon. For instance, Bickel (2010, p. 348) notes that the log rule, but not the Brier rule, ensures that *higher* probability assignments to the truth will result in *lower* penalties.[2] He regards this as compelling reason to prefer the log rule over the Brier rule. By contrast, Selten (1998, pp. 49–50) prefers the Brier rule, because he thinks that in some situations the log rule is more sensitive to small differences between probability assignments than is warranted by intuition, and in other situations not sensitive enough.[3]

That different authors have given different weights to intuitions regarding scoring has partly to do with the fact that they have focused on different examples, and the problem is that the features of the examples which have been used to make one or the other rule look appealing are not always generalizable.

To see how different examples may steer our intuitions in different directions, note that in our own example there is no sense in which either of the false hypotheses (whether A or B) is closer to the truth than the other, again given the information provided. That does not make this hypothesis partition special. But the designated feature is also not altogether general, and by assuming otherwise one might be implicitly favoring some scoring rules over others. For suppose David, Emma, and Frank were wondering how well a student of theirs is going to do on an exam, and the probabilities given in Table 1 are the probabilities that the student will receive an A, a B, or a C, possibilities we may take to be described by hypotheses A, B, and C, respectively. If B turns out true, then the other hypotheses would seem equally far from the truth. But if either A or C turns out true, then, although false, B would be closer to the truth than whichever is the other false hypothesis. Suppose C is true indeed. Would we still want to agree with the Brier score that David and Emma do equally well, given their probabilities for the hypotheses at issue, or even agree with the log score that all three colleagues do equally well? While the three colleagues assign the same probability to the truth, we are tempted to say that David still is closer to the mark, given that he assigns a higher probability than the others to the false hypothesis that is closest to the truth (viz., B) and a lower probability than the others to the false hypothesis that is most distant from the truth (A).

An entire program in the philosophy of science is devoted to making the notion of truthlikeness (or "verisimilitude") formally precise, and by now numerous measures of

[2] That the Brier rule cannot guarantee this is a direct consequence of the general fact mentioned two paragraphs back.

[3] Selten instead prefers the Brier rule, mainly because, as he proves, it is the only scoring rule (up to positive linear transformations) that satisfies each of what he considers to be four important desiderata for such rules, which Selten presents as axioms. According to the first axiom, the ordering of the hypotheses should not influence the score. According to the second, the score should not be affected by the introduction of an additional hypothesis that receives zero probability. The third axiom is the requirement of strict propriety. The fourth axiom, finally, concerns a type of situation that we do not consider in this essay, namely, when a probability assignment is scored in light of another probability assignment rather than in light of the truth of one hypothesis; the axiom requires that, in this situation, the score should be the same regardless of which probability assignment is considered to be the "true" one.

truthlikeness exist; see, for instance, Schurz (1987, 1991, 2011, 2014), Niiniluoto (1998, 1999), Kuipers (2000, 2001, 2014), and Cevolani, Festa, and Kuipers (2013). Here, we will not commit ourselves to any particular such measure and just note that all measures of truthlikeness currently advocated in the literature will do for the purposes of this paper.

The idea that truthlikeness may matter to scoring has been mentioned in the literature, but mostly only to be set aside as a problem that can easily be dealt with by using the so-called quadratic scoring rule (also known as "weighted least squares metric"). If H_j is the true element of hypothesis partition $H = \{H_i\}_{i=1}^n$, this rule assigns a penalty of $\mathcal{Q}_j(\mathbf{p}) \coloneqq \sum_{i=1}^n w_i(\delta_{ij} - p_i)^2$ to someone whose probabilities for the elements of H are given by $\mathbf{p}$. The only general constraints on the weights w_i are that $w_i > 0$ for all i and that $\sum_{i=1}^n w_i = 1$, so that we actually have a *schema* here, with the Brier score as the special case where all hypotheses are weighted equally.

For instance, Rosenkrantz (1981, Ch. 2, p. 1) calls the above-mentioned property of the Brier score to penalize more heavily when the probabilities assigned to the false hypotheses are unequal (ceteris paribus) "attractive when all false alternatives are regarded as 'equidistant' from the truth." He further states, however, that "[w]here false answers are not equally far from the truth and we wish to weight them differently, we can use the weighted least squares metric . . . ." In the same vein, Greaves and Wallace (2006, p. 628) claim that the quadratic scoring rule "can take account of the value of verisimilitude . . . by a judicious choice of the [weights]"; specifically, their proposal is to assign weight to a proposition depending on the extent to which it represents "a set of 'close' states" (*ibid.*).

As it stands, however, this proposal will not work. Which set or sets of states are close depends on which state is *actual*; equivalently, how far from the truth a false alternative is depends on which hypothesis is *true*. And the weights of the quadratic scoring rule lack this dependence. To be sure, if we know which hypothesis is true, the dependence can be built in by hand. For instance, given that (we said) the student will receive a C, so that A is more distant from the truth than B, we can weight A more heavily than B. That might give the desired result, and probably this is the kind of use of the quadratic scoring rule that the aforementioned authors had in mind.

But which weights are we to assign when we want to use the rule *not* knowing the truth, as when we would like to calculate our *expected* score? To calculate expected scores, we consider all possibilities of truth, calculate for each individual possibility the penalty we would incur were that possibility to be actual, and then take a weighted average of those penalties, the weights being our probabilities for the possibilities. When using the quadratic scoring rule, however, there is a second set of weights involved. And the problem is that, while truthlikeness relations shift from one possibility to the next—for instance, in the possibility in which the student receives a C, hypothesis C is closer to the truth than hypothesis A, but in the possibility in which the student receives a B, hypotheses A and C are equally distant from the truth—the weights attributed by the quadratic scoring rule stay the same whichever possibility is considered. Consequently, a set of weights that adequately reflects truthlikeness relations under one supposition of where the truth lies may well fail to do so under another such supposition.

To avoid this problem, we may adapt the quadratic scoring rule by *doubly* instead of *singly* indexing each weight, where the additional index then refers to the true hypothesis, thereby obtaining what we shall call a *verisimilitude-sensitive scoring rule*, or VS rule for short. This rule imposes a penalty of $\mathscr{V}_j(\mathbf{p}) \coloneqq \sum_{i=1}^{n} w_{ij}(\delta_{ij} - p_i)^2$ on someone assigning probabilities $\mathbf{p}$ to the elements of hypothesis partition H, where H_j is the true element of that partition, and where w_{ij} is the distance from H_i to the truth. (Again, this is really a schema, yielding different rules for different weighting functions.)

To make the difference between the rules vivid, suppose that David, not knowing which grade his student will get, wishes to calculate his expected score, given his probabilities $\mathbf{d} = (.1, .5, .4)$ for hypotheses A, B, and C. First assume that he uses an instance of the quadratic scoring rule. Without loss of generality, let $w_{\mathrm{A}} = .1$, $w_{\mathrm{B}} = .3$, and $w_{\mathrm{C}} = .6$. David will then find that[4]

$$\begin{aligned}\mathbb{E}_{\mathbf{d}}[\mathscr{Q}(\mathbf{d})] \;=\; & .1\big((.1)(.9^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .5\big((.1)(.1^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .4\big((.1)(.1^2) + (.3)(.5^2) + (.6)(.6^2)\big) \;=\; 0.228.\end{aligned}$$

Now suppose that David instead assumes a VS rule, for instance, one that assigns a weight of .1 to the truth and that weights the other hypotheses proportionally to their distance from the truth. Suppose also that, in the present case, distances are given by the ordering of the hypotheses; so if A is true, then C is twice as far from the truth as B, and so on. Then David's expected score is

$$\begin{aligned}\mathbb{E}_{\mathbf{d}}[\mathscr{V}(\mathbf{d})] \;=\; & .1\big((.1)(.9^2) + (.3)(.5^2) + (.6)(.4^2)\big) \\ & + .5\big((.45)(.1^2) + (.1)(.5^2) + (.45)(.4^2)\big) \\ & + .4\big((.6)(.1^2) + (.3)(.5^2) + (.1)(.6^2)\big) \;\approx\; 0.123.\end{aligned}$$

We see how the doubly indexed truthlikeness weights of the VS rule vary per possibility—as they must do, for the reason previously mentioned—while the singly indexed weights of the quadratic scoring rule stick to their propositions across all three possibilities.

3. Truthlikeness and impropriety. While VS rules offer a seemingly straightforward way to take account of truthlikeness relations, these rules face what may appear to be a devastating objection. We can start to bring out the problem by considering that, relative to his current probabilities, David minimizes his expected quadratic score by having those very probabilities; formally, $\arg\min_{\mathbf{p}} \mathbb{E}_{\mathbf{d}}[\mathscr{Q}(\mathbf{p})] = \mathbf{d}$. This is no coincidence. It is known that not only the Brier score but *all* instances of the quadratic scoring rule schema are proper, and even strictly proper (Rosenkrantz 1981, Ch. 2). On the other hand, relative to David's probabilities, 0.123 is *not* the minimum VS score he can incur,

[4] To be entirely precise, instances of the quadratic scoring rule and the VS rule would have to be embellished with a super- or subscript to indicate the weighting function that is being assumed. We will not be so fussy, however.

for minimizing the function

$$.1\big(.1(1-x)^2+.3y^2+.6z^2\big)+.5\big(.45x^2+.1(1-y)^2+.45z^2\big)+.4\big(.6x^2+.3y^2+.1(1-z)^2\big),$$

subject to $x+y+z=1$, yields 0.118, where this minimum is reached at (.146, .548, .306). Hence the VS rule David assumes is improper. This might not be a matter of great concern; perhaps David just made an unfortunate choice of truthlikeness weights. But the problem is more fundamental.

The intuition behind VS rules is that, although it is bad in general to assign a positive probability to a false hypothesis, it is worse the further the hypothesis is from the truth. So, say that a VS rule's weights *reflect truthlikeness in a minimally adequate sense* iff hypotheses are weighted as a function of their distance from the truth, with hypotheses further from the truth being weighted more heavily than hypotheses closer to the truth. Then we have:

Theorem 1 *Every VS rule whose weights reflect truthlikeness in a minimally adequate sense is improper.*

(For a proof, see Appendix A.) In other words, it is not just that David was unlucky in the weighting function he picked; he could have picked none that would not have led to the same problem of impropriety, or at least none that is minimally adequate.

How damning is the above result for the class of VS rules? While the mainstream holds that impropriety is *totally* damning, I would like to argue that the answer should be: it all depends on the purpose of our scoring.

As mentioned, scoring rules were initially meant for *eliciting* probabilities. To serve that purpose, they better not encourage the subject whose probabilities one would like to elicit to lie about those probabilities. But if we are assessed by means of an improper scoring rule, then by revealing our actual probabilities we may expect to incur a larger penalty than if we pretend to have different probabilities, as was seen in the case of David. That can make it disadvantageous to be truthful.

It is by now generally recognized that scoring rules may also be used for the purpose of *self-assessment*. If David adheres to the VS rule considered above, and he wonders about the accuracy of his current probabilities, he may conclude that he should replace those probabilities with those that were found to minimize his expected penalty. That appears problematic; as Moss (2011, p. 1057) notes, it implies that self-assessment by means of an improper scoring rule "could motivate you to raise or lower your credences *ex nihilo*, in the absence of any new evidence whatsoever."

In response, it could be argued that finding out that our probabilities do not minimize our expected penalty is itself new information, so that when we thereupon adapt those probabilities, we are *not* acting in the absence of new evidence. This would in effect be an instance of what Lombrozo (2017) calls "learning by thinking," occasioned by what she calls an "observation generated inside the head." But some might want to reply that the evidence should bear, in an intuitively clear sense, on the hypotheses to which our probabilities are assigned. The intuitively clear sense may be difficult to pin down

formally, but let that pass.[5] For using a VS rule, or any other improper scoring rule, for self-assessment seems to face a threat more severe than the one Moss points toward.

Above we found that, given David's probabilities in Table 1, he would actually minimize his expected penalty by raising his probability for A to .146, raising his probability for B to .548, and lowering his probability for C to .306. Ironically, if he shifts his probabilities accordingly, he will find that he actually minimizes his expected penalty by setting his probabilities for A, B, and C to .164, .593, and .243, respectively. It does not end there, for relative to *these* probabilities, David minimizes his expected penalty by setting his probabilities for the hypotheses to .166, .630, and .204. And it goes on. And on? That would seem to prevent David from ever having stable probabilities for A, B, and C, unless he decides to arbitrarily stop self-assessing at some point.

However, the iterative minimization process, should David engage in it, turns out to reach a fixed point. To be precise, David would, after 442 steps, arrive at probability assignment $\mathbf{d}^* = (.083, .833, .083)$,[6] a probability assignment that is "strongly self-recommending" (Greaves and Wallace 2006, p. 619) in that $\arg\min_{\mathbf{p}} \mathbb{E}_{\mathbf{d}^*}[\mathcal{Q}(\mathbf{p})] = \mathbf{d}^*$. (For completeness, we note that David's expected penalty at the fixed point equals 0.057, whereas for his initial probabilities it was 0.118, as we saw.)

We get a first sense of what may be the real problem here if we repeat our calculations for Emma and Frank and find that they arrive at exactly the same fixed point as David! Not only that; varying the weighting function a little, we find that the three colleagues now arrive at a different fixed point, but that this fixed point is again the same for the three of them. We can illustrate this graphically by noting that vectors in the standard unit simplex (or probability simplex) of dimensionality $n-1$ can be interpreted as probability distributions on an n-element hypothesis partition, with the i-th vector component representing the probability of the i-th hypothesis (de Finetti 1962). Figure 1 shows three two-dimensional simplexes, in each of which the initial probability assignments of David, Emma, and Frank as given in Table 1 are represented by medium-sized dots. The smaller dots represent their consecutive probability assignments while they go through the VS-rule-governed minimization process, the differences among the simplexes being only that the VS rules that are assumed in that process assign different truthlikeness weights: the left simplex represents the process that relies on the VS rule that was assumed in the example above; the processes shown in the other two simplexes used VS rules with slightly different weighting functions (the details are unimportant here).

The finding is again no coincidence but follows from

[5] In this connection, it is also worth mentioning that, at least according to some influential Bayesian statisticians (Gelman and Hill 2007; Gelman and Shalizi 2012, 2013; Kruschke 2013), raising or lowering probabilities in the absence of the kind of evidence with "direct bearing" is accepted as legitimate practice, most notably, as resulting from a so-called posterior predictive check in which a statistical model may be rejected because it is found unsatisfactory (according to informal criteria) in light of simulated data. If rejected, the model is to be replaced by a new one, which requires, among other things, a specification of new prior probabilities. The simulated data that can motivate this kind of model revision—including probability revision—is presumably not the kind of new evidence that Moss has in mind.

[6] This result was obtained by means of the FixedPointList function from *Mathematica* and therefore holds only up to machine precision. However, that the process would reach a fixed point (even if perhaps not after 442 steps) is guaranteed by Theorem 2, to be stated shortly.

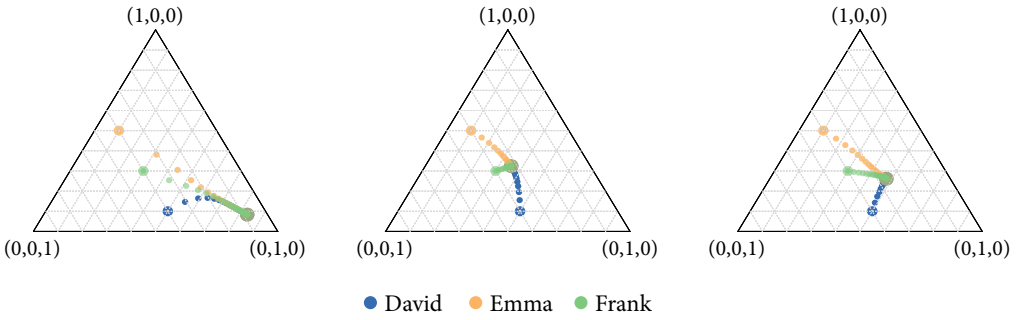

Figure 1: Three probability simplexes showing that David, Emma, and Frank all reach the same fixed points (large dots) by iteratively adapting their probabilities on the basis of VS score minimization, and illustrating the fact that these points depend on the weights assumed by the given VS rule. (Initial assignments are marked by medium-sized dots, intermediate assignments by small dots.)

Theorem 2 *Let S be the standard unit $(n-1)$-simplex, let $\mathbf{p}$ and $\mathbf{p}^*$ range over vectors in S, and let $m\colon S \to S$ be defined as follows:*

$$m(\mathbf{p}) \;\coloneqq\; \arg\min_{\mathbf{p}^*} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i w_{ij} (\delta_{ij} - p_j^*)^2,$$

with δ_{ij} the Kronecker delta, and with $w_{ij} > 0$ for all i, j, and $\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} = 1$. Then there is a $\mathbf{p}^+ \in S$ such that (i) $m(\mathbf{p}^+) = \mathbf{p}^+$, (ii) $\mathbf{p}^+$ is unique, and (iii) $\mathbf{p}^+$ depends only on the w_{ij}.

(See Appendix B for a proof.) To see how this theorem bears on the issue at hand, it suffices to observe that, for any VS rule $\mathscr{V}$,

$$\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] \;=\; \arg\min_{\mathbf{p}^*} \sum_{i=1}^{n} \sum_{j=1}^{n} p_i w_{ij} (\delta_{ij} - p_j^*)^2,$$

with the w_{ij} being provided by the weighting function of the given rule. In other words, the function by which David, Emma, and Frank iteratively update their probabilities in the represented processes is an instance of m. That, as a result, their probability assignments reach a fixed point follows from clause (i) of the theorem, and that, despite their different initial probability assignments, this fixed point is the same for the three of them, provided they use the same VS rule, follows from clauses (ii) and (iii).

Surely we have hit upon an absurd result here, one which offers a compelling reason to refrain from self-assessment by means of a VS rule. As Winkler (1996) notes, however, scoring rules not only have a (what he calls) ex ante use; they can also be used ex post, for evaluating performance. Potential problems with elicitation or self-assessment

are beside the point if one is going to use an improper scoring rule for assessing people's performance without disclosing that rule to them. And often there will be no requirement for disclosure. An investment company might hope to identify job applicants with special capacities to predict the stock market by scoring their answers to test questions via some custom-built scoring rule without informing the applicants about the rule being used. Similarly, if a television network wants to hire a new weather forecaster and is now retroactively analyzing, by means of an improper scoring rule, the performance of various candidates being considered for the job, it will make no difference whether or not the network makes publicly known how it is conducting the analysis.[7]

In short, depending on the purpose of our scoring, improper scoring rules may be admissible. But for those who remain concerned about impropriety, it will be good to know that verisimilitude-sensitivity *need* not come at the expense of propriety, for there is a scoring rule that is both verisimilitude-sensitive *and* proper. This is the so-called ranked probability score (RPS), first proposed in Epstein (1969) and shown to be strictly proper in Murphy (1969). This rule has received hardly any attention from philosophers,[8] nor is it widely discussed outside philosophy (O'Hagan et al. 2006, p. 169). Rather than comparing a probability distribution $\mathbf{p}$ on a partition of hypotheses with the vector $\mathbf{v}$ of truth values of those hypotheses, it compares the cumulative distribution function of $\mathbf{p}$ with the cumulative distribution function of $\mathbf{v}$. Given a partition $\{H_i\}_{i=1}^n$ and a probability distribution $\mathbf{p}$ on this partition, and supposing H_j to be true, the ranked probability score associated with $\mathbf{p}$ is defined as

$$\mathscr{R}_j(\mathbf{p}) \;\coloneqq\; \frac{\sum_{k=1}^{n}\left(\sum_{i=1}^{k} p_i - \gamma_{kj}\right)^2}{n-1},$$

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

To illustrate, given David's probabilities $\mathbf{d}$ and still supposing C to be the true hypoth-

[7] A real-life example of this kind of usage is found in recent work on forecasting carried out by a group of psychologists from various American universities (Mellers et al. 2015; Tetlock and Gardner 2015). These researchers have organized, over a period of several years, a number of prediction tournaments, mostly concerning geopolitical questions. They found that some otherwise ordinary people were much more accurate forecasters than even professional intelligence analysts. A key objective of the research was to determine what distinguishes the most accurate forecasters from the rest of the population. The researchers used a number of different scoring rules for evaluating their participants' performance, including the Brier score but also the so-called AUROC, which is known to be an improper scoring rule (see, e.g., Agresti 2007, Ch. 5, or Hastie, Tibshirani, and Friedman 2009, Ch. 9, for details). Given that the participants were never told what the evaluation process consisted of, the use of an improper scoring rule in that process will not have affected their responses. (Note that, although in this research both proper and improper scoring rules were used for the purposes of selection, one could also use an improper scoring rule to select participants while at the same time scoring them via a proper scoring rule to determine their compensation in the experiment. Letting participants know how they will be compensated will then encourage them to post their true probabilities, while the improper scoring rule—the use of which is *not* disclosed to the participants—may still yield more useful information.)

[8] In fact, to the best of my knowledge, Konek (2016) contains the only reference to the rule (actually, the continuous version of the RPS rule) in the entire philosophical literature.

esis (whence $j = 3$), David's ranked probability score is

$$\mathscr{R}_3(\mathbf{d}) \;=\; \frac{(.1-0)^2 + (.1+.5-0)^2 + (.1+.5+.4-1)^2}{2} \;=\; 0.185.$$

In the same way, we find that Emma's ranked probability score equals 0.305, and Frank's, 0.225.[9] In the example where the three colleagues' probabilities are about what grade their student will receive, these outcomes make perfect sense: as we said, David appears to do best in this case, because he assigns a higher probability to the false hypothesis that is closer to the truth than to the one further from the truth. By the same token, Emma would seem to do worst, given that she does exactly the opposite. Frank steers a sort of middle course between his colleagues in assigning the two false hypotheses equal probability.

The VS rules helped to make some points about impropriety. However, this family of rules has some interest even in the presence of the RPS rule, which is sensitive to truthlikeness while also being proper. The independent interest derives from the fact that VS rules are flexible in a way that the RPS rule is not.[10] Authors concerned with truthlikeness have advanced a number of different ways of measuring the distance from the truth, and more generally the distance between hypotheses, and there is disagreement about which of those is most reasonable, and even about whether there is a unique best such measure (see, e.g., Niiniluoto 1984, Ch. 7). While VS rules can account for different ways of measuring truthlikeness—by allowing users to choose different weighting functions—the RPS rule can *not*: given this rule, relations of truthlikeness are completely fixed by the ordering of hypotheses in the hypothesis partition.[11]

4. Conclusion. Our discussion has pointed toward ways in which questions concerning scoring can be context-dependent. We saw that scoring may serve different purposes in different contexts, and depending on the purpose, we may want to impose different

[9] Because, as noted, the RPS rule is strictly proper, it satisfies Selten's third axiom (see note 3). To see that it also satisfies his fourth axiom, note that, for comparing a probability assignment $(p_1, \ldots, p_n)$ with a "true" probability distribution $(p_1^*, \ldots, p_n^*)$, the RPS rule takes this form:

$$\frac{(p_1 - p_1^*)^2 + ((p_1+p_2) - (p_1^*+p_2^*))^2 + \cdots + ((p_1+\cdots+p_n) - (p_1^*+\cdots+p_n^*))^2}{n-1}.$$

The symmetry required by the fourth axiom then follows from the fact that the addends in the numerator are all squared. Furthermore, the fact that David's and Emma's rank probability scores are different, as seen in the main text, is enough to show that the rule does *not* satisfy Selten's first axiom. Finally, to show that neither does it satisfy the second axiom, we can add to the partition consisting of hypotheses A, B, and C the hypothesis that the student will receive a C–, where this has zero probability for David. Keeping his probabilities for A, B, and C as they were, David's rank probability score then becomes (approximately) 0.243, and hence the addition of the zero-probability alternative did affect the score.

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

[11] It might be said that the VS rule used in this section does not do quite as well with respect to the grading example as the RPS rule. Although David does better than Emma—David having a score of 0.117, and Emma, of 0.189—he incurs the same penalty as Frank. However, this result depends on the particular weights we chose for the example. It is easy to choose weights which could still be said to reflect truthlikeness relations but which would lead to qualitatively the same result as the RPS rule.

requirements on scoring rules. By itself, this does nothing to undermine the idea of there being one true scoring rule. Even if the goal of our scoring is selection, so that we may have no reason to require propriety, it might still be the case that some proper scoring rule is most conducive to our goal. However, our findings also suggest that, in contexts in which the hypotheses at issue can be said to be closer to or further from the truth, the RPS rule, and possibly also some rules from the VS family, not only make sense but yield the more reasonable results, while if relations of truthlikeness are absent, those rules are best avoided. The idea of a unique scoring rule appropriate in any context—an idea shared by virtually all authors who have considered the matter of scoring[12]—is thence called into question.[13]

Appendix A

Recall that the weights of a VS rule are all positive and add up to 1, and that they are said to reflect truthlikeness in a minimally adequate sense iff hypotheses are assigned weights as a function of their distance from the truth, with hypotheses farther from the truth being assigned larger weights than hypotheses closer to the truth.

Theorem 1 *Every VS rule whose weights reflect truthlikeness in a minimally adequate sense is improper.*

Proof: Without loss of generality, consider a hypothesis partition of three hypotheses, H_1, H_2, and H_3. Then, where $\mathscr{V}$ is some VS rule and $\mathbf{p} = (p_1, p_2, p_3)$ is a given person's probability assignment to the aforementioned hypotheses, with p_i the probability assigned to H_i, this person's expected $\mathscr{V}$-score for a probability assignment $\mathbf{p}^*$ to the same hypotheses is given by the function

$$\begin{aligned}\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] \;=\; & p_1\big(w_{11}(1-p_1^*)^2 + w_{21}(p_2^*)^2 + w_{31}(p_3^*)^2\big) \\ & + \; p_2\big(w_{12}(p_1^*)^2 + w_{22}(1-p_2^*)^2 + w_{32}(p_3^*)^2\big) \\ & + \; p_3\big(w_{13}(p_1^*)^2 + w_{23}(p_2^*)^2 + w_{33}(1-p_3^*)^2\big).\end{aligned}$$

Again without loss of generality, assume that the hypotheses are ordered by their distances from each other, with H_2 being equally far from H_1 and H_3, and H_1 and H_3 being twice as far from each other as they are from H_2. Then $w_{11} = w_{33}$, $w_{21} = w_{23}$, $w_{31} = w_{13}$, and $w_{12} = w_{32}$, so that we can simplify notation by defining $w_1 \coloneqq w_{11} = w_{33}$; $w_2 \coloneqq w_{21} = w_{23}$; $w_3 \coloneqq w_{31} = w_{13}$; $w_4 \coloneqq w_{12} = w_{32}$; and $w_5 \coloneqq w_{22}$. For $\mathscr{V}$ to be proper, it must hold that $\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] = \mathbf{p}$, for any distribution $\mathbf{p}$ on $\{H_1, H_2, H_3\}$. To

[12] To my knowledge, the only other author explicitly open to the possibility of "scoring rule pluralism" is Schurz (2018).

[13] I am greatly indebted to Eric Raidl, Christopher von Bülow, Verena Wagner, Sylvia Wenmackers, and two anonymous referees for valuable comments on previous versions of this paper. Thanks also to Lieven Decock, Samuel Fletcher, and Jos Uffink for helpful discussions. Versions of this paper were presented at the Universities of Düsseldorf and Konstanz and at the IHPST (Paris). I thank the audiences on those occasions for stimulating questions and remarks.

see whether this does hold, we use the method of Lagrange multipliers. Specifically, where $f(\mathbf{p}^*) = p_1^* + p_2^* + p_3^*$, we must find values for p_1^*, p_2^*, p_3^*, and λ such that $\nabla \mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] = \lambda \nabla f(\mathbf{p}^*)$ and $f(\mathbf{p}^*) = 1$. Calculating the first-order partial derivatives of $\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)]$, we find

$$\begin{aligned}
(\partial/\partial p_1^*)\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] &= -2w_1p_1(1-p_1^*) + 2w_3p_3p_1^* + 2w_4p_2p_1^*;\\
(\partial/\partial p_2^*)\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] &= -2w_5p_2(1-p_2^*) + 2w_2p_1p_2^* + 2w_2p_3p_2^*;\\
(\partial/\partial p_3^*)\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] &= -2w_1p_3(1-p_3^*) + 2w_3p_1p_3^* + 2w_4p_2p_3^*.
\end{aligned}$$

Because $\nabla f(\mathbf{p}^*) = \mathbf{1}$, we have $(\partial/\partial p_i^*)\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] = \lambda$ for all $i \leqslant 3$. So in particular, expanding the partial derivatives in p_1^* and p_3^* and dividing both by 2, we have

$$-w_1p_1 + w_1p_1p_1^* + w_3p_3p_1^* + w_4p_2p_1^* = -w_1p_3 + w_1p_3p_3^* + w_3p_1p_3^* + w_4p_2p_3^*,$$

and hence

$$w_1p_1p_1^* + w_3p_3p_1^* + w_4p_2p_1^* - w_1p_3p_3^* - w_3p_1p_3^* - w_4p_2p_3^* - w_1p_1 + w_1p_3 = 0.$$

Suppose that $\mathscr{V}$ is proper, so that $\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)]$ reaches its minimum if $p_1 = p_1^*$, $p_2 = p_2^*$, and $p_3 = p_3^*$. Then there must be values for the w_i such that

$$w_1(p_1)^2 + w_3p_3p_1 + w_4p_2p_1 - w_1(p_3)^2 - w_3p_1p_3 - w_4p_2p_3 - w_1p_1 + w_1p_3 = 0.$$

However, factoring the left-hand side yields

$$(p_1 - p_3)(-w_1 + w_1p_1 + w_4p_2 + w_1p_3).$$

This equals 0 iff either (i) $p_1 = p_3$ or (ii) $w_1 = w_4$, where the latter follows from the fact that the condition that the right-hand factor equals 0 can be rewritten as $w_1(1 - p_1 - p_3) = w_4p_2$, in conjunction with the fact that the p_i sum to 1. Because, as said, for $\mathscr{V}$ to be proper, it must hold for all $\mathbf{p}$ that $\arg\min_{\mathbf{p}^*} \mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)] = \mathbf{p}$, we may pick a $\mathbf{p}$ such that $p_1 \neq p_3$, thereby violating (i). As for (ii), note that whichever precise values the w_i assume, w_1 must be smaller than 1/3 (given that it is assigned to the supposed truth) and w_4 must be greater than 1/3 (given that it is assigned to the two hypotheses supposed false). Consequently, on the supposition that $\mathscr{V}$ is proper, we can minimize $\mathbb{E}_{\mathbf{p}}[\mathscr{V}(\mathbf{p}^*)]$ subject to the given constraint iff the truthlikeness weights assigned by the rule do *not* reflect truthlikeness in a minimally adequate sense. By assumption, the weights do reflect truthlikeness in a minimally adequate sense. Given that we made no further assumptions about $\mathscr{V}$, it follows that every VS rule is improper if it assigns truthlikeness weights in a minimally adequate fashion. □

Remark The above proof proceeds by constructing a specific counterexample involving three hypotheses that are assumed to stand in specific relations of truthlikeness to each other. To see that this assumption does not undermine the generality of the proof, we note that the said relations are perfectly possible according to all modern measures of truthlikeness (see page 5 for references). As a matter of fact, one can think of our ear-

lier example concerning the possible grades (A, B, or C) a given student may receive as instantiating exactly the relations of truthlikeness that are assumed to hold in the counterexample. It is also to be noted, however, that not *all* known measures of truthlikeness will do for the purposes of the proof. Most famously, Tichý (1974) discovered that on Popper's (1963) measure all false theories are equally far from the truth, contrary to what Popper had hoped to achieve with his measure.

Appendix B

In this appendix we prove

Theorem 2 *Let S be the standard unit $(n-1)$-simplex, let $\mathbf{p}$ and $\mathbf{p}^*$ range over vectors in S, and let $m\colon S \to S$ be defined as follows:*

$$m(\mathbf{p}) \;\coloneqq\; \arg\min_{\mathbf{p}^*} \sum_{i=1}^{n}\sum_{j=1}^{n} p_i w_{ij}(\delta_{ij} - p_j^*)^2,$$

with δ_{ij} the Kronecker delta, and with $w_{ij} > 0$ for all i, j, and $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} = 1$. Then there is a $\mathbf{p}^+ \in S$ such that (i) $m(\mathbf{p}^+) = \mathbf{p}^+$, (ii) $\mathbf{p}^+$ is unique, and (iii) $\mathbf{p}^+$ depends only on the w_{ij}.

Proof: Clause (i) follows from Brouwer's (1911) fixed-point theorem, which (in one version) states that every continuous function from a simplex onto itself has a fixed point. It does not follow from Brouwer's theorem that the fixed point is unique.

To prove clause (ii), then, one first verifies that the function that is being minimized at each step on the way to the fixed point has the Hessian

$$\begin{bmatrix} 2(p_1w_{11} + \cdots + p_nw_{1n}) & 0 & \cdots & 0 \\ 0 & 2(p_1w_{21} + \cdots p_nw_{2n}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2(p_1w_{n1} + \cdots + p_nw_{nn}) \end{bmatrix}$$

This is a diagonal matrix, so its eigenvalues are the diagonal elements, which, given the constraints on the p_i and w_{ij}, can be seen to be all necessarily positive. Therefore, the Hessian is positive definite everywhere, and given that a simplex is a convex set, it follows that the function that is minimized is strictly convex, and hence the minimum it reaches is unique. So, at each step toward the fixed point, a unique minimum is reached. As a result, the minimum reached at the fixed point is unique as well.

For clause (iii), finally, note that at the fixed point the function that is being minimized is of the form

$$m^+(\mathbf{p}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n} p_i w_{ij}(\delta_{ij} - p_j)^2.$$

Because the fixed point $\mathbf{p}^+$ is a minimum, it holds that $\nabla m^+(\mathbf{p}^+) = \mathbf{0}$. We obtain a system of n polynomial equations with n variables and with the w_{ij} as coefficients by setting

$(\partial/\partial p_i)m^+(\mathbf{p}^+) = 0$, for all $i \leqslant n$. This system has a unique solution (in virtue of the first two clauses), which is bound to be strictly in terms of the coefficients. □

References

Agresti, A. (2007) *An Introduction to Categorical Data Analysis*. Hoboken NJ: Wiley.

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

Bernardo, J. M. and 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.

Bickel, J. E. (2010) "Scoring rules and decision analysis education." *Decision Analysis* 7: 346–357.

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

Brouwer, L. E. J. (1911) "Über Abbildungen von Mannigfaltigkeiten." *Mathematische Annalen* 71: 97–115.

Cevolani, G., Festa, R., and Kuipers, T. A. F. (2013) "Verisimilitude and belief change for nomic conjunctive theories." *Synthese* 190: 3307–3324.

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

de Finetti, B. (1962) "Does it make sense to speak of 'good probability appraisers'?" In I. J. Good (ed.), *The Scientist Speculates: An Anthology of Partly-baked Ideas*. New York: Basic Books, pp. 357–364.

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

Gelman, A. and Hill, J. (2009) *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge: Cambridge University Press.

Gelman, A. and Shalizi, C. R. (2012) "Philosophy and the practice of Bayesian statistics in the social sciences." In H. Kincaid (ed.), *The Oxford Handbook of Philosophy of Social Science*. Oxford: Oxford University Press, pp. 259–273.

Gelman, A. and Shalizi, C. R. (2013) "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology* 66: 8–38.

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

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

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

Konek, J. (2016) "Probabilistic knowledge and cognitive ability," *Philosophical Review* 125: 509–587.

Kruschke, J. K. (2013) "Posterior predictive checks can and should be Bayesian." *British Journal of Mathematical and Statistical Psychology* 66: 45–56.

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. (2014) "Empirical progress and nomic truth approximation revisited." *Studies in History and Philosophy of Science* 46: 64–72.

Leitgeb, H. and Pettigrew, R. (2010) "An objective justification of Bayesianism I: Measuring inaccuracy," *Philosophy of Science* 77: 201–235.

Levinstein, B. A. (2012) "Leitgeb and Pettigrew on accuracy and updating," *Philosophy of Science* 79: 413–424.

Lombrozo, T. (2017) " 'Learning by thinking' in science and in everyday life." In P. Godfrey-Smith and A. Levy (eds.), *The Scientific Imagination*. Oxford: Oxford University Press, in press.

McCarthy, J. (1956) "Measures of the value of information." *Proceedings of the National Academy of Sciences* 42: 654–655.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., and Tetlock, P. (2015) "Identifying and cultivating superforecasters as a method of improving probabilistic predictions." *Perspectives on Psychological Science* 10: 267–281.

Moss, S. (2011) "Scoring rules and epistemic compromise." *Mind* 120: 1053–1069.

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

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*, Hoboken NJ: Wiley.

Popper, K. R. (1963) *Conjectures and Refutations*. London: Routledge and Kegan Paul.

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

Schurz, G. (1987) "A new definition of verisimilitude and its applications." In P. Weingartner and G. Schurz (eds.), *Logic, Philosophy of Science and Epistemology* (Proceedings of the 11th International Wittgenstein Symposium). Vienna: Hölder-Pichler-Tempsky, pp. 177–184.

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

Schurz, G. (2011) "Verisimilitude and belief revision." *Erkenntnis* 75: 203–221.

Schurz, G. (2014) *Philosophy of Science: A Unified Approach*. New York: Routledge.

Schurz, G. (2018) *The Optimality of Meta-induction: A New Approach to Hume's Problem*. Manuscript.

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

Tetlock, P. and Gardner, D. (2015) *Superforecasting: The Art and Science of Prediction*. London: Penguin Random House.

Tichý, P. (1974) "On Popper's definition of verisimilitude." *British Journal for the Philosophy of Science* 25: 155–160.
Winkler, R. L. (1969) "Scoring rules and the evaluation of probability assessors." *Journal of the American Statistical Association* 64: 1073–1078.
Winkler, R. L. (1996) "Scoring rules and the evaluation of probabilities." *Test* 5: 1–60.
Winkler, R. L. and Murphy, A. H. (1968) " 'Good' probability assessors." *Journal of Applied Meteorology* 7: 751–758.