In Favor of Logarithmic Scoring

Randall G. McCutcheon*†

Shuford, Albert, and Massengill proved, a half century ago, that the logarithmic scoring rule is the only proper measure of inaccuracy determined by a differentiable function of probability assigned the actual cell of a scored partition. In spite of this, the log rule has gained less traction in applied disciplines and among formal epistemologists that one might expect. In this article we show that the differentiability criterion in the Shuford et al. result is unnecessary and use the resulting simplified characterization of the logarithmic rule to give novel arguments in favor of it.

1. Introduction: Scoring Rules. Measures of epistemic utility (or disutility), that is, scoring rules, are used in various disciplines to elicit faithful report of, and to measure the accuracy (or inaccuracy) of, probabilistic predictions. Given a partition $A_1, \ldots, A_n$ of an event space, we take a scoring rule for this partition to be a function $\Psi : \Omega \times \{1, \ldots, n\} \rightarrow [0, \infty]$, where Ω is the set of n-tuples of nonnegative reals summing to 1.[1] We will look on $\Psi(\mu, j)$ as the *inaccuracy*—a negatively oriented quantity that one seeks to minimize—of the forecast $\mu = (Cr(A_1), \ldots, Cr(A_n))$ when A_j obtains.

One scoring rule with a long and storied history is the *quadratic rule* (Brier 1950),

$$B((x_1, \ldots, x_n), j) = \sum_{i=1}^{n} (x_i - y_i)^2,$$

Received June 2017; revised April 2018.

*To contact the author, please write to: Department of Mathematical Sciences, University of Memphis, Memphis, TN; e-mail: rmcctchn@memphis.edu.

†Thanks to Steve Kalikow and the anonymous referees and editors at *Philosophy of Science*.

1. We assume that agents adopt credence functions obeying the probability axioms. In particular, we will not concern ourselves with vindications of probabilism by way of accuracy considerations. Strictly speaking we would say that, for an agent submitting an incoherent credence function, inaccuracy ought simply to remain undefined.

Philosophy of Science, 86 (April 2019) pp. 286–303. 0031-8248/2019/8602-0005$10.00
Copyright 2019 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

where $y_i = 1$ for $i = j$ and $y_i = 0$ otherwise.[2] Other well-known scoring rules include

$$S((x_1, \ldots, x_n), j) = 1 - \frac{x_j}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$

(the so-called *spherical rule*), as well as the *logarithmic rule* (Good 1952),

$$L((x_1, \ldots, x_n), j) = -\log x_j.$$

Although the quadratic rule is the most popular ("by far," say Fallis and Lewis 2015, 576) scoring rule, many investigators adopt an implicitly pluralistic attitude. Each of several contenders, on this view, has good points and bad, with suitableness to a given application depending on a weighing of various considerations.

2. A Heuristic Argument for the Log Rule. We contend, to the contrary, that the logarithmic rule has sufficient virtues (and other rules sufficient defects) that it should be looked on as at least the clear favorite (and probably as the only serious contender). Although this conviction is bolstered in part by the arguments we develop below, it was, for us, in fact cemented by some austere, first-blush information-theoretic heuristics. Since we find these considerations as convincing now as ever, we rehearse a version here.

Imagine a tedious game of 20 questions in which we think of a number from 1 to 1,024. Obviously you can figure out our number in 10 questions if you like. First you ask whether the number is greater than 512 (say). Regardless of how we answer, you get 1 bit of information, winnowing the pool of live numbers to 512. If we say no, you next ask whether the number is greater than 256, and so on. What is going on? You gain 1 bit of information when we first answer "no, the number is not greater than 512" because your posterior probability (1/512) in the actual number is twice as great as your prior (1/1,024). This multiplier, 2, represents your information gain. To convert a multiplicative quantity to an additive one, one takes the logarithm of 2. (It is customary to use logarithms base 2.)

You might have started with a riskier question, say "is the number greater than 256?" If we had answered no, your risky behavior would have paid off; credence in the actual number would have quadrupled, from 1/1,024 to 1/256. Hence you would have gained 2 bits of information ($\log_2 4 = 2$). But if we had answered yes, your risky behavior would have cost you. Credence in the actual number would have jumped from 1/1,024 to merely

2. For two cell partitions, the square difference between prior and posterior is the same for each cell, so it is more common to use a "half Brier" score, equal to one square difference rather than their sum.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

1/768. The multiplier here is 4/3, and $\log_2(4/3) \approx .415$. Since the probability of this disappointment is 4/3, the expected information gain of the riskier path is only $(3/4)(.415) + (1/4)2 \approx .811$. So it is more prudent to ask, initially, whether the number is greater than 512.

The log rule is based on just this sort of "information counting." Suppose a weatherman is asked his credence in the proposition that it will rain tomorrow. If he answers 1/2, then regardless of whether it rains, he will gain 1 bit of information upon seeing the actual outcome. Namely, his credence in it will double, increasing from 1/2 to 1. If he has initial credence in rain of 1/4 and it rains, he will gain 2 bits of information. Namely, his credence in the actual outcome will quadruple, increasing from 1/4 to 1. If it does not rain, however, his credence in the actual outcome will increase by a factor of 4/3 (from 3/4 to 1). This, as we have seen, gives him $\approx .415$ bits of information.

The weatherman seeks to adopt credences that anticipate the actual to the greatest extent possible, in the sense of minimizing expected information gain. That is, he wants as much of the information that will be reflected in his posterior credences to be reflected already in his prior credences (the less he learns tomorrow, the more he knows today). The fact is general. It is rational to want your credences to reflect as much of your knowledge as possible. Using the log rule to measure inaccuracy captures this intuition; inaccuracy simply corresponds to the amount of posterior information not reflected in the priors, that is, $-\log_2 x$ (bits gained, or *surprisal*, in information-theoretic parlance), where x is prior credence in the relevant actual outcome.

3. Shuford, Albert, and Massengill on Logarithmic Scoring. Our more theoretical arguments for logarithmic scoring meanwhile are based on a result from Shuford, Albert, and Massengill (1966). Consider a partition (E_1, E_2, E_3) of an outcome space and an agent who announces a prior of $(x_1, x_2, 1 - x_1 - x_2)$ on that partition. Under the log rule, this agent's expected inaccuracy is then given by

$$L(x_1, x_2) = -p_1 \log x_1 - p_2 \log x_2 - p_3 \log(1 - x_1 - x_2),$$

where p_i is the actual probability (objective chance or ideal epistemic probability) of E_i.

We note two features: (1) The minimum expected inaccuracy occurs at $x_1 = p_1$, $x_2 = p_2$. (2) Inaccuracy is a differentiable (on (0, 1)) function of the agent's credence in the actual cell alone. For its satisfaction of the first property, one says that the logarithmic rule is *proper*—there is an incentive to have credences equal to the actual probabilities. Although this is plainly a desirable feature, there are other proper scoring rules (among them the Brier

https://doi.org/10.1086/702028 Published online by Cambridge University Press

score and the spherical rule). What sets the log rule apart is what Shuford et al. (1966) proved, namely, that the logarithmic is the only scoring rule (up to a constant multiple) satisfying both 1 and 2.[3] In light of this result, we take it that the case for the logarithmic rule (as against pluralism) turns on whether 2 can be established as a necessary desideratum.

We argue for this in two stages. First, we strengthen the result of Shuford et al. by eliminating the differentiability condition. (This technical portion is relegated to an appendix.) Second, we give two arguments that inaccuracy ought to be a function of the agent's credence in the actual cell alone.[4]

3.1. Inaccuracy and Likelihood. A more compelling feature of logarithmic scoring that has been noted in the literature (see, e.g., Bernardo 1979) is that it promotes a strong relationship between accuracy and Bayesian confirmation. In typical applications, one is interested in assigning a score to a probabilistic model for a random variable having unknown distribution (in response to a random sample R taken from it). When the distribution is unknown, R provides evidence that may confirm one candidate model at the expense of another. For A and B in the support of one's prior distribution over the "true chances," A receives greater confirmation by R than does B when A better fits the evidence, that is, when $\Pr(R|A) > \Pr(R|B)$.[5] It is plausible that accuracy should mirror confirmation, that is, that A should be deemed more accurate than B, given sample R, precisely when A receives greater confirmation by R.

Only a rule for which score is a function of credence assigned the actual cell alone can have this property. To illustrate, consider an agent C and a random experiment with outcome space $\{E_1, E_2, E_3\}$. We suppose that C views the experiment's true probability function (c_1, c_2, c_3) as a random variable

3. To see this, consider a scoring rule assigning value $f(p)$, where p is the agent's credence in the actual cell. (By definition $f(1) = 0$; inaccuracy is zero when the agent's prior reflects certainty in the actual outcome.) The expected score is $S(q_1, q_2) = p_1 f(q_1) + p_2 f(q_2) + (1 - p_1 - p_2) f(1 - q_1 - q_2)$. Since the function S is differentiable and has a global minimum at $q_1 = p_1$, $q_2 = p_2$, its partial derivatives $S_{q_1} = p_1 f'(q_1) - (1 - p_1 - p_2) f'(1 - q_1 - q_2)$ and $S_{q_2} = p_2 f'(q_2) - (1 - p_1 - p_2) f'(1 - q_1 - q_2)$ are both equal to zero there. This yields $p_1 f'(p_1) = p_2 f'(p_2)$. But this should hold for all $p_1 > 0$, $p_2 > 0$ with $0 < p_1 + p_2 < 1$. In other words $xf'(x)$ is constant on (0, 1), and one quickly determines $f(x)$ to be some constant multiple of $\log x$.

4. At least one set of authors, Knab and Schoenfield (2015), explicitly state (as part of an argument that quadratic scoring can give "strange" results) that "a probabilistic agent's accuracy . . . at world w should be determined solely by the amount of credence she invests in the true theory at w, and the amount of credence she invests in false theories at w." As we restrict to agents that obey the probability axioms, that is precisely what we are arguing for here.

5. We take the degree of confirmation of A by R to be $\Pr(A|R)/\Pr(A)$ when $\Pr(A) > 0$.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

having a continuous distribution, the statistics of which may be described by a probability density function $f(x, y, z)$ defined on triples (x, y, z) of nonnegative reals summing to 1. (C's actual credence in E_i is of course the expectation of c_i under this distribution.) The confirmation afforded a given triple $A = (x, y, z)$ by random sample $R = E_1$ is proportional to $\Pr(R|A) = \Pr(E_1|(x, y, z)) = x$. If greater accuracy is to correspond to greater confirmation, then, the inaccuracy of $A = (x, y, z)$ upon observation of E_1 must be a function of $x = A(E_1)$.

One might think that this argument only works if both A and B lie in the support of C's distribution over the true chances; if one starts out knowing that the true chances are either (1/3, 1/2, 1/6) or (1/6, 1/2, 1/3) then one cannot justify scoring, say, (1/4, 1/4, 1/2) in any particular way by its confirmation by evidence, since one's prior probability that this triple corresponds to the actual chances (or even the actual ideal epistemic probability) is zero. (Thanks to an anonymous referee for this point.) Recall though that inaccuracy scores are functions of credences and outcomes alone; they do not further depend on, say, one's distribution over the true chances. Justification in cases in which both A and B are in the support of C's distribution over the chances therefore generalizes to cases in which they are not.

There is precedent for both acknowledging the existence and denying the force of such considerations. Selten (1998, 44) in particular writes: "The logarithmic scoring rule has a close connection to the maximum likelihood principle. However, in spite of this theoretical advantage, the logarithmic scoring rule is not really recommendable." He then goes on to detail several objections against logarithmic scoring, two of which we examine in the next section. For now, we move to our second argument in favor of 2, which is more difficult to answer.

3.2. Inaccuracy and Untested Conditional Probabilities. Consider again an experiment having outcome space partitioned as $\{E_1, E_2, E_3\}$. A credence function Cr over this space is wholly determined by (*a*) the restriction of Cr to the subspace generated by $\{E_1, E_2 \vee E_3\}$ and (*b*) the conditional probability $Cr(E_2|E_2 \vee E_3)$. Suppose that there is a scoring rule S and credence functions $A = (a_1, a_2, a_3)$ and $B = (b_1, b_2, b_3)$ with $b_1 = a_1$ such that $S(A, 1) > S(B, 1)$; that is, A is judged less accurate than B when $R = E_1$ is observed. Since Cr is wholly determined by a and b, $S(Cr, 1)$ is wholly determined by a and b as well. Note that the restrictions of A and B to the subspace generated by $\{E_1, E_2 \vee E_3\}$ are identical; a, therefore, plays no part in the difference of $S(A, 1)$ and $S(B, 1)$. The reasons for this difference must therefore be found in b, that is, in the fact that the conditional probabilities $A(E_2|E_2 \vee E_3)$ and $B(E_2|E_2 \vee E_3)$ disagree.

To bring out the oddness of this, we can again take advantage of the fact that we are free to choose the experiment in any way we like. Here is our choice.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

First, a coin of uncertain bias is tossed. If the coin comes up heads, stop. If tails, a six-sided die of uncertain bias is then rolled. Let E_1 = heads, E_2 = tails ∧ six, and E_3 = tails ∧ ¬six. For emphasis, let A = (7/10, 1/10, 1/5) and B = (7/10, 3/20, 3/20).

The outcomes of the toss and the potential roll are (by stipulation) causally independent. Among agents respecting this stipulation, then, A would be adopted by, and only by, agents for whom the coin has expected propensity 7/10 to land heads and the die has expected propensity 1/3 to land six. Meanwhile, B would be adopted by, and only by, agents for whom the coin has expected propensity 7/10 to land heads and the die has expected propensity 1/2 to land six.

Suppose that the experiment is run, and the coin lands heads; that is, E_1 is realized. An agent adopting A (call her Amy) will be scored as less accurate than an agent adopting B (Beatrice). Why? Not on account of their attitudes toward heads; these are equivalent. Because Amy believes that the die has expected propensity 1/3 to land six, then, whereas for Beatrice the expected propensity in six is 1/2. But, Amy will no doubt protest, the die was not even rolled; no evidence bearing in any way on the propensity of the die to land six was gathered.

Assuming that no satisfactory response to Amy's complaint is forthcoming, all the pluralist can now do is acknowledge that the example does tell against rules that do not score by the actual cell alone while holding out hope that other considerations might tell equally against proper rules (constant multiples of the log rule by Shuford et al. [1966] and our appendix) that do. That project succeeding, it might then be thought that this is just one of those situations in which "one can't have it all" and that the weight of various considerations ought to determine the best choice of rule in specific applications—just what pluralism recommends.

Alas, this project is not as promising as one might hope. For, as we show in the next section, several well-regarded objections to the log rule miss their mark.

4. Objections to Logarithmic Scoring Answered. In this section we answer three objections to logarithmic scoring. This set of objections is surely not exhaustive, but we believe it to be fairly representative.

4.1. Convexity. Joyce (1998) defends a constraint on scoring rules S, according to which, for every pair of distinct credence functions c_1, c_2 and outcome j, if $S(c_1,j) = S(c_2,j)$ then $S((1/2)c_1 + (1/2)c_2, j) < S(c_1,j)$. That is, when two distinct credence functions are judged to be equally inaccurate for a given outcome, the midpoint of the two must be judged strictly more accurate than either. So, for example, since (7/10, 3/20, 3/20) is the midpoint of (7/10, 1/10, 1/5) and (7/10, 1/5, 1/10) (and since every scoring rule that mer-

https://doi.org/10.1086/702028 Published online by Cambridge University Press

its consideration is invariant under the permutation of cells), an advocate for this constraint would score (7/10, 3/20, 3/20) as strictly more accurate than (7/10, 1/5, 1/10) when E_1 obtains. Joyce (2009) gives new arguments in favor of this constraint, which he terms *convexity* (note, however, that he also backs away from his earlier position somewhat, choosing to treat convexity as an "optional constraint"):

> Suppose that a single ball will be drawn at random from an urn containing nine white balls and one black ball. On the basis of this evidence, a person might reasonably settle on a credence of $b = 0.1$ for the proposition that the black ball will be drawn and a credence of $b^- = 0.9$ for the proposition that a white ball will be drawn. Suppose that the ball is drawn, and that *we* learn that it is black. We are then asked to advise the person, without telling her which ball was drawn, whether or not to take a pill that will randomly raise or lower her credence for a black draw, with equal probability, by 0.01, while leaving her credence for a white draw at 0.9. If our only goal is to improve the person's epistemic utility, then our advice should depend on the convexity of the score for truths at credence 0.1. For a rule that is convex here . . . the pill's disadvantages outweigh its advantages. (Joyce 2009, 283)

Note that the pill induces probabilistic incoherence; it changes credence in black while leaving credence in white the same. That fact diminishes the force of the argument, as it opens the door for a critic to claim that incoherence, rather than concavity, is responsible for any encountered disutility. It would be better, we think (and we assume this going forward), to allow the agent's credence in white to vary in the expected way—become .89 when credence in black becomes .11, and so on.

Subsequent these changes we do at least agree with Joyce's claim that a scoring rule ought to deem use of the pill epistemically undesirable in the mean (the logarithmic rule does, as it satisfies convexity for two-cell partitions). The best explanation we see for this is that, if one assigns probabilities .09 and .11 to two independent events A and B, then one assigns probability .0099 to the conjunction $A \wedge B$, whereas if one assigns probabilities .1 and .1 to A and B, respectively, then one assigns probability $.01 > .0099$ to the conjunction. We accept 2 so, given any finite set of actual (independent) outcomes, we think one's inaccuracy with respect to the corresponding experiments ought to be a strictly decreasing function of the probability one assigns to the conjunction of those outcomes (so it is epistemically worse to have assigned half of them probability .09 and half of them probability .11 than it is to have assigned all of them probability .1).

That reasoning is unavailable for partitions having more than two cells. If an agent has credences (7/10, 3/20, 3/20) over a three-cell partition (E_1, E_2, E_3) and we know that E_1 is the case, the agent is exposed to no epistemic

https://doi.org/10.1086/702028 Published online by Cambridge University Press

risk, from our perspective, if she takes a pill that will fix her credence in the true outcome E_1 but induce small offsetting random perturbations in her credences for E_2 and E_3. What Joyce saw as epistemically undesirable was the employment of "a random process that has just as much chance of moving her away from the truth as it has of moving her toward it" (Joyce 2009, 284). This does not preclude indifference to the employment of a process inducing movements known (by us) to be orthogonal to the truth.[6]

4.2. Hypersensitivity. A number of objections to logarithmic scoring congregate around the fact that the log rule gives an inaccuracy score of infinity to an agent with zero credence in a realized cell of a scored partition. Even Shuford et al. (1966, 136–37), summarizing their important positive results for the rule, write: "In review, the 'logarithmic' scoring system is the only one which has the property that the student's score depends only on the probability that he assigns to the correct answer when there are more than two possible answers. All other (proper scoring rules) lack this property. We find, however, that the logarithmic scoring system is unbounded and thus impossible to realize in practice, e.g. how can one give a student a score of minus infinity?" Selten (1998, 51), meanwhile, writes: "The use of the logarithmic scoring rule implies . . . that wrongly describing something extremely improbable as having zero probability is an unforgivable sin."

We believe, to the contrary, that the log rule is both simple to realize (note that even a one-in-a-trillion outcome that comes out actual incurs an inaccuracy score of only 40 bits or so) and equitable in its judgments. As to the log rule's no-forgiveness policy regarding zero credences in actual outcomes, we find this reasonable. Offending agents would, in theory, take and lose arbitrarily many bets against a zero credence actuality, perhaps taking hallucination that the bets were going against them as a likely explanation for their mounting debts.

6. A further indication that this example is not harmful to the log rule is that the (standard piecemeal version of; see below) quadratic rule is more likely than the logarithmic rule to favor the pill taker over the nontaker over longish, finite sequences of independent draws from the urn. Letting Δ_q and Δ_l represent the greater inaccuracy incurred by the pill taker under the quadratic and logarithmic rules, respectively, Δ_q takes on values (.0021, −.0019, −.0179, .0181) and Δ_l takes on values ≈ (.016119665, −.01594154, −.13750352, .15200309) with probabilities (.45, .45, .05, .05). So $\sqrt{\text{Var}(\Delta_q)}/E(\Delta_q) = \sqrt{.000036}/.0001 = 60$, whereas $\sqrt{\text{Var}(\Delta_l)}/E(\Delta_l) \approx \sqrt{.00120056}/.0005580758 \approx 59.96898$. In 10^4 trials the expectations would increase 10,000 fold and the standard deviations 100 fold, so a case in which the pill taker had lesser measured inaccuracy would lie $\approx 1/.5996898 \approx 1.66753$ standard deviations from the mean under logarithmic scoring (occurs with frequency $p \approx .0477$) but only $1/.6 \approx 1.66667$ standard deviations under piecemeal quadratic scoring ($p \approx .0478$).

https://doi.org/10.1086/702028 Published online by Cambridge University Press

Selten (1998, 44) makes a related complaint, saying of the log rule that "it is too sensitive with respect to differences between very small probabilities." Precisely, Selten calls a scoring rule *S hypersensitive* if for every $\varepsilon > 0$ and every $M > 0$ there are probability distributions (over n-cell partitions, $n \geq 2$) r and p assigning positive measure to each cell such that the Euclidean distance from r to p is at most ε but the r-expectation of $S(p)$ exceeds the r-expectation of $S(r)$ by at least M. As Selten notes, the log rule is hypersensitive. As to why this is a problem, he writes: "in general, it will be very difficult to judge how small a very small probability should be. Usually there will be no good theoretical reasons to specify a probability as 10^{-5} rather than 10^{-10}. . . . Such differences can be of crucial importance for the comparison of the two theories" (51).

One needs to discriminate between two types of case, however. In the first type of case, when one is scoring a partition with a few large cells and a few (or one) small exceptional cell(s), hypersensitivity fails to manifest in logarithmic scoring provided one assigns even (very) modestly realistic credences. If one is scoring the toss of a fair coin, with outcomes heads, tails, and other (*other* being a conjunction of such unlikely scenarios as "lands on edge," "flies off into space," "unreadable," etc.), it matters little (in the mean) if one assigns *other* credence 10^{-10} or even 10^{-50} in a case in which the true probability is 10^{-5}; the unlikeliness of the outcome dwarfs the magnitude of the penalty. Mean inaccuracy will of course increase dramatically if one assigns *other* an excessively low probability, such as 10^{-10^9}, but these are just deserts for such an unconscionably impoverished estimate.

In the other type of case, in which there are many small, unexceptional cells, such as when a respondent fills out a multiquestion survey or reports credences about the outcome of a large single-elimination tournament, there typically are "good theoretical reasons" to specify one probability over another. In the case of a 64-competitor single-elimination tournament, there are 2^{63} possibilities for the final bracket, the most likely of which may have true probability $\approx 2^{-30}$. Even so, it is easy enough to specify an accurate prior for the realized bracket using the piecemeal approach of assigning credences to each contest, conditional (where applicable) on results to date, and multiplying. We return to this point below.

4.3. Neutrality. The final objection we will consider is based on symmetry considerations. Given a scoring rule S and a probability measure p on a partition of event space, Selten (1998) defines $V(p|q)$ to be the q-expectation of $S(p)$ (i.e., the expected inaccuracy of p in a case in which q gives the true probabilities) and defines the *expected score loss* of p at q by $L(p|q) = V(q|q) - V(p|q)$. (This is a measure of the greater mean inaccuracy incurred by choosing p rather than the true probability function q.) Selten then formu-

https://doi.org/10.1086/702028 Published online by Cambridge University Press

lates an axiom of *neutrality*, which states that $L(p|q) = L(q|p)$ for any q and p. He writes:

> The interpretation of (*Neutrality*) becomes clear if one looks at the hypothetical case that one and only one of two theories p and q is right, but it is not known which one. The expected score loss of the wrong theory is a measure of how far it is from the truth. It is only fair to require that this measure is "neutral" in the sense that it treats both theories equally. If p is wrong and q is right, then p should be considered to be as far from the truth as q in the opposite case that q is wrong and p is right. (Selten 1998, 54)

Having defended the logarithmic rule against the charges of no forgiveness and hypersensitivity, it may seem odd that we are choosing to address this charge last; if no forgiveness and hypersensitivity are justified, then neutrality clearly is not. We think, however, that there is value in looking at an argument against neutrality that does not bring in near-zero probabilities; more so, in that it will serve well as an introduction to the next section.

Selten introduces four axioms in all (of which neutrality is the last), then shows that, together, these axioms characterize the quadratic rule.[7] The logarithmic rule, meanwhile, satisfies the first three axioms but fails neutrality. Since we accept the first three axioms, then, for us neutrality and the quadratic rule are simply equivalent. Our argument against the former, then, will consist in showing how the latter engenders a deficient notion of "expected score loss."

Consider two agents, p and q. We assume that p has credence 1/2 in A, which has true probability 1/4, while q has credence 1/4 in an independent event B for which the true probability is 1/2. According to the reasoning behind neutrality, p is "as far from the truth" regarding A as q is regarding B. That is, inaccuracy should be scored in such a way that their "expected score losses" are equal. (And so they are, according to the quadratic scoring rule.) Suppose that we now attempt to flesh out p and q's epistemic attitudes without introducing further expected score loss: p and q correctly deem A and B to be independent, and both p's credence in B and q's credence in A are aligned to the true probabilities.

7. Although several good arguments against the quadratic rule appear in the literature of the past decade or so, few (if any) authors have offered wholesale endorsement of the logarithmic rule in its stead. Leitgeb and Pettigrew (2010) show that the quadratic rule is not consistent with Jeffrey conditionalization (Jeffrey 1965) but seem more willing to jettison the latter than the former. Levinstein (2012), responding to Leitgeb and Pettigrew, shows that the logarithmic rule does cohere with Jeffrey conditionalization but stops short of embracing it. Fallis and Lewis (2015) show that the quadratic rule does not even cohere with standard conditionalization. They do not, however, endorse the logarithmic rule.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

The situation is now as follows. Over the partition

$$W = \{A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B\},$$

p's and q's credence functions are (1/4, 1/4, 1/4, 1/4) and (1/16, 3/16, 3/16, 9/16), respectively. The true probabilities, meanwhile, are given by $r = (1/8, 1/8, 3/8, 3/8)$. It is easy to see that A's quadratic score over W is always $(3/4)^2 + 3(1/4)^2 = 3/4$. We leave it to the reader to verify that the true expectation (i.e., the r-expectation) of B's quadratic score is, however, 49/64; in particular, q now has higher expected score loss, according to the quadratic rule.

Since "expected score loss" should mean, roughly, "expected amount of gratuitous inaccuracy," one ought to reject any scoring rule according to which either p or q incurs any additional expected score loss in fleshing out their attitudes as they do, that is, in the ideally rational manner—let alone different amounts of it. The log rule, by contrast, is not subject to this objection. We therefore judge the quadratic scoring rule to be unacceptable. Concomitantly, we reject neutrality.

5. On the Computational Intractability of Competing Rules. Although we think that the arguments of the previous two sections provide compelling reasons to prefer proper rules that score by the actual cell alone (i.e., logarithmic rules), some readers will of course still insist on clinging to their favorite alternatives. In this section we suggest that the theoretical disadvantages these readers will meet with are the least of their worries. Indeed, once one advances beyond toy examples, scoring a credence function by a rule that does not score by the actual cell alone is apt to become computationally intractable.

Let us return to the single-elimination tournament example. In practice, it would be extraordinarily tedious to specify probabilities for all 2^{63} possible brackets. For the log rule this is not a problem. The agent simply assigns probabilities for each "first round" contest (outcomes from a given round may not be independent conditional on results of past rounds in general, so this constitutes a simplifying assumption) then, after learning which competitors prevailed in those contests (but nothing else), assigns probabilities for each "second round" contest, and so on. At the end, one may, by working backward, figure out the agent's prior probability in the actual bracket, which is sufficient to compute the agent's inaccuracy under the log rule. (This, owing to the identity $\log Cr(A \cap B) = \log Cr(A) + \log Cr(B|A)$, is equal to the sum of the individual inaccuracy scores for the 63 contests for which the agent provided credences.)

For scoring rules that do not score by credence in the actual cell alone, though, this shortcut will not serve. In order to score the 2^{63}-cell partition aris-

https://doi.org/10.1086/702028 Published online by Cambridge University Press

ing from the elimination tournament with the quadratic rule, for example, one requires credences for every cell. Unlike with the log rule, then, there is no simple, equivalent way to compute the score "piecemeal." Natural-looking attempts (such as computing quadratic scores for each contest and adding them) can give conflicting results.

To illustrate, suppose that we predict rain with probabilities 1/2 in New York (NY) and 1/2 in Tokyo, whereas your probabilities are 1/3 and 4/5, respectively. Suppose further that we agree that these are independent events. If it rains in both cities, then the sum of our quadratic scores for the two-cell partition determined by the weather in the two cities, respectively, is $(1/2)^2 + (1/2)^2 = 1/2$, while yours is $(2/3)^2 + (1/5)^2 = 109/225 < 1/2$. You are more accurate, then, according to this piecemeal approach.

But, your initial credence function on the smallest common refinement of the two independent partitions considered, namely,

$$\{(\text{NY}, \text{Tokyo}), (\text{NY}, \neg\text{Tokyo}), (\neg\text{NY}, \text{Tokyo}), (\neg\text{NY}, \neg\text{Tokyo})\},$$

was (4/15, 1/15, 8/15, 2/15). So your actual quadratic score over the common refinement is

$$\left(\frac{11}{15}\right)^2 + \left(\frac{1}{15}\right)^2 + \left(\frac{8}{15}\right)^2 + \left(\frac{2}{15}\right)^2 = \frac{190}{225}.$$

Our credence function on the refinement meanwhile was (1/4, 1/4, 1/4, 1/4), yielding a quadratic score of $(3/4)^2 + 3(1/4)^2 = 3/4 < 190/225$. We are therefore more accurate, according to this more "holistic" approach. So these are different rules.

Indeed, piecemeal versions of the quadratic rule depend further on the generating sequence of partitions employed. Suppose you are scored first on the event that it rains in either both or neither of the cities in question (in which you have credence 2/5). Upon learning the truth of this event, you would come to have credence 2/3 in NY. If you are then scored on NY, your running score would be $(3/5)^2 + (1/3)^2 = 106/225$. But imagine an agent whose priors are 1/2 in NY and .52 in Tokyo. That agent's running score would be .4804 (which lies strictly between 106/225 and 109/225) by either of the piecemeal methods we have considered. So, again, these are different rules.

Since scoring rules that are not functions of the actual cell alone generally depend on all cells, this is a problem that may plague any proper scoring rule that fails to satisfy 2, that is, all proper scoring rules, except for the logarithmic rule. For large partitions, such rules are difficult to evaluate, and seemingly natural piecemeal variants fail to be equivalent—both to the target rule and to each other.

https://doi.org/10.1086/702028 Published online by Cambridge University Press

6. Conclusion. In light of the compelling heuristics in its favor and the results of Shuford et al. (1966), it is surprising that the logarithmic scoring rule has lagged in popularity. Our goal has been to render it more palatable or at least a "necessary evil."

Other scoring rules, meanwhile, contradict confirmation, pay heed to untested differences in conditional probabilities, attribute increases in expected score loss to agents who have extended their credences ideally, and are likely to either depend arbitrarily on a choice of generating partitions or make computation intractable. In light of this, epistemologists and others who employ scoring rules to evaluate the accuracy of credences and have neglected the logarithmic rule would do well to reconsider its merits.

Appendix

Theorem 1. Let f, taking values in the extended reals, be strictly decreasing on [0, 1] with $f(1) = 0$. If for every $p, q \geq 0$ with $p + q \leq 1$ the function

$$H(x, y) = pf(x) + qf(y) + (1 - p - q)f(1 - x - y)$$

has a strict global minimum at $x = p$, $y = q$ then f is differentiable on (0, 1).

Remark. We adhere to the convention that $0 \cdot \infty = 0$, where applicable. Given that, theorem 1 as formulated immediately generalizes to versions with greater numbers of cells. For example, one may establish that if, for every $p, q, r \geq 0$ with $p + q + r \leq 1$, $H'(x, y, z) = pf(x) + qf(y) + rf(z) + (1 - p - q - r)f(1 - x - y)$ has a strict global minimum at $x = p$, $y = q$, $z = r$ then f is differentiable on (0, 1). (The proof is immediate; just set $r = 0$ and apply theorem 1.) The theorem does not, however, admit of a two-cell version. If, for example,

$$f(x) = \begin{cases} 3 + (1 - x)^2 & 0 \leq x < \frac{1}{3} \\ 1 + (1 - x)^2 & \frac{1}{3} \leq x \leq \frac{2}{3} \\ (1 - x)^2 & \frac{2}{3} < x \leq 1 \end{cases}$$

then $H''(x) = xf(x) + (1 - x)f(1 - x)$ has a strict global minimum at $x = p$, for every $0 \leq p \leq 1$ is not differentiable (or even continuous) on (0, 1) (cf. Predd et al. 2009, sec. 7.2.2).

https://doi.org/10.1086/702028 Published online by Cambridge University Press

Proof of Theorem 1. Suppose $0 < x, y$ and $x + y < 1$. Then for any $0 < \varepsilon < y$,

$$\begin{aligned} &xf(x) + yf(y) + (1 - x - y)f(1 - x - y) \\ &\quad < xf(x + \varepsilon) + yf(y - \varepsilon) + (1 - x - y)f(1 - x - y), \end{aligned}$$

so that

$$x(f(x + \varepsilon) - f(x)) > y(f(y) - f(y - \varepsilon)). \tag{A1}$$

Define $\theta(w, \varepsilon) = f(w + \varepsilon) - f(w)$ for $w, \varepsilon > 0$ with $w + \varepsilon < 1$. Then

$$x\theta(x, \varepsilon) > y\theta(y - \varepsilon, \varepsilon), \quad 0 < x, y, \quad x + y < 1, \quad 0 < \varepsilon < y.$$

Making the substitution $z = y - \varepsilon$, one has

$$x\theta(x, \varepsilon) > (z + \varepsilon)\theta(z, \varepsilon), \quad x, z, \varepsilon > 0, \quad x + z + \varepsilon < 1.$$

Switching the roles of x and z,

$$z\theta(z, \varepsilon) > (x + \varepsilon)\theta(x, \varepsilon) > \frac{x + \varepsilon}{x}(z + \varepsilon)\theta(z, \varepsilon),$$

or

$$\frac{z}{x + \varepsilon}\theta(z, \varepsilon) > \theta(x, \varepsilon) > \frac{z + \varepsilon}{x}\theta(z, \varepsilon), \quad x, z, \varepsilon > 0, \quad x + z + \varepsilon < 1. \tag{A2}$$

We claim that $\liminf_{\varepsilon \to 0^+}(\theta(x, \varepsilon)/\varepsilon) > -\infty$. Otherwise, dividing (A1) by ε and choosing a "bad" sequence of ε tending to zero, one could conclude that

$$\liminf_{\varepsilon \to 0^+} \frac{f(y) - f(y - \varepsilon)}{\varepsilon} = -\infty$$

for any $y \in [(1 - x)/2,\ 1 - x]$. Letting $M > 0$ be arbitrary and

$$t = \inf\left\{ y \in \left[\frac{1 - x}{2}, 1 - x\right] : f(y) > f(1 - x) + M(1 - x - y) \right\},$$

if $t > (1 - x)/2$ then choosing ε small so that $f(t) - f(t - \varepsilon) < -M\varepsilon$, one obtains

$$\begin{aligned} f(t - \varepsilon) &> f(t) + M\varepsilon \geq f(1 - x) + M(1 - x - t) + M\varepsilon \\ &= f(1 - x) + M(1 - x - (t - \varepsilon)). \end{aligned}$$

https://doi.org/10.1086/702028 Published online by Cambridge University Press

So $t = (1 - x)/2$ and $f((1 - x)/2) \geq f(1 - x) + M((1 - x)/2)$. But M is arbitrary, so this is absurd.

Suppose now that $0 < x < z < 1$. For any $\varepsilon, y > 0$ with $y + z + \varepsilon < 1$, from (A2)

$$\theta(x, \varepsilon) > \frac{y + \varepsilon}{x}\theta(y, \varepsilon)$$

and

$$\theta(y, \varepsilon) > \frac{z + \varepsilon}{y}\theta(z, \varepsilon).$$

It follows that

$$\theta(x, \varepsilon) > \frac{(y + \varepsilon)(z + \varepsilon)}{xy}\theta(z, \varepsilon). \tag{A3}$$

Similarly from (A2)

$$\theta(x, \varepsilon) < \frac{y}{x + \varepsilon}\theta(y, \varepsilon)$$

and

$$\theta(y, \varepsilon) < \frac{z}{y + \varepsilon}\theta(z, \varepsilon),$$

from which follows

$$\theta(x, \varepsilon) < \frac{yz}{(x + \varepsilon)(y + \varepsilon)}\theta(z, \varepsilon). \tag{A4}$$

Letting $y = (1 - z)/2$, (A3) and (A4) give

$$\frac{[(1 - z)/2]z}{(x + \varepsilon)\{[(1 - z)/2] + \varepsilon\}}\theta(z, \varepsilon) > \theta(x, \varepsilon) > \frac{\{[(1 - z)/2] + \varepsilon\}(z + \varepsilon)}{x[(1 - z)/2]}\theta(z, \varepsilon) \tag{A5}$$

whenever $0 < x < z < 1$ and $\varepsilon < (1 - z)/2$.

Fix now $x \in (0, 1)$, and let $B > 1$ be arbitrarily close to 1. Fix $\varepsilon > 0$ sufficiently small that $x + 3\varepsilon < 1$,

$$\frac{[(1 - z)/2]z}{(x + \gamma)\{[(1 - z)/2] + \gamma\}} > B^{-1/3} \quad \text{and} \quad \frac{\{[(1 - z)/2] + \gamma\}(z + \gamma)}{x[(1 - z)/2]} < B^{1/3}$$

https://doi.org/10.1086/702028 Published online by Cambridge University Press

for every $z \in [x, x + \varepsilon]$ and $0 < \gamma < \varepsilon$. Next let $0 < \gamma < \varepsilon$ be any number so small that, choosing N such that $\varepsilon/N \geq \gamma > \varepsilon/(N + 1)$, one has $(N + 1)/N < B^{1/6}$. Now for $j = 0, 1, \ldots, N$, set $z_j = x + j\gamma$. From (A5),

$$\frac{[(1 - z_j)/2]z_j}{(x + \gamma)\{[(1 - z_j)/2] + \gamma\}}\theta(z_j, \gamma) > \theta(x, \gamma) \rightarrow \theta(z_j, \gamma) > \theta(x, \gamma)/\frac{[(1 - z_j)/2]z_j}{(x + \gamma)\{[(1 - z_j)/2] + \gamma\}}.$$

Summing from 0 to N (and recalling that $\theta(x, \gamma) < 0$) yields

$$\theta(x, \varepsilon) > \sum_{j=0}^{N}\theta(z_j, \gamma) > \sum_{j=0}^{N}\theta(x, \gamma)/\frac{[(1 - z_j)/2]z_j}{(x + \gamma)\{[(1 - z_j)/2] + \gamma\}} > B^{1/3}\theta(x, \gamma)(N + 1).$$

Since $\varepsilon/N \geq \gamma$, one then has

$$\left(\frac{\theta(x, \varepsilon)}{\varepsilon}\right)(B^{-1/2}) > \left(\frac{\theta(x, \gamma)}{\varepsilon}\right)B^{-1/6}(N + 1) > \frac{N\theta(x, \gamma)}{\varepsilon} \geq \frac{\theta(x, \gamma)}{\gamma}.$$

Letting γ tend to zero, one obtains

$$\limsup_{\gamma \rightarrow 0^+}\left(\frac{\theta(x, \gamma)}{\gamma}\right) \leq \left(\frac{\theta(x, \varepsilon)}{\varepsilon}\right)(B^{-1/2}). \tag{A6}$$

Similarly, from (A5)

$$\theta(x, \gamma) > \frac{\{[(1 - z_j)/2] + \gamma\}(z_j + \gamma)}{x[(1 - z_j)/2]}\theta(z, \gamma) \rightarrow \theta(x, \gamma) / \frac{\{[(1 - z_j)/2] + \gamma\}(z_j + \gamma)}{x[(1 - z_j)/2]} > \theta(z, \gamma).$$

Summing from 0 to $N - 1$ yields

$$B^{-1/3}\theta(x, \gamma)N > \sum_{j=0}^{N-1}\theta(x, \gamma)/\frac{\{[(1 - z_j)/2] + \gamma\}(z_j + \gamma)}{x[(1 - z_j)/2]} > \sum_{j=0}^{N-1}\theta(z_j, \gamma) > \theta(x, \varepsilon).$$

Since $\varepsilon(N + 1) < \gamma$,

$$\frac{\theta(x, \gamma)}{\gamma} > \frac{\theta(x, \varepsilon)B^{1/3}}{N\gamma} > \frac{\theta(x, \varepsilon)(N + 1)B^{1/3}}{\varepsilon N} \geq \left(\frac{\theta(x, \varepsilon)}{\varepsilon}\right)(B^{1/2}).$$

https://doi.org/10.1086/702028 Published online by Cambridge University Press

Letting γ tend to zero, one obtains

$$\liminf_{\gamma \to 0^+} \left(\frac{\theta(x, \gamma)}{\gamma} \right) \geq \left(\frac{\theta(x, \varepsilon)}{\varepsilon} \right) \left(B^{1/2} \right). \tag{A7}$$

Since B may be taken arbitrarily close to 1 and $\limsup_{\varepsilon \to 0^+} (\theta(x, \varepsilon)/\varepsilon) < \infty$,

$$f'_+(x) = \lim_{\gamma \to 0^+} \left(\frac{\theta(x, \gamma)}{\gamma} \right)$$

exists for arbitrary $x \in (0, 1)$ by (A6) and (A7). One may show similarly that

$$f'_-(x) = \lim_{\gamma \to 0^-} \left(\frac{\theta(x, \gamma)}{\gamma} \right)$$

exists as well. (In particular, f is continuous on (0, 1).)

Dividing by ε on both sides of (A1) and letting ε tend to zero,

$$xf'_+(x) \geq yf'_-(y) \text{ for every } x, y > 0 \text{ with } x + y < 1. \tag{A8}$$

We claim that equality holds in (A8). Assume for contradiction that $xf'_+(x) > yf'_-(y)$ for some $x, y > 0$ with $x + y < 1$. Choose $T < 1$ such that $xf'_+(x) > Tyf'_-(y)$. Then for all sufficiently small $\varepsilon > 0$,

$$x(f(x + \varepsilon) - f(x)) > Ty(f(y) - f(y - \varepsilon)). \tag{A9}$$

Let

$$\begin{aligned} S = \sup \{ \gamma \in [0, \varepsilon] : & y(f(y - \varepsilon + \gamma) - f(y - \varepsilon)) \\ & > T^{-1}x(f(x + \varepsilon) - f(x + \varepsilon - \gamma)) \}. \end{aligned}$$

Note that (by continuity)

$$y(f(y - \varepsilon + S) - f(y - \varepsilon)) \geq T^{-1}x(f(x + \varepsilon) - f(x + \varepsilon - S)). \tag{A10}$$

Suppose that $S < \varepsilon$. Since $(y - \varepsilon + S)f'_+(y - \varepsilon + S) \geq (x + \varepsilon - S)f'_-(x + \varepsilon - S)$, for all sufficiently small $\gamma > 0$ one has $S + \gamma < \varepsilon$, and

$$\begin{aligned} & (y - \varepsilon + S)(f(y - \varepsilon + S + \gamma) - f(y - \varepsilon + S)) \\ & \quad > T^{-1/2}(x + \varepsilon - S)(f(x + \varepsilon - S) - f(x + \varepsilon - S - \varepsilon)) \\ & \quad \to y(f(y - \varepsilon + S + \gamma) - f(y - \varepsilon + S)) \\ & \quad > T^{-1}x(f(x + \varepsilon - S) - f(x + \varepsilon - S - \varepsilon)), \end{aligned}$$

https://doi.org/10.1086/702028 Published online by Cambridge University Press

the implication for ε sufficiently small (ε is chosen after T). Adding this to (A10),

$$y(f(y-\varepsilon+S+\gamma)-f(y-\varepsilon)) \geq T^{-1}x(f(x+\varepsilon)-f(x+\varepsilon-S-\gamma)).$$

Thus, $S+\gamma$ is in the set, the supremum of which is S. This contradiction establishes that $S=\varepsilon$. Thus, (A10) says that $y(f(y)-f(y-\varepsilon)) \geq T^{-1}x(f(x+\varepsilon)-f(x))$, contradicting (A9) and establishing that equality does in fact hold in (A8). Taking then $x=y \in (0,\ 1/2)$, $xf'_-(x)=xf'_+(x)$, so f is differentiable on $(0,\ 1/2)$. Finally for $x \in [1/2,\ 1)$, choose $y \in (0, 1-x)$ and note that

$$xf'_+(x)=yf'_-(y)=yf'_+(y)=xf'_-(x),$$

so f is differentiable on $(0,\ 1)$.

REFERENCES

Bernardo, J. M. 1979. "Expected Information as Expected Utility." *Annals of Statistics* 7:686–90.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78:1–3.

Fallis, Don, and Peter J. Lewis. 2015. "The Brier Rule Is Not a Good Measure of Epistemic Utility (and Other Useful Facts about Epistemic Betterness)." *Australasian Journal of Philosophy* 94:576–90.

Good, I. J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society* B 14:107–14.

Jeffrey, R. 1965. *The Logic of Decision*. New York: McGraw-Hill.

Joyce, J. M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65 (4): 575–603.

———. 2009. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief." In *Degrees of Belief*, ed. F. Huber and C. Schmidt-Petri. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 342. Dordrecht: Springer.

Knab, Brian, and Miriam Schoenfield 2015. "A Strange Thing about the Brier Score." *M-Phi*, March 12. http://m-phi.blogspot.nl/2015/03/a-strange-thing-about-brier-score.html.

Leitgeb, H., and R. Pettigrew. 2010. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77:236–72.

Levinstein, Benjamin Anders. 2012. "Leitgeb and Pettigrew on Accuracy and Updating." *Philosophy of Science* 79:413–24.

Predd, J., Robert Seiringer, Elliott H. Lieb, Daniel N. Osherson, H. Vincent Poor, and Sanjeev R. Kulkarni. 2009. "Probabilistic Coherence and Proper Scoring Rules." *IEEE Transactions on Information Theory* 55 (10): 4786–92.

Selten, Reinhard. 1998. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics* 1:43–62.

Shuford, Emir H., Jr., Arthur Albert, and H. Edward Massengill. 1966. "Admissible Probability Measurement Procedures." *Psychometrika* 31 (2): 125–45.

https://doi.org/10.1086/702028 Published online by Cambridge University Press