Bayesian Measures of Confirmation from Scoring Rules

Steven J. van Enk*†

I show how scoring rules, interpreted as measuring the inaccuracy of a set of degrees of belief, may be exploited to construct confirmation measures as used in Bayesian confirmation theory. I construct two confirmation measures from two particular standard scoring rules. One of these measures is genuinely new, the second is trivially ordinally equivalent to the difference measure. These two measures are tested against three well-known measures of confirmation in a simple but illuminating case that contains in a natural way the problem of irrelevant conjunction. The genuinely new measure emerges, arguably, as the best.

1. Introduction. Consider a Bayesian agent assigning degrees of belief to various propositions. There are two well-known concepts in this context that, I shall demonstrate, can be fruitfully connected: scoring rules and confirmation measures. Scoring rules (see, e.g., Winkler 1969; Bickel 2007; Gneiting and Raftery 2007) can play at least three different roles: (i) to force agents to make their degrees of belief follow the rules of probability calculus, (ii) to force their degrees of belief in given hypotheses to equal the probabilities they assign to those hypotheses, and (iii) to quantify the inaccuracy of their degrees of belief.

Irrespective of its use, a scoring rule associates with each proposition H (chosen from a partition of hypotheses over some domain) a penalty, the amount of which depends on the agent's degree of belief (or credence) b in that proposition and on the proposition's truth value t (with $t = 1$ for true and $t = 0$ for false). For an illustrative example, consider the "quadratic

Received June 2013; revised August 2013.

*To contact the author, please write to: Department of Physics, University of Oregon, Eugene, OR 97401; e-mail: svanenk@uoregon.edu.

†I thank the anonymous referees for their very useful comments.

Philosophy of Science, 81 (January 2014) pp. 101–113. 0031-8248/2014/8101-0003$10.00
Copyright 2014 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/674205 Published online by Cambridge University Press

scoring rule"—chosen because it is the simplest, although not necessarily the best—which assigns a penalty

$$Q_H(t,b) := (t-b)^2. \tag{1}$$

Suppose our agent assigns a probability p to the truth of proposition H. For that agent, the expected value of the penalty incurred on that hypothesis is then

$$\begin{aligned}\langle Q_H \rangle &= p(1-b)^2 + (1-p)b^2 \\ &= p(1-p)^2 + (1-p)p^2 + (p-b)^2.\end{aligned} \tag{2}$$

The second line shows that the expected penalty is minimized by choosing the degree of belief b equal to the probability p. This property makes the scoring rule *proper*. Furthermore, the penalty can be minimized only by choosing $b = p$, and that makes the quadratic scoring rule a *strictly proper* one. And thus agents have, to the extent that they care about minimizing expected penalties, an incentive to align degrees of belief with probabilities. This demonstrates use and interpretation ii of the scoring rule.

One can view the above argument as a Dutch Book argument, in that an agent who is betting on his or her beliefs can be made to lose money if those degrees of belief do not equal probabilities (see De Finetti 1990; Schervish, Kadane, and Seidenfeld 2003). The same argument, using the same scoring rule, can be interpreted differently (Joyce 1998; see also Leitgeb and Pettigrew 2010; Levinstein 2012): one can view the expected penalty as a measure of the inaccuracy of the agent's belief (interpretation iii). With this interpretation, the agent must align credences with probabilities in order to be the least inaccurate. The philosophical advantage of interpreting the scoring rule this way is that with it the agent has an epistemic rather than a pragmatic reason for adhering to certain beliefs. I will stick with this interpretation in the following, and I will assume that the agent chooses $b = p$ for that epistemic reason. (I simply accept that a Bayesian agent's credence assignments conform to probability calculus, and interpretation i of scoring rules plays no role in this study.)

Suppose our agent learns a new piece of evidence E concerning hypothesis H, which makes the agent update his or her degree of belief from $b = p$ to $b' = p'$ (in different notation, $p = \Pr(H)$ and $p' = \Pr(H|E)$). In the light of the new probability assigned to H, the agent's previous degree of belief b is now revealed to be less accurate. Indeed, according to the quadratic scoring rule, the change in inaccuracy thanks to the new evidence is negative and equals

$$\Delta Q_H := \langle Q'_H \rangle - \langle Q_H \rangle = -(b-b')^2. \tag{3}$$

https://doi.org/10.1086/674205 Published online by Cambridge University Press

Here the expected value of the penalty incurred on the previous belief b is based on the updated probability p'.

The basic observation of the present study is now this: equation (3) quantifies how much the evidence E has changed the agent's belief in the hypothesis H, but Bayesian confirmation measures are meant to do exactly this, too. So why not set the degree of confirmation equal to ΔQ_H?

2. Two Confirmation Measures from Scoring Rules. To answer this question, we first have to overcome a technical obstacle: whereas a scoring rule does not distinguish between evidence being undermining or confirming belief in H, a confirmation measure does make this distinction, namely, by including a sign in the amount of confirmation. That is, if we denote a generic measure of confirmation by $C(H;E)$, then, by convention, $C(H;E) < 0$ indicates the agent's belief in the truth of H is undermined by E, while if $C(H;E) > 0$ it is supported. Fortunately, in the case of the quadratic scoring rule, this obstacle is easily overcome by explicitly including this sign in the definition of the degree of confirmation (by setting it equal to either $+\Delta Q_H$ or $-\Delta Q_H$). I will in addition show in this section how the same simple solution works just as well for another well-established scoring rule. (I relegate to the appendix the discussion of two further scoring rules for which this simple solution does not work. Although one can still construct confirmation measures connected to these scoring rules, the measures are no longer uniquely determined by the scoring rules and thereby lose some appeal.)

2.1. Quadratic Confirmation Measure. Denoting the confirmation measure based on ΔQ_H by Q we include a (negative) positive sign for (dis)confirming cases by setting Q equal to

$$Q(H;E) := \begin{cases} +(b - b')^2 & \text{if } b' > b \\ -(b - b')^2 & \text{otherwise} \end{cases}, \tag{4}$$

such that

$$|Q(H;E)| = -\Delta Q_H. \tag{5}$$

This measure of confirmation, which we could call the "quadratic confirmation measure," is trivially ordinally equivalent to the difference measure $d(H;E) = b' - b$ (as proposed by Carnap [1962]). For example, the measures d and Q always agree on (i) which of any two pieces of evidence supports the belief in a given proposition more than the other and (ii) which of any pair of propositions is supported more than the other by a given piece of evidence. The measure Q has at least one advantage over d (apart from being derivable from a scoring rule): the linear dependence on b' and b leaves d, unlike Q, open to attacks like (one version of) the Popper-Miller argument against inductive inferences (Fitelson 1999). On the other hand, all positive

https://doi.org/10.1086/674205 Published online by Cambridge University Press

features of the difference measure discussed in Fitelson (1999) and Eells and Fitelson (2002) are possessed by Q as well.

The quadratic scoring rule is not optimal according to the pragmatic criteria put forward in Bickel (2007), and there is no reason to expect the corresponding confirmation measure Q to be optimal, either. So the next order of business will be to discuss a second scoring rule and see whether and how one can construct a (possibly better) confirmation measure from it.

Before doing this, however, there is one more issue to address, and that is to give the reason for using the difference, in equation (3), between the old and new inaccuracies for quantifying the change in accuracy, rather than, say, their ratio, or the logarithm of their ratio (for confirmation measures that same choice is open for debate). The reason is this. Two scoring rules are considered equivalent if and only if they are related to each other via a positive affine transformation, that is, via multiplication with a positive constant and addition of another constant. Under a positive affine transformation, the difference between old and new penalties changes by just the multiplication factor, and that transformation in turn leaves confirmation measures equivalent. The ratio of old and new penalties, in contrast, changes in quite arbitrary ways and can, for instance, be made infinite (and thereby meaningless) by an appropriate choice of the affine transformation (provided old and new penalties are not equal). Only by taking the difference between old and new penalties do equivalent scoring rules lead to equivalent confirmation measures.

2.2. Logarithmic Confirmation Measure. In 1952 Good proposed a logarithmic scoring rule of the form

$$G_H(t, b) := -t \ln b - (1 - t) \ln (1 - b). \tag{6}$$

Using the inaccuracy interpretation of this scoring rule, updating an agent's degree of belief from b to b' makes that belief less inaccurate. This can be shown by applying the Gibbs inequality ($\ln x \leq x - 1$ with equality only for $x = 1$) twice, so that

$$\Delta G_H := \langle G_H' \rangle - \langle G_H \rangle = b' \ln \frac{b}{b'} + (1 - b') \ln \frac{1 - b}{1 - b'} \leq 0, \tag{7}$$

with equality only for $b' = b$. This also shows that this rule is strictly proper. By following the same steps as before, we construct the corresponding measure of confirmation, which we denote by L, by setting

$$L(H; E) = \begin{cases} b' \ln \dfrac{b'}{b} + (1 - b') \ln \dfrac{1 - b'}{1 - b} & \text{if } b' > b \\ b' \ln \dfrac{b}{b'} + (1 - b') \ln \dfrac{1 - b}{1 - b'} & \text{otherwise} \end{cases}, \tag{8}$$

https://doi.org/10.1086/674205 Published online by Cambridge University Press

such that

$$|L(H;E)| = -\Delta G_H. \tag{9}$$

This measure (as well as Q) satisfies the Hypothesis Symmetry (Eells and Fitelson 2002). That is, under the replacement of a hypothesis H by the catchall ^{-}H, the confirmation measure changes sign. In the notation used here, this means that L changes sign under the simultaneous replacements $b \leftrightarrow 1 - b$ and $b' \leftrightarrow 1 - b'$. In words, evidence confirms H just as much as it disconfirms ^{-}H. Eells and Fitelson regard this as a good symmetry for confirmation measures to possess. On the other hand, Eells and Fitelson also discuss two symmetries that, they argue, confirmation measures should not possess, and it turns out that, indeed, neither L nor Q satisfy Evidence Symmetry (i.e., $C(H;E) = -C(H;{}^{-}E)$ for all evidence E and any hypothesis H) or Commutativity Symmetry (i.e., $C(H;E) = C(E;H)$).

I end with two technical observations. First, L increases with b' for fixed b with $b' > b$ and decreases with b' for fixed b when $b' < b$, as a confirmation measure indeed should. Second, for relatively small changes in belief, that is, for $b' = (1 + \varepsilon)b$ with $\varepsilon \ll 1$, we have (for not-too-large values of b)

$$L(H;E) \approx \begin{cases} +\dfrac{(b'-b)^2}{2b} & \text{if } b' > b \\ -\dfrac{(b'-b^2)}{2b} & \text{otherwise} \end{cases}. \tag{10}$$

This observation helps one to see that the two measures Q and L are not ordinally equivalent. Just consider two cases of supporting evidence, case 1 where $b = 0.1$ and $b' = 0.101$, and case 2 where $b = 0.2$ and $b' = 0.2011$. According to measure Q the first case provides less support than does the second, but according to measure L case 1 provides more support.

3. A Test: The Bit String Guessing Game. In order to see whether there is any merit to at least one of the confirmation measures just constructed, I consider a simple game. The goal is to guess a secret M-bit string, which has been picked at random from all $N = 2^M$ possible M-bit strings. Corresponding to each possible string k there is the proposition H_k which states that string k is the secret one. Suppose our agent discovers the first piece of evidence, namely, that the first bit is "0." The agent updates her or his degrees of belief from $b_k = 1/N$ for all N propositions to (i) $b_k' = 2/N$ for the propositions about those bit strings that start with "0" and (ii) $b_k' = 0$ for the remaining ones. Subsequently discovering the next n bits would be described in an analogous way, by just replacing N by $N/2^n$, and so on.

3.1. Irrelevant Conjunction. This simple game provides an instance of a problem that has been dealt with successfully within Bayesian confirmation theory, namely, the problem of irrelevant conjunction (see, e.g., Haw-

https://doi.org/10.1086/674205 Published online by Cambridge University Press

thorne and Fitelson 2004). For, given that the evidence concerns only the first bit, each proposition H_k can be viewed as a conjunction of a relevant statement about the first bit with irrelevant statements about the other bits. The problem of irrelevant conjunction is, briefly, this: if evidence E confirms some hypothesis H, in the sense of increasing the probability assigned to the truth of H, then it also confirms, in the same sense, the composite hypothesis H & X, where X is any irrelevant proposition. Intuitively, that does not seem quite right. The way Bayesian confirmation theory has solved this problem is by showing that, at least for some confirmation measures, the degree of confirmation, although still positive, is diminished by adding irrelevant propositions. I will demand the same feature of the measures of confirmation considered here.

3.2. Desiderata. What do the various Bayesian measures of confirmation, both previously known ones and the ones proposed here, have to say about the evidence obtained in the bit-string guessing game? I will judge the confirmation measures on four criteria. I state the criteria in terms of a generic measure of confirmation denoted as C, and I abbreviate $C(H_k; E) =: C_k$.

(I) The evidence *is* literally 1 bit of information. The total amount of confirmation, defined here as the sum $\sum_k |C_k|$, ought to reflect this fact, in that, taken over all values of M, this sum ought to be bounded from above, and it ought to be bounded from below away from zero. In other words, the total amount of confirmation should never be zero nor infinite, not even in the limit of $M \to \infty$.

(II) A necessary property of C, if it is to handle the problem of irrelevant conjunction properly, is that for any k the absolute value $|C_k|$ strictly decrease with increasing number of bits M, unless $C_k = 0$, when it should stay equal to zero for all M.

(III) For those propositions for which the evidence is undermining, it is decisive in that no further evidence can change the agent's degree of belief. But the confirming evidence is not decisive, unless $N = 2$. Thus, for $N > 2$, the amount of disconfirmation brought about by the one bit of evidence ought to be larger (in absolute value) than the amount of confirmation.

(IV) For $N = 2$ the amount of confirming evidence ought to equal the amount of undermining evidence, as in both cases the evidence is decisive and the situation is symmetric under the interchange $H \leftrightarrow {}^{-}H$.

3.3. Test Results. I apply these criteria to five different measures of confirmation, with the following results.

https://doi.org/10.1086/674205 Published online by Cambridge University Press

Q. According to the quadratic measure Q we have the following degrees of confirmation:

$$Q(H_k; E) = \begin{cases} +1/N^2 & \text{for all } k : b_k' = 2/N \\ -1/N^2 & \text{for all } k : b_k' = 0 \end{cases}. \tag{11}$$

This measure passes requirements II and IV but fails I and III. It fails I because in the limit $M \to \infty$ we have $\sum_k |Q(H_k; E)| \to 0$.

L. According to the logarithmic measure L derived from Good's scoring rule, we have

$$L(H_k; E) = \begin{cases} 2 \ln 2/N - (1 - 2/N) \ln \dfrac{N-1}{N-2} & \text{for all } k : b_k' = 2/N \\ \ln(1 - 1/N) & \text{for all } k : b_k' = 0 \end{cases}. \tag{12}$$

This measure passes all four tests. For $N = 2$ the degree of confirmation is $\pm \ln 2$, precisely quantifying the one bit of information that has been obtained as evidence (one bit equals ln2 nats [natural units]).

l. Consider the well-known measure l, the log likelihood ratio (as advocated by Good [1983]). Here the evidence is quantified as

$$l(H_k; E) = \begin{cases} +\ln 2 + \ln \dfrac{N-1}{N-2} & \text{for all } k : b_k' = 2/N \\ -\infty & \text{for all } k : b_k' = 0 \end{cases}. \tag{13}$$

This measure passes III and IV but fails I. The confirmation part passes II, but the disconfirmation part does not. (More about how and why l fails here follows below.)

d. The difference measure satisfies

$$d(H_k; E) = \begin{cases} +1/N & \text{for all } k : b_k' = 2/N \\ -1/N & \text{for all } k : b_k' = 0 \end{cases}, \tag{14}$$

and the total amount of confirmation, $\sum_k |d(H_k; E)|$, equals 1. This measure passes I, II, and IV, but it fails III.

r. The log probability ratio, denoted by r, was advocated as the best Bayesian confirmation measure in Milne (1996) and originates from Keynes (2004). It yields

https://doi.org/10.1086/674205 Published online by Cambridge University Press

$$r(H_k;E)=\begin{cases}+\ln 2 & \text{for all } k: b'_k = 2/N \\ -\infty & \text{for all } k: b'_k = 0\end{cases}. \tag{15}$$

This measure passes only III and fails the other three criteria.

3.4. L versus l. Only one measure of the five considered here satisfies all four criteria, namely, the measure L. The reason that the measure l fails some criteria is because it assigns an infinite amount of confirmation to decisive evidence, but there is a good argument for doing so. No finite amount of future evidence is able to reverse the verdict that a degree of belief is zero or that a degree of belief equals unity. One could say, then, that the main difference between measures L and l is, roughly, that the former directly quantifies the present evidence, whereas the latter quantifies the future evidence needed to undo the present evidence.

A similar characterization of the difference between L and l presents itself when examining the following argument (from Fitelson 2001) given in favor of l. The argument states that decisive evidence for or against a hypothesis H should always count as the maximal amount of evidence, irrespective of the prior probability assigned to H. The latter condition makes this "maximality" requirement incompatible with criterion II. The measure l satisfies this maximality requirement by always assigning an infinite degree of confirmation to decisive evidence. In particular, it does this irrespective of how many bits the evidence is composed of. In contrast, L always assigns a finite amount $\ln(1/b)$ to decisive evidence, with b the prior probability assigned to the hypothesis that is now conclusively confirmed. This implies, for instance, that, for the test case considered here, M decisive bits taken together confer a degree of confirmation of $M\ln 2$ according to L, thus precisely capturing the amount of information contained in that evidence.

3.5. Q versus d. Although the measures Q and d are trivially ordinally equivalent, one way in which Q performs better than d was mentioned earlier, in section 2.1. The bit string test, on the other hand, gives an argument in favor of d and against Q: whereas d fails only requirement III, Q fails in addition requirement I.

3.6. L versus Q. The fact that the confirmation measure L derived from Good's logarithmic scoring rule was found to be superior to the measure Q derived from the quadratic scoring rule happily agrees with the former scoring rule being considered superior to the latter rule on the basis of pragmatic criteria discussed by Bickel (2007). We may draw the conclusion, also, that as an epistemic measure of inaccuracy, the logarithmic scoring rule is better than the quadratic scoring rule. This conclusion agrees with Levinstein (2012) but is contra Leitgeb and Pettigrew (2010).

https://doi.org/10.1086/674205 Published online by Cambridge University Press

4. Conclusions. The point of this article was to demonstrate a connection between scoring rules and Bayesian confirmation measures by showing how to construct measures of confirmation from two particular scoring rules. The idea behind this construction is as follows. First, within Bayesian confirmation theory, how much a given piece of evidence supports or undermines a hypothesis is determined by how much the agent's belief in that hypothesis changed. Second, this change in belief may be quantified by how much less inaccurate that belief has become. Third, the decrease in inaccuracy is appropriately measured by a scoring rule (in Joyce's [1998] interpretation).

One surprise—given the plethora of existing measures of confirmation—is that one of these measures is genuinely new. Another surprise is that in a simple test that new measure (based on a scoring rule proposed by Good in 1952) emerges as the best, even when compared to established confirmation measures. In standard notation its magnitude is defined as

$$|L(H;E)| = \Pr(H|E)\ln\frac{\Pr(H|E)}{\Pr(H)} + \Pr(\neg H|E)\ln\frac{\Pr(\neg H|E)}{\Pr(\neg H)}. \qquad (16)$$

This measure shares several good properties with the log likelihood ratio (also connected to Good's name), which in the same notation can be written so as to bring out the similarities between the two, as

$$l(H;E) = \ln\frac{\Pr(H|E)}{\Pr(H)} - \ln\frac{\Pr(\neg H|E)}{\Pr(\neg H)}. \qquad (17)$$

I argued that l can be seen as quantifying the future evidence needed to undo evidence E, whereas L directly quantifies the information contained in E.

The above legitimizes L as a valid Bayesian confirmation measure. Further tests in other relevant contexts are an obvious avenue of research.

Appendix

This appendix will show how two other scoring rules can be used to construct two confirmation measures in less straightforward ways. The measures obtained here are by no means unique. Such measures may have their uses, but given the strength of the measure L, the results here are presented merely for the sake of completeness. (Neither measure satisfies all criteria of sec. 3.2.)

Asymmetric Logarithmic Rule. In the case of the two scoring rules discussed in the main text, the decrease of inaccuracy only depends on the old and new degrees of belief in the proposition under consideration. In particular, it does not depend on any other degrees of belief. For other scoring

https://doi.org/10.1086/674205 Published online by Cambridge University Press

rules this is not necessarily the case. In order to treat the role of other propositions in those scoring rules, we suppose, for simplicity and convenience, that we can create a countable partition of our domain of interest. That is, we assume we have a set of N mutually exclusive propositions $\mathbf{H} = \{H_1, H_2, \ldots H_N\}$ such that exactly one H_i must be true (we may take the limit of $N \to \infty$ at the end of a calculation). Moreover, it will be advantageous to also assume that all propositions are simple, in that the probability of any evidence to occur is fully determined by specifying which of the propositions in the partition is true. We denote our agent's old and new degrees of belief in proposition H_i by b_i and b_i', respectively. We have, then, $\sum_i b_i = \sum_i b_i' = 1$.

One may define an asymmetric version of Good's rule (see Bickel [2007] or Gneiting and Raftery [2007]) by setting the penalty as

$$L_H(t, b) := -t \ln b. \tag{A1}$$

So in this case a penalty is incurred only for the proposition H that turns out to be true. The expected change in inaccuracy upon changing the degree of belief from b to b' is now

$$\Delta L_H := \langle L_H' \rangle - \langle L_H \rangle = b' \ln(b/b'). \tag{A2}$$

The inaccuracy may actually increase, namely, when $b' < b$. At first sight this may seem to automatically overcome the technical obstacle mentioned above, in that it may seem that we could just define a confirmation measure by setting $L := -\Delta L_H$. However, this fails because the amount of disconfirmation would equal zero whenever $b' = 0$, whereas it should be strictly negative.

The scoring rule (A1) is, in spite of appearances perhaps, strictly proper, when one takes into account that, when $b' < b$, at least one of the other degrees of belief must increase in such a way that the total expected change in inaccuracy actually always does decrease. We can see this by examining the following auxiliary quantity:

$$A_i := b_i' \left[\ln \frac{b_i}{b_i'} - \frac{b_i}{b_i'} + 1 \right]. \tag{A3}$$

By the Gibbs inequality this quantity is never positive and equal to zero only if $b_i = b_i'$. Moreover, we have

$$\sum_i A_i = \sum_i b_i' \ln \frac{b_i}{b_i'}. \tag{A4}$$

This sum represents the change in the sum of the inaccuracies of our agent's beliefs about all the propositions in the partition (cf. eq. [A2]). This sum

https://doi.org/10.1086/674205 Published online by Cambridge University Press

will be referred to as "the total inaccuracy." (It in fact equals the Kullback-Leibler divergence between the old and the new probability distributions, which has the interpretation as the amount of information lost when the old probability distribution is used instead of the updated one.) The change in total inaccuracy is always negative unless none of the agent's credences change.

After these preliminary remarks, we are ready to define a measure of confirmation as

$$A(H;\, E) = \begin{cases} b' \ln \dfrac{b}{b} + b - b' & \text{if } b' > b \\ b' \ln \dfrac{b}{b'} + b' - b & \text{otherwise} \end{cases}. \tag{A5}$$

We might denote A as the "asymmetric logarithmic measure." Even though we had to invoke other propositions to demonstrate the strict properness of the scoring rule, A depends only on the old and new credences in the proposition under consideration. This measure is constructed so as to satisfy

$$\sum_i |A(H_i;\, E)| = -\sum_i \Delta L_{H_i}. \tag{A6}$$

This constraint is (much) weaker than (5) or (9), as it restricts only the sum, and not each individual term. As such, the choice of measure A is not unique and one might even say it is ad hoc. The measure satisfies criteria I, II, and III, but not IV. It does not satisfy the Hypothesis Symmetry of Eells and Fitelson (2002).

Spherical Scoring Rule. The spherical scoring rule depends explicitly on the distribution of values $\{b_j\}$ through the Euclidean length of the vector

$$\vec{b} := (b_1, b_2, \ldots b_N),\ |\vec{b}| = \sqrt{\textstyle\sum_j b_j^2}.$$

For proposition H_i the penalty is set as

$$S_{H,i} = t\left(1 - \frac{b_i}{|\vec{b}|}\right). \tag{A7}$$

Again, a penalty is incurred only if proposition H_i turns out to be true, and the penalty increases from 0 for $b_i = 1$ to a maximum of 1 for $b_i = 0$. Upon changing beliefs, the change in expected value of the penalty incurred on proposition H_i is

https://doi.org/10.1086/674205 Published online by Cambridge University Press

$$\Delta S_{H_i} := \langle S'_{H,i} \rangle - \langle S_{H,i} \rangle = b'_i \left[\frac{b_i}{|\vec{b}|} - \frac{b'_i}{|\vec{b}'|} \right]. \tag{A8}$$

Analogously to the case of the asymmetric logarithmic scoring rule, ΔS_{H_i} may be positive for some values of i, but the total inaccuracy never increases. The latter statement follows from the Cauchy-Schwarz inequality

$$\sum_i \Delta S_{H_i} = \frac{\vec{b}' \cdot \vec{b}}{|\vec{b}|} - |\vec{b}'| \leq 0, \tag{A9}$$

with equality only for $\vec{b}' = \vec{b}$. Now we can rewrite the change in total inaccuracy as

$$\sum_i \Delta S_{H_i} = -\frac{1}{2} |\vec{b}'| \sum_i \left(\frac{b'_i}{|\vec{b}'|} - \frac{b_i}{|\vec{b}|} \right)^2. \tag{A10}$$

The right-hand side is a sum of all nonpositive terms, but the conditions under which the individual terms in the sum are zero are the wrong conditions for confirmation measures, namely, $b'_i |\vec{b}| = b_i |\vec{b}'|$, instead of what would be the correct conditions $b'_i = b_i$. To fix this problem, let us define for $\vec{b}' \neq \vec{b}$ the following positive "global" quantity, which depends on the whole probability distribution,

$$\gamma_E = \frac{|\vec{b}'||\vec{b}| - \vec{b}' \cdot \vec{b}}{|\vec{b}||\vec{b}' - \vec{b}|^2}, \tag{A11}$$

where the subscript E indicates that the value of γ depends explicitly on $\vec{b}'$ and thereby on the evidence E. Then we can rewrite

$$\sum_i \Delta S_{H_i} = -\gamma_E \sum_i (b'_i - b_i)^2 \tag{A12}$$

as a sum of nonpositive terms that vanish only when $b'_i = b_i$. This enables us to define

$$S(H; E) = \gamma_E Q(H; E) \tag{A13}$$

so as to satisfy the constraint

$$\sum_i |S(H_i; E)| = -\sum_i \Delta S_{H_i}. \tag{A14}$$

https://doi.org/10.1086/674205 Published online by Cambridge University Press

The two measures Q and S are obviously similar, but they are not ordinally equivalent. While they always agree on which of any pair of propositions is more confirmed by a given piece of evidence, they do not necessarily agree on which of two pieces of evidence provides more confirmation for a given proposition, thanks to the dependence of γ_E on E.

Finally, the measure S fails criteria I and III and does not satisfy Eells and Fitelson's Hypothesis Symmetry.

REFERENCES

Bickel, J. Eric. 2007. "Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules." *Decision Analysis* 4 (2): 49–65.

Carnap, Rudolf. 1962. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

de Finetti, Bruno. 1990. *Theory of Probability: A Critical Introductory Treatment*. Vols. 1 and 2. Trans. Antonio Machi and Adrian Smith. Reprint. London: Wiley.

Eells, Ellery, and Branden Fitelson. 2002. "Symmetries and Asymmetries in Evidential Support." *Philosophical Studies* 107 (2): 129–42.

Fitelson, Branden. 1999. "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity." *Philosophy of Science* 66: S362–S378.

———. 2001. "Studies in Bayesian Confirmation Theory." PhD diss., University of Wisconsin.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78.

Good, Irving John. 1952. "Rational Decisions." *Journal of the Royal Statistical Society*. Series B (Methodological) 14:107–14.

———. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

Hawthorne, James, and Branden Fitelson. 2004. "Discussion: Re-solving Irrelevant Conjunction with Probabilistic Independence." *Philosophy of Science* 71 (4): 505–14.

Joyce, James M. 1998. "A Nonpragmatic Vindication of Probabilism." *Philosophy of Science* 65:575–603.

Keynes, John Maynard. 2004. *A Treatise on Probability*. Mineola, NY: Courier Dover.

Leitgeb, Hannes, and Richard Pettigrew. 2010. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77 (2): 201–35.

Levinstein, Benjamin Anders. 2012. "Leitgeb and Pettigrew on Accuracy and Updating." *Philosophy of Science* 79 (3): 413–24.

Milne, Peter. 1996. "log[$P(h/eb)/P(h/b)$] Is the One True Measure of Confirmation." *Philosophy of Science* 63 (1): 21–26.

Schervish, Mark J., Joseph B. Kadane, and Teddy Seidenfeld. 2003. "Measures of Incoherence: How Not to Gamble If You Must." In *Bayesian Statistics 7: Proceedings of the 7th Valencia Conference on Bayesian Statistics*, ed. J. M. Bernardo et al. Oxford: Oxford University Press.

Winkler, Robert L. 1969. "Scoring Rules and the Evaluation of Probability Assessors." *Journal of the American Statistical Association* 64 (327): 1073–78.

https://doi.org/10.1086/674205 Published online by Cambridge University Press