A Pragmatist's Guide to Epistemic Utility

Benjamin Anders Levinstein*†

We use a theorem from M. J. Schervish to explore the relationship between accuracy and practical success. If an agent is pragmatically rational, she will quantify the expected loss of her credence with a strictly proper scoring rule. Which scoring rule is right for her will depend on the sorts of decisions she expects to face. We relate this pragmatic conception of inaccuracy to the purely epistemic one popular among epistemic utility theorists.

1. Introduction. Accuracy is an important epistemic good. Indeed, according to accuracy-first epistemology, accuracy is the only epistemic good. The higher your credences in truths and the lower your credences in falsehoods, the better off you are, all epistemic things considered. Given this alethic monism, recent proponents of accuracy-first epistemology argue for a variety of epistemic norms by co-opting the resources of practical decision theory, with inaccuracy playing the role of epistemic disutility.[1] For instance, Joyce (1998, 2009) argues that agents should have credences that obey the axioms of the probability calculus by appeal to the decision-theoretic norm of dominance avoidance. On Joyce's favored measures of inaccuracy, any credence function that is not probabilistically coherent will be less accurate than some fixed probabilistically coherent alternative function at every world.[2]

Received April 2016; revised December 2016.

*To contact the author, please write to: Department of Philosophy, 1 Seminary Place, Rutgers University, New Brunswick, NJ 08901; e-mail: balevinstein@gmail.com.

†Thanks to Seamus Bradley, Catrin Campbell-Moore, Greg Gandenberger, James Joyce, Richard Pettigrew, Patricia Rich, and audiences in Bristol and Munich. I was supported by the European Research Council (ERC) starting grant Epistemic Utility Theory: Foundations and Applications during some of the work on this article. I received funding from the ERC under the European Union's Horizon 2020 research and innovation program (grant 669751).

1. Recent examples of the epistemic utility approach include Joyce (1998, 2009), Leitgeb and Pettigrew (2010a, 2010b), Pettigrew (2016a), and Konek and Levinstein (2017).

2. Other decision-theoretic norms appealed to include minimizing expected inaccuracy to establish conditionalization (Greaves and Wallace 2006; Leitgeb and Pettigrew 2010b),

Philosophy of Science, 84 (October 2017) pp. 613–638. 0031-8248/2017/8404-0001$10.00
Copyright 2017 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

Unlike with traditional Dutch-book arguments, the appeal to accuracy considerations appears to be nonpragmatic. Joyce claims that his argument for probabilism brings in no practical considerations whatever and is instead purely epistemic. Indeed, epistemic utility theory (i.e., this decision-theoretic, accuracy-first approach to epistemology) tries to eschew pragmatic considerations entirely.

Such philosophical scruples lead to two difficult problems. First, as we will see, Joyce's argument and the arguments of other epistemic utility theorists only work for a certain class of measures called *strictly proper scoring rules*.[3] This class excludes some extremely natural measures, and it is hard to see why only those measures of inaccuracy are legitimate. Second, if inaccuracy measures are to play the role of epistemic disutility functions for rational agents, it is not clear how to determine which particular measure is right for which agent.[4] It is doubtful that any intuitive notion of accuracy could render one measure objectively correct for all agents, and it is also hard to see what reasons an agent would have to choose one measure over another.

We will provide an answer to both these questions below but from a starting point anathema to a pure epistemic utility theorist. Avoiding appeal to intrinsic epistemic goodness entirely, we will assume that all value is ultimately grounded in practical value. In particular, credences have value based on their connection to practical success. For us, the first question is how a practically rational agent goes about assigning value to her own credences and the credences of others (i.e., how does she assign value to doxastic states?).

One initial advantage we have over the pure epistemic utility theorist is that we can assume such an agent will be probabilistically coherent, for otherwise she's vulnerable to Dutch books. Indeed, we assume such an agent will be an expected utility maximizer.

From this starting point of expected utility maximization, we can understand accuracy's practical role by repurposing a representation theorem from Schervish (1989). Here is the idea in brief. Suppose you have a credence of .3 that it will rain. You may end up having to make a decision at some point on the basis of this credence, such as whether to bring an umbrella, whether to drive instead of walk, or whether to accept a monetary bet that pays off just in case it in fact rains. You do not yet know for sure which particular decisions you will have to make, but you do know that the less accurate your

minimax to establish the principle of indifference (Pettigrew 2016b), Hurwicz criteria to make sense of Jamesian epistemology (Pettigrew 2016c), and chance-dominance avoidance to establish the Principal Principle (Pettigrew 2013).

3. In fact, a few other structural restrictions are needed as well. For details, see Pettigrew (2016a).

4. Some epistemic utility theorists will see this issue as less important than the first, but others see it as necessary at least for the argument for probabilism. See sec. 2.2.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

credence is, the more likely it is that you will make what turns out to be the wrong decision (relative to your desires). So, you can assign your credence an expected loss (i.e., negative expected utility) by averaging over the values of the possible good and bad decisions you might make based on it. As we will see from Schervish's theorem, under some natural assumptions, this method generates exactly the sort of measures of inaccuracy that epistemic utility theorists find acceptable. That is, the expected loss function a rational agent uses to assign practical value to her own credence or to evaluate another agent's credence simply is a proper scoring rule. Moreover, Schervish's theorem will allow us to represent an agent with a single measure of inaccuracy that reflects her expectations of the kinds of practical decisions she will make.

Although our measures of inaccuracy will ultimately be generated from practical considerations, they are nonetheless in a derivative sense epistemic. They reflect an agent's valuation of her credence before she has any particular purpose for it in mind (i.e., before she knows which decisions she will end up making). This allows us to treat epistemic value as quasi-separable from practical value since we do not need to reference any specific practical decision when evaluating how well-off an agent is epistemically. We are in agreement with the pure epistemic utility theorist that the only epistemic good is accuracy as measured in accord with a proper scoring rule. We simply disagree about the ultimate source of this value.

This practical approach to epistemic utility will give us a further advantage over the pure epistemic utility theorist. Because inaccuracy measures function, for us, as summary statistics of expected practical disutility, we can use them to explain an agent's *practico-epistemic* behavior—practical actions that are performed for the sake of epistemic gain, such as evidence gathering, paying for information, and conducting experiments. Understanding this sort of behavior is extremely important to epistemology but nonetheless falls outside of the domain of the purely epistemic.[5]

To be clear, this practical approach does not entail that the project of the pure epistemic utility theorist is doomed. Despite the current difficulties, there may well be a satisfactory account of why proper scoring rules are the only reasonable measures of epistemic utility that do not invoke any practical considerations whatever. The point instead is to investigate the valuation of doxastic states from a practical perspective and to see why and how pragmatically rational agents will use proper scoring rules for such a valuation.

Indeed, epistemic utility theorists themselves may still find this discussion of interest even if they reject the practical foundations. In addition to the independent usefulness of the technical methods used for generating

5. See Gibbard (2007) for another approach that aims to make sense of accuracy in terms of its consequences for practical success.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

measures of inaccuracy, the relationship between what epistemic utility theorists claim is accuracy's purely epistemic value and its ultimate practical value should concern them, especially when it comes to practico-epistemic behavior.

Moreover, many philosophers will be drawn to the claim that the practical value of accuracy is the primary or even sole value of accuracy and that there is no such thing as purely epistemic utility. For instance, functionalists think that we simply cannot divorce doxastic states entirely from their effects on our behavior. Beliefs only make sense insofar as they interact with desires to produce action. Such philosophers will then generally prefer pragmatic arguments for epistemic norms (such as Dutch books) and will likewise prefer a pragmatic basis for valuing accuracy. Like the epistemic utility theorists, however, they too should be interested in understanding the notion of accuracy and how it relates to practical success.

So, although we here appeal ultimately only to pragmatic instead of pure epistemic value, our approach will also have significant payoffs. These, in brief, include

1. A new justification of the standard measures of inaccuracy.
2. A new explanation of why and when to use one measure over another.
3. A better understanding of the connection among accuracy, practical success, epistemic evaluation, and practico-epistemic behavior such as evidence gathering.

Section 2 introduces the basic tools for measuring inaccuracy and the difficulties of epistemic utility theory. Section 3 explains how to determine the practical value of a credence, presents Schervish's theorem, and discusses its significance. Section 4 briefly relates Schervish's theorem to the value of information, evidence gathering, and evaluation of other agents. Section 5 wraps up.

2. Inaccuracy and Scoring Rules. In this section, we look at the two important questions identified in the introduction: what are the general constraints on plausible candidate measures of inaccuracy, and which measure in particular is right in a given context? We will approach these questions for now from the point of view of the epistemic utility theorist. That is, we will see how we might try and answer them if we want a purely epistemic notion of inaccuracy.

Let us start with credences in individual propositions. A measure of inaccuracy, or scoring rule, is meant to quantify how close a credence in a proposition is to its truth-value at a world. At the very least, a higher credence in a true proposition should not count as more inaccurate than a lower credence in that same proposition. We can use this minimal constraint to define the class of functions of interest:

https://doi.org/10.1086/693444 Published online by Cambridge University Press

Definition 1 A function $G:[0, 1] \times \{0, 1\} \to [0, \infty]$ is a *(local) scoring rule* if $G(x, 1)$ is monotonically decreasing and $G(x, 0)$ is monotonically increasing.

Often we will write a scoring rule G as (g_1, g_0) where $g_i(x) = G(x, i)$. By requiring $g_1(x)$ and $g_0(x)$ to be monotonically decreasing and increasing respectively, we guarantee that as a credence gets closer to a truth-value, its score will not get worse. Note that for now, we do not even require the monotonicity to be strict.

We can generalize this idea to constrain good measures of inaccuracy for entire credence functions. Let Ω be a finite set of worlds and $\mathcal{F}$ be a subset of the power set of Ω; $\text{bel}(\mathcal{F})$ is the set of belief functions over $\mathcal{F}$, where a belief function assigns some number $x \in [0, 1]$ to each proposition in $\mathcal{F}$. Note that probability functions form a subset of the belief functions. For $w \in \Omega$ and $X \in \mathcal{F}$, let $w(X) = 1$ if $w \in X$ and $= 0$ otherwise. To constrain the class of relevant functions, we first define an analogous weak monotonicity constraint:

Definition 2 A function $G: \text{bel}(\mathcal{F}) \times \Omega \to [0, \infty]$ is *weakly truth-directed* if for any $b, c \in \text{bel}(\mathcal{F})$, if $|w(X) - b(X)| \leq |w(X) - c(X)|$ for every $X \in \mathcal{F}$, then $G(b, w) \leq G(c, w)$.

Weak truth-directedness says that if b's credence is always at least as close to the truth as c's credence, then b is no more inaccurate than c. We then say:

Definition 3 A function $G: \text{bel}(\mathcal{F}) \times \Omega \to [0, \infty]$ is a *(global) scoring rule* if it is weakly truth-directed.

These definitions of scoring rules are too weak to carve out a good class of inaccuracy measures, but they will be useful below. The most obvious way to strengthen them is to require stronger monotonicity conditions. We say:

Definition 4 A function $G:[0, 1] \times \{0, 1\} \to [0, \infty]$ is a *(strict local) scoring rule* if $G(x, 1)$ is strictly decreasing and $G(x, 0)$ is strictly increasing.

Likewise, we define a stronger notion of truth-directedness:

Definition 5 A function $G: \text{bel}(\mathcal{F}) \times \Omega \to [0, \infty]$ is *truth-directed* if for any $b, c \in \text{Prob}(\mathcal{F})$, if

1. $|w(X) - b(X)| \leq |w(X) - c(X)|$ for every $X \in \mathcal{F}$ and
2. $|w(X) - b(X)| < |w(X) - c(X)|$ for some $X \in \mathcal{F}$,

then $G(b, w) < G(c, w)$.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

In turn,

> **Definition 6** A function $G : \text{bel}(\mathcal{F}) \times \Omega \to [0, \infty]$ is a *(strict global) scoring rule* if it is truth-directed.

2.1. Propriety. Any strict scoring rule is in some sense a measure of inaccuracy. However, from the point of view of the epistemic utility theorist, even some strict scoring rules fail to generate the results she wants.

Joyce (1998, 2009), for instance, argues that epistemic agents should be probabilistically coherent. In schematic terms, Joyce's argument runs as follows. There is some class $\mathcal{I}$ of reasonable measures of inaccuracy. For any scoring rule $G \in \mathcal{I}$ and any nonprobabilistic belief function b, there exists an alternative probability function c that is less inaccurate than b according to G at every possible world. Furthermore, according to G, for any probabilistically coherent function c and any belief function b, c is less inaccurate than b at some world. In other words, all and only the nonprobability functions are dominated according to every reasonable measure of inaccuracy.

The major weakness of this argument is that some measures of inaccuracy that seem perfectly reasonable do not yield this result. Consider the absolute-value measure, for instance:

$$\text{abs}(b, w) = \sum_{X \in \Omega} |b(X) - w(X)|.$$

Here, abs is clearly truth-directed and at least seems natural. Nevertheless, it yields absurd verdicts about the relative accuracy of two belief functions. Imagine an urn contained a red, a green, and a blue ball, one of which will be drawn at random (i.e., with a 1/3 chance). According to abs, an agent with a credence of 0 in red, green, and blue counts as less inaccurate than an agent with a credence of 1/3 in each proposition, regardless of which ball is actually drawn.[6]

Surprisingly, it is relatively easy to identify exactly what further major restriction on measures of inaccuracy is needed to generate Joyce's results: every probability function must assign itself minimum expected inaccuracy.[7]

6. Note that the agent with a credence of 0 in each proposition will receive a total score of 1, since her credence in two of the propositions will be perfectly accurate, while her credence in one proposition will be off by 1. The agent with a credence of 1/3 in each proposition will be off by 2/3 in one proposition and by 1/3 in the remaining two, for a total score of 4/3.

7. Joyce (2009) himself derives propriety from truth-directedness along with the weaker principle of Coherent Admissibility, which requires every probability function to be nondominated.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

That is, the argument requires scoring rules to be *strictly proper* according to the following definition:

> **Definition 7** A scoring rule G is a *proper scoring rule* if for all probability functions b and belief functions c, $E_b(G(c))$ is minimized at $b = c$, where E_b denotes b's expectation function. If this minimum is unique, then G is a *strictly proper scoring rule*. If G is proper but not strictly proper, then we say that G is a *merely proper scoring rule*.

Note that if G is a local scoring rule, then G is (strictly) proper if for all $x, y \in [0, 1]$, $yg_1(x) + (1 - y)g_0(x)$ is (uniquely) minimized at $x = y$.

Propriety is a curious property. On the one hand, it is a crucial constraint necessary for the success of the epistemic utility program. In addition to Joyce's argument, nearly every other argument in the epistemic utility literature requires this restriction as well.[8] Without it probabilistically coherent credence functions would be self-undermining. That is, they would face a kind of Moorean paradox: 'I assign credence x to X, but I think a credence of x' in X would be more/at least as accurate.' On the other hand, it seems hard to justify on the basis of reflection on the notion of inaccuracy alone. Propriety simply does not seem to stem from alethic monism on its own.[9] Furthermore, propriety rules out two of the most obvious measures of inaccuracy right from the bat, that is, abs and the euclidean measure:

$$\mathrm{euc}(b, w) = \left(\sum_{X \in \mathcal{F}} (b(X) - w(X))^2 \right)^{1/2}.$$

Two of most common measures of distance are abs and euc, and inaccuracy is supposed to be a measure of proximity to truth. Both are truth-directed, yet neither is proper.[10]

Fortunately for the epistemic utility theorist, other scoring rules are relatively natural as well and do turn out to be proper. Three common strictly proper global rules include

8. See, e.g., n. 2.

9. There are a number of arguments that try to independently motivate restrictions on the class of reasonable inaccuracy measurements that entail propriety. Discussing each would substantially lengthen this article, but I refer the interested reader to Joyce (1998), D'Agostino and Sinigaglia (2010), Leitgeb and Pettigrew (2010a), and Pettigrew (2016a). For further doubts about the plausibility of propriety stemming from alethic monism, see Gibbard (2007).

10. We have already seen that abs is improper. To see that euc is improper, suppose an agent assigns credence .9 to X and .1 to $\neg X$. She expects the credence function that assigns 1 to X and 0 to $\neg X$ to be less inaccurate than she is according to euc.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

Brier Score

$$\mathrm{BS}(b, w) = \frac{1}{|\mathcal{F}|}\sum_{X\in\mathcal{F}}(w(X) - b(X))^2.$$

Log Score

$$\mathrm{Log}(b, w) = -\frac{1}{|\mathcal{F}|}\sum_{X\in\mathcal{F}}\ln(|(1 - w(X)) - b(X)|).$$

Spherical Score

$$\mathrm{Sph}(b, w) = \frac{1}{|\mathcal{F}|}\sum_{X\in\mathcal{F}}1 - \frac{|1 - w(X) - b(X)|}{\left(b(X)^2 + (1 - b(X))^2\right)^{1/2}}.$$

Each of these rules is *additive*. That is, each is simply the (normalized) sum of a local strictly proper rule:

Local Brier $\mathrm{BS}(x, i) = (i - x)^2$.
Local Log $\mathrm{Log}(x, i) = -\ln(|(1 - i) - x|)$.
Local Spherical $\mathrm{Sph}(x, i) = 1 - |1 - i - x|/(x^2 + (1 - x)^2)^{1/2}$.

Later on, we primarily focus on local rules and then see how they relate to additive global rules.[11]

So, despite its theoretical importance, propriety itself is in need of some additional explanation. We provide one below—when we walk through Schervish's theorem we will gain a new understanding of what makes proper scoring rules so special. Our solution will not satisfy the austere scruples of those who want inaccuracy to be a purely epistemic notion with no appeal to pragmatic considerations but instead will explain why pragmatically rational agents use them to determine the value of their own credences.

2.2. Which Scoring Rule to Use? A second question is which scoring rule serves as the best measure of inaccuracy in a given context. Even if we insist on strict propriety, we have infinitely many rules left to choose from.

11. The Spherical Score looks odd at first, but it is more natural when understood geometrically. For given credence function c, proposition X, and world w, let $\| c_X \|$ be the length of the vector $c_X = \langle c(X), 1 - c(X) \rangle$. Let $\theta_{X,w}$ be the angle between c_X and $\langle w(X), w(\neg X) \rangle$. The local spherical score of a credence $c(X)$ is then $\| c_X \| \cos \theta_{X,w}$. That is, it is determined by the length of the vector c_X and the angle between c_X and the actual truth-value of X at w. For a more thorough discussion, see Jose (2007).

https://doi.org/10.1086/693444 Published online by Cambridge University Press

It is not immediately clear why someone would opt for the Brier or the Log or the Spherical rule.

Some epistemic utility theorists may regard this question as less pressing. As we saw, Joyce establishes an *accuracy-dominance* argument for probabilism. As long as the scoring rule in question is strictly proper—and meets a few other structural assumptions[12]—all and only probability functions are undominated, so it appears in this case that there is no need to choose any single measure. However, as Bronfman (2009) and Pettigrew (2016a, chap. 5) point out, which functions dominate which others is scoring-rule dependent. In particular, if b is not a probability function, then there may be no probability function that dominates it on every rule that is considered legitimate. So, if an agent adopts b as her credence function but does not adopt any particular measure of inaccuracy, then any probability function c will do worse than b at some world according to some measure. Both authors argue that if we do not choose a single rule with which to measure an agent's inaccuracy, then the normative force of accuracy-dominance arguments for probabilism is undermined.[13]

In response, one may be a subjectivist and claim that the scoring rule merely reflects an agent's subjective epistemic values, just as in practical contexts rational agents may adopt alternative credence functions.[14] One may also be an objectivist and claim that a single rule is correct.[15]

Schervish's theorem will enable us to provide a new kind of answer. An agent's scoring rule will not reflect her epistemic values, but instead it will represent the kinds of decision problems she expects to face. In full generality, any proper scoring rule could be correct in a given context. Furthermore, an agent's global scoring rule will usually be built out of different local scoring rules for different propositions.

3. The Pragmatic Evaluation of Credences. Let us now put aside this notion of pure epistemic utility unsullied by practical value and return to the

12. Namely, as long as the rule is truth-directed, continuous, strictly proper, and additive (i.e., the sum of local scoring rules), the result that all and only probability functions are undominated goes through.

13. I harbor doubts as to whether this objection is actually successful, but I mention it here to note that epistemic utility theorists themselves consider this issue an important problem. It is worth acknowledging as well that the class of admissible measures of inaccuracy need not be narrowed all the way down to a singleton to avoid the Bronfman objection.

14. Joyce (2009) at least leans in this direction.

15. This position is perhaps the most popular among epistemic utility theorists, with the Brier usually being the rule of choice (Rosenkrantz 1981; Leitgeb and Pettigrew 2010a; Pettigrew 2016a).

https://doi.org/10.1086/693444 Published online by Cambridge University Press

world of hard-nosed pragmatism. We wish now to understand how a practically rational agent will evaluate her own doxastic state and possible alternative doxastic states. For instance, we will try to determine how much expected utility an agent assigns her credence of .6 that it will rain.

Our task is a bit easier than the epistemic utility theorist's, as we already have some understanding of practical rationality. We will assume that practically rational agents are *expected utility maximizers*.[16] In particular, they have probabilistically coherent credence functions, since otherwise they would be subject to Dutch books. This starting point will give us some initial traction.

We make a few additional assumptions. First, unlike in causal or evidential decision theory, we will only look at situations in which acts and states are independent. That is, in the situations we consider below, whether an agent performs an action has no bearing on whether an event of interest occurs. For instance, whether you bring an umbrella does not by itself (at least normally) affect your credence that it will rain. Because the actions do not affect outcome, we will often refer to actions as 'bets'.

This may seem unduly restrictive, but we are interested in evaluating credences in propositions, not credences conditionalized on or imaged on the performance of action. Your credence that you will get a promotion is different from your credence that you will get a promotion supposing you bribe your boss, and the two in turn have different values.

Second, we will assume that credences and states are independent. That is, the probability of events of interest does not depend on an agent's credences. For example, the chance a coin will land heads will not be affected by your belief that the coin will land heads. Third, we assume that the value of an outcome is not itself affected by an agent's credence. In other words, agents do not themselves assign direct value to the beliefs they hold. For example, we will not try to account for the utility you gain from your high credence that your colleagues are fond of you. These last two assumptions are for the sake of simplification.

3.1. The Practical Value of a Credence. With this background out of the way, let us now see how an agent may evaluate her own credence in terms of expected practical value. Let R be the proposition that it will rain today. There are a number of different bets on R that an agent, let us call her

16. In particular, I assume that agent's doxastic states are (or are representable by) a unique probability function and that she has a utility function that is unique up to positive affine transformation. Both of these idealizations are necessary for Schervish's theorem to generate a unique scoring rule. An important question that I will not explore here is what happens when these assumptions are relaxed.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

Alice, might take that will affect her utility. Suppose the possible actions are bringing an umbrella (u), wearing a raincoat (w), or staying home (s).

Suppose we want to know whether Alice will bring an umbrella. Given that she is an expected utility maximizer, she will only if

$$\mathrm{EU}(u) \geq \max(\mathrm{EU}(w), \mathrm{EU}(s)) = \mathrm{EU}(\neg u),$$

where EU is Alice's expected utility function. In other words, we can see whether she will bring an umbrella by looking at her decision between two actions, bringing an umbrella or not bringing an umbrella, even though the action space itself is more fine grained. This will allow us to treat each of Alice's decision problems as if there were only two options from now on: whether to u or $\neg u$.[17]

Imagine Alice's payoff matrix is as given in table 1, which shows how much utility Alice gets, depending on whether she brings an umbrella when it rains or does not rain. By a simple calculation of Alice's expected utility, we can determine how high her credence x in R must be before she decides to bring an umbrella.

$$\begin{aligned} \mathrm{EU}(u) &= x(-1) + (1 - x)(-2) \\ &= x - 2. \end{aligned} \tag{1}$$

$$\begin{aligned} \mathrm{EU}(\neg u) &= x(-4) + (1 - x)(0) \\ &= -4x. \end{aligned} \tag{2}$$

We then have (1) > (2) if and only if $x > 2/5$. So, Alice will bring an umbrella if $x > 2/5$ and not bring an umbrella if $x < 2/5$. For ease, we will conventionally decide that Alice will bring the umbrella if and only if $x > 2/5$.

17. An important issue in decision theory is the relationship between *small-* and *grand-*world decision problems. In small-world problems, an agent does not partition the space of outcomes, states, and acts maximally finely. In the current (small-world) decision problem, for instance, Alice does not distinguish between outcomes in which her umbrella breaks and outcomes in which her umbrella remains in tact, even though those clearly result in different rewards. Ideally, an agent would always deliberate using a maximally fine-grained partition (if such there be), but that requirement is so unrealistic that it would render decision theory of little guiding value. I agree, then, with Joyce (1999) that when an agent deliberates using a small-world partition and selects action a from that set of actions, she is committed to the view that her "fully considered beliefs and desires would sanction the choice of a from among the alternatives listed" (74). In other words, "we can think of a rational agent's attitudes toward the states, outcomes, and acts in a small-world decision problem as *her best estimates* of the attitudes that she would hold regarding those states, outcomes, and acts in the grand-world context" (75, emphasis mine).

https://doi.org/10.1086/693444 Published online by Cambridge University Press

TABLE 1. PAYOFF MATRIX

	R	$\neg R$
u	-1	-2
$\neg u$	-4	0

All that matters for how much utility Alice ends up getting is (i) whether R and (ii) whether her credence is greater than 2/5.

3.2. Reformulation. We need to reformulate this problem to suit our aims of constructing proper scoring rules. Since scoring rules are loss functions, we will describe Alice as an expected loss minimizer instead of as an expected utility maximizer. This choice is purely conventional. Table 2 reexpresses table 1 in terms of losses instead of gains.

Notice that if it rains, Alice is sure to incur a loss of at least 1 no matter what she does. As far as her decision is concerned, this loss is irrelevant since it is merely a result of the state of the world. So, we can normalize table 2 by subtracting the minimum loss that is sure to result at each state. We then arrive at table 3.

Finally, we rewrite table 3 as table 4 by dividing the loss in each cell by the sum of the total losses in each cell. In this case, the sum is 5, since Alice will lose 2 if she performs u and $\neg R$ and will lose 3 if she performs $\neg u$ and R.

When considering this problem in isolation, we can forget about the sum of the losses. It represents the "stakes" of the problem, but it will not affect Alice's decision. So, for now, we can rewrite the payoff matrix as table 5, which captures this single decision problem conspicuously. As before, all that matters for much (dis)utility Alice ends up getting is (i) whether R and (ii) whether her credence is greater than 2/5.

3.3. Scoring Rules and the Problem of the Umbrella. We can now design a scoring rule that will track how much of a loss (under our normalization) a credence of x in R will bring Alice. That is, given that her credence is currently x, we determine how much she expects to lose from her bet on rain.

TABLE 2. LOSS MATRIX

	R	$\neg R$
u	1	2
$\neg u$	4	0

https://doi.org/10.1086/693444 Published online by Cambridge University Press

TABLE 3. Loss Matrix, First Normalization

	R	$\neg R$
u	$0 = 1 - 1$	$2 = 2 - 0$
$\neg u$	$3 = 4 - 1$	$0 = 0 - 0$

First, let us determine how much she would lose if R and if $\neg R$, respectively. Per table 5, If R and $x \leq 2/5$, Alice will lose 3/5, since she will not bring an umbrella. If $x \geq 2/5$ and $\neg R$, she will lose 2/5, since she will bring an umbrella. Otherwise she loses nothing.

Consider $G = (g_1, g_0)$, where

$$g_1(x) = \begin{cases} 3/5 & \text{if } x \leq 2/5 \\ 0 & \text{if } x > 2/5, \end{cases}$$

and

$$g_0(x) = \begin{cases} 0 & \text{if } x \leq 2/5 \\ 2/5 & \text{if } x > 2/5. \end{cases}$$

Given Alice's credence and R's truth-value, G returns the amount Alice will lose.

Now suppose Alice wishes to evaluate a credence of y given her credence x. That is, she wants to determine how much she would expect to lose if she had decided whether to bring an umbrella on the basis of a credence of y in R.

We then have

$$\begin{aligned} \mathrm{E}_x(G(y)) &= x \times g_1(y) + (1 - x)g_0(y), \\ &= \begin{cases} x \times 3/5 & \text{if } y \leq 2/5 \\ (1 - x) \times 2/5 & \text{if } y > 2/5, \end{cases} \end{aligned}$$

where $\mathrm{E}_x(G(y))$ is minimized exactly when

$$x, y \leq 2/5$$

or

TABLE 4. Loss Matrix, Second Normalization

	R	$\neg R$
u	0	$(2/5) \times 5$
$\neg u$	$(3/5) \times 5$	0

https://doi.org/10.1086/693444 Published online by Cambridge University Press

TABLE 5. Loss Matrix, Stake-Free Version

	R	$\neg R$
u	0	2/5
$\neg u$	3/5	0

$$x, y > 2/5.$$

So, G is a merely proper scoring rule.

3.4. The General Case. More schematically, we can represent the agent as choosing between two options, d_1 and d_0, where d_1 is better to perform if P and d_0 is better to perform if $\neg P$. That is,

$$\mathrm{L}(d_1, P) \leq \mathrm{L}(d_0, P)$$
$$\mathrm{L}(d_1, \neg P) \geq \mathrm{L}(d_0, \neg P).$$

We again normalize losses in the same way we did in table 3 by setting

$$L(d_1, P) = L(d_0, \neg P) = 0.$$

Then, in accord with table 4, we can now construct a loss matrix as expressed in table 6, where $q \in [0, 1]$ and $W \in (0, \infty]$. Alice's cutoff for performing d_1 is represented by q. That is, if and only if Alice's credence in $P < q$ will she perform d_1. Otherwise, she will perform d_0. The *weight* or stakes of the problem is represented by W. Again, we ignore W for now and focus only on the cutoff points for deciding whether to d_1.

Definition 8 A *q-problem with respect to P* is a two-decision problem such that $\mathrm{L}(d_1, \neg P) = W \times q$.

For any particular q, Alice sees no difference in expected value between two forecasts on the same side of q since those forecasts will lead to the exact same action.

In this more general case, we can determine Alice's valuation of a credence x with the merely proper scoring rule $G = (g_1, g_0)$:

TABLE 6. Loss Matrix, Second Normalization

	P	$\neg P$
d_1	0	$q \times W$
d_0	$(1 - q) \times W$	0

https://doi.org/10.1086/693444 Published online by Cambridge University Press

$$g_1(x) = \begin{cases} 1 - q & \text{if } x \leq q \\ 0 & \text{if } x > q, \end{cases}$$

and

$$g_0(x) = \begin{cases} 0 & \text{if } x \leq q \\ q & \text{if } x > q. \end{cases}$$

3.5. Uncertainty about the Bet. We are often uncertain what bets we are actually going to face in the future. Supposing you are going to bet on P, you may still be uncertain whether you will face a q- or q'-problem. For instance, Alice may know she will be offered some bet that will return \$1 if it rains, and \$0 otherwise, but not yet know what price the bookie will offer her.

We will handle the more general case in a moment, but for now assume that there is some finite set $\mathcal{Q} \subset [0, 1]$ such that Alice has credence 1 that she will face some q-problem, where $q \in \mathcal{Q}$. We will treat Q as a random variable representing the q-value of the decision problem Alice faces and use $\Pr(q_0)$ as an abbreviation for $\Pr(Q = q_0)$.

So, if Alice is uncertain about the value of Q, what expected loss does she assign her credence x in P? Ignoring the stakes, we find

$$h_1(x) := \mathrm{EL}(x|P) = \sum_{\substack{q \in \mathcal{Q} \\ x \leq q}} (1 - q) \times \Pr(q), \tag{3}$$

$$h_0(x) := \mathrm{EL}(x|\neg P) = \sum_{\substack{q \in \mathcal{Q} \\ q < x}} q \times \Pr(q), \tag{4}$$

where h_1 represents Alice's expected loss of having credence x conditional on P, and h_0 represents Alice's expected loss of having credence x conditional on $\neg P$. That is, h_i represents Alice's expected loss given the truth-value of P but with the q-value of the bet still undetermined.

We calculated h_1 as follows: if P is the case and Alice's credence x turns out to be less than or equal to Q, she will lose $1 - Q$. If her credence turns out to be greater than Q, she will not lose anything. So, h_1 just is the sum of $1 - q$ discounted by Alice's credence $\Pr(Q = q)$ for every $q \geq x$. The formula for h_0 is determined similarly.

We can then determine Alice's unconditional expected loss of credence x in P by

https://doi.org/10.1086/693444 Published online by Cambridge University Press

$$\mathrm{EL}_x(x) = x \times h_1(x) + (1 - x)h_0(x),$$

where $\mathrm{EL}_x(x)$ is the loss Alice currently expects to suffer from her credence of x before she learns whether P is true or false. More generally, we can calculate Alice's unconditional expected loss of alternative forecast y in P by

$$\mathrm{EL}_x(y) = x \times h_1(y) + (1 - x)h_0(y),$$

where $\mathrm{EL}_x(y)$ is the expected loss Alice assigns to using another forecast instead of her own, but with her utility function held fixed. It is easy to check that $H = (h_1, h_0)$ is in fact a merely proper scoring rule.

3.6. Letting the Stakes Count. As we noticed, two q-problems can have very different stakes. The same hand in blackjack matters a lot more at the \$1,000 table than the \$1 table, even though the probabilities remain the same.

Furthermore, an agent may expect that if she faces a q-problem it will present her with stakes different from a q'-problem. Suppose, for instance, she is 50% confident she will face the low-stakes problem in the top panel of table 7 and 50% confident she will face the high-stakes problem in the bottom panel of table 7. Because the latter problem has much higher stakes, having her credence on the right side of 2/3 is significantly more important than having her credence on the right side of 1/2. In general, for any given q, the expected value of the stakes W can vary.

Let $\mathrm{E}(W|q)$ be the expected value of W given that Alice faces a q-problem. Letting g_1 represent the expected loss of credence x conditional on P and g_0 be the expected loss of credence x conditional on $\neg P$, we have

$$g_1(x) = \sum_{\substack{q \in \mathcal{Q} \\ x \leq q}} (1 - q) \times \mathrm{E}(W|q) \times \Pr(q), \tag{5}$$

$$g_0(x) = \sum_{\substack{q \in \mathcal{Q} \\ q < x}} q \times \mathrm{E}(W|q) \times \Pr(q). \tag{6}$$

TABLE 7. q-Problems with Different Stakes

	P	$\neg P$
Low stakes:		
d_1	0	1/2
d_0	1/2	0
High stakes:		
d_1^*	0	(2/3) × 15
d_0^*	(1/3) × 15	0

https://doi.org/10.1086/693444 Published online by Cambridge University Press

The overall expected loss Alice assigns forecast y if her credence is x is then

$$\mathrm{EL}_x(y) = x \times g_1(y) + (1 - x) \times g_0(y). \tag{7}$$

Again, G is a merely proper scoring rule. Let $\mathcal{Q} = \{q_1, \ldots, q_n\}$. If Alice's credence is x where $q_i < x \leq q_{i+1}$, then she assigns any y in the same interval the same expected loss she assigns herself. Otherwise, she assigns y a greater expected loss.

To determine G, what matters then is both how high the stakes are expected to be given that a q-problem is faced and how likely Alice thinks it is that she will face a q-problem. That is, what matters is the quantity

$$M(q) := \mathrm{E}(W|q) \times \Pr(q).$$

We can then reformulate equations (5) and (6) above as

$$g_1(x) = \sum_{\substack{q \in \mathcal{Q} \\ x \leq q}} (1 - q) \times M(q), \tag{8}$$

$$g_0(x) = \sum_{\substack{q \in \mathcal{Q} \\ q < x}} q \times M(q), \tag{9}$$

where $M(q)$ measures how important Alice thinks q-problems are to get right. Note that (i) $M(q) \geq 0$ and (ii) $M(q) = 0$ if and only if $\Pr(q) = 0$. Neither of these facts is surprising: it is never good to get a q-problem wrong, and if you think you might face a q-problem, then getting it right matters at least a little bit.

3.7. The Continuous Case. So far, we have assumed that q is in some finite set $\mathcal{Q} \subset [0, 1]$ to keep things discrete. However, Alice could potentially face a q-problem for any $q \in [0, 1]$. The main difficulty is that once we make this generalization, the probability of any particular q-problem is 0.[18]

The natural way to handle this problem is to trade out $M(q)$ for a function $m(q)$ that measures the *probability density* of facing a q-bet factored by the expected stakes of that bet. That is,

$$m(q) := p(q)\mathrm{E}(W|q),$$

where $p(q)$ is the probability density Alice assigns to the claim that she will face a q-problem. We call such an m Alice's *support function*. The support

18. With the exception of at most countably many values of Q.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

function measures how much relative importance is assigned to each possible value of Q.[19]

To get a handle on m, note that (i) $m(q) \geq 0$ and (ii) m is constantly 0 over some region $[\alpha, \beta]$ just in case Alice is certain that she will not face a q-problem for any $q \in [\alpha, \beta]$. As with M, the first condition reflects the fact that it is never good to get a q-problem wrong. The second condition means that if Alice thinks it is possible she will face a q-problem for q in some region, then getting such a problem right matters at least a little bit.

By switching out M for m and the sums for integrals in equations (8) and (9), we can generate scoring rules when $\mathcal{Q} = [0, 1]$. With some reformulation and loss of generality, Schervish's theorem is then:

Theorem 1 (Schervish). Let $m(q)$ be a support function, and let

$$g_1(x) = \int_x^1 (1 - q)m(q)\mathrm{d}q,$$

$$g_0(x) = \int_0^x q \times m(q)\mathrm{d}q.$$

Then $G = (g_1, g_0)$ is a proper scoring rule. If m is strictly positive almost everywhere, then G is a *strictly* proper scoring rule.

With some generalization, this method gets us every proper and strictly proper scoring rule, aside from uninteresting ones that, for example, assign every region infinite importance.

Note that the condition that m be positive almost everywhere is a kind of *regularity* condition. Alice will set $m(q) = 0$ if and only if she is certain she will not face a q-problem for that q-value. So, if she leaves open the possibility (however remote) of facing any q-problem whatever, then she will use a strictly proper scoring rule to evaluate her credence.

3.8. Examples. Above, we defined the local versions of the Brier Score, Log Score, and Spherical Score. Let us see how these each encodes very different expectations about the bets Alice may face.

If Alice assigns constant weight to every point, we have

$$m(q) = c,$$

which generates the Brier Score (times $c/2$).

19. For theories that allow $\mathrm{E}(W|q)$ to be defined even when $\Pr(q) = 0$, see Rényi (1955) and Popper (1959).

https://doi.org/10.1086/693444 Published online by Cambridge University Press

If Alice cares about points in [0,1] closer to 0 or 1 more than she cares about other points, she might set

$$m(q) = \frac{1}{q \times (1 - q)},$$

which generates the Log Score.

If Alice weights points near .5 more than she weights other points, she may set

$$m(q) = \frac{1}{(2q^2 - 2q + 1)^{3/2}},$$

which generates the Spherical Score.

Figure 1 provides a visual representation of each of these scoring rules along with their corresponding support functions.

The Brier Score is the most egalitarian of all scoring rules in terms of the decisions the user expects to make. Imagine you knew you were going to be offered a bet on X that paid \$2 if X and \$0 otherwise. The price of the bet will be chosen at random (i.e., by the uniform distribution on [0,2]). In is case, the Brier Score is the right scoring rule. The stakes of the bet are constant; that is, $\mathrm{E}(W|q) = 2$ for all q. Furthermore, since the value of $\mathcal{Q}$ follows a uniform distribution, $p(q) = 1$ for all q. So, $m(q) = 2$.

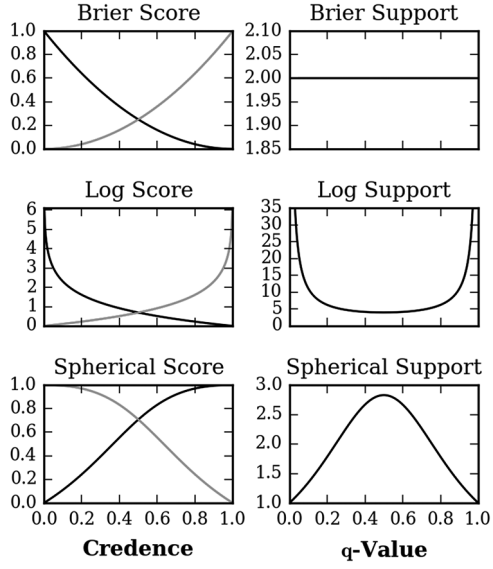


Figure 1. Brier, log, and spherical scores along with their support functions. *Left*, ascending curves represent $g_0(x)$ and descending curves represent $g_1(x)$ for the respective scoring rules. Color version available as an online enhancement.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

The Logarithmic Score is approximately right when, in expectation, bets will be concentrated near the end points of the unit interval and when the stakes are high near those points. The Spherical Rule is best when the success of your decisions will likely depend on correct unbiased binary classification, that is, guessing whether X or $\neg X$ is true depending on whether your credence in X is greater or less than .5.

In real life, of course, our views over the bets we will face on propositions are a lot messier. Suppose I am wondering whether my house will burn down in the next year. I do not yet know exactly which bet I will face on that proposition, but in expectation a credence of .01 is very different from a credence of .0001. In the former case, I will likely pay quite a bit for an insurance policy. So, relatively close to 0, my loss function will behave a lot like the Log Rule. However, a credence of 10^{-6} and a credence of 10^{-7} in this same proposition are roughly equivalent as far as my real-world success is concerned. Despite the value of my house, if my credence is low enough that it will burn down, I am effectively morally certain that it will not burn down. I assign, at some point, negligible practical weight to the possibility that I will wrongly bet that it will remain in tact for the next year. So the Log Rule will not be a perfect fit.

Similarly, the Spherical Score is a good approximation of the right score for my credence in whether it will rain. Whether I make the right decision about bringing an umbrella, buying baseball tickets in advance, or canceling my picnic hinges on whether my credence is above or below (approximately) .5. I doubt any decision I make will depend on whether my credence is on the correct side of .05 or .95. At that point, I am effectively certain it will not (will) rain as far as my decision making goes. Therefore, the correct support function will place more weight on middling regions of the unit interval and less weight on extreme regions.

3.9. Global Scoring Rules. Let us now briefly turn our attention to scoring global credence functions. The idea is the same: the expected score of an entire credence function is the expected loss an agent would incur from using that credence function to make bets.

The score of an individual credence x in proposition X at a world is the expected loss. That is, the score measures the expected loss given the actual truth-value of X before it is known exactly which bet on X the agent will face. Likewise, when we score an entire credence function b at a world, we plug in the actual truth-values of the propositions in $\mathcal{F}$, but we do not yet plug in the actual bets the agent faces. Indeed, she may not face a bet on some propositions at all.

The easiest way to determine this global score is simply to identify it with the (normalized) sum of the local scores:

https://doi.org/10.1086/693444 Published online by Cambridge University Press

$$G(b, w) = \alpha \times \sum_{X \in \mathcal{F}} G_X(b(X), w(X)),$$

where α is positive, and G_X is the scoring rule used for proposition X. In general, as we have seen, the scoring rule for the individual propositions will vary heavily, depending on the expected bets the agent may face. Furthermore, because we are much more likely to face bets on some propositions than on others, and because the stakes of those bets will vary heavily, some propositions will count much more to b's overall score at w. For instance, it is much more likely that we will face a bet on whether it will rain than on whether there are an odd number of grains of sand in the Sahara. So, $G(b,w)$ will give extra weight to the former proposition.

For this to work (i.e., for us to be able to simply add up a bunch of local scores) we need a way of localizing bets. That is, given your entire credence function, we need a way to determine which single proposition your decision to perform act u or $\neg u$ is a bet on. You might, for instance, want to bring an umbrella if your credence in rain is high enough. But you may also decide to bring an umbrella if your credence that it will snow is high. So, your decision is not really a bet on rain, nor is it a bet on snow, but instead it is a bet on (rain or snow).

To do so, the best method seems to be to partition the set of states of the world into those in which you would rather perform u and those in which you would rather perform $\neg u$. With your utility assignments fixed, the proposition u is taken to be a bet on the disjunction of all the states in which performing u is in expectation better than performing $\neg u$.[20] In this way, Schervish's method can be used to generate an additive scoring rule that represents the expected loss of an agent's entire doxastic state.

4. The Value of Information. One useful application of scoring rules is that they allow us to quantify the value of information for various propositions. We will now see how this may work in some detail.

20. As a referee points out, this introduces another important grand-world/small-world problem. To see this, suppose Alice would rather bring an umbrella if she learned it was going to rain but would want to leave it home otherwise. That is, on the partition $\{R, \neg R\}$, Alice considers u a bet on R. However, suppose that if it will rain only very lightly (L), she would still rather leave her umbrella at home. So, when she considers $\{R\&L, R\&\neg L, \neg R\}$, she takes u to be a bet on just $R\&\neg L$. Even if all of Alice's preferences are ultimately partition invariant, then, what actions are a bet on which states of the world will vary depending on how the set of states is carved up. In turn, Schervish's method will yield different results depending on this partition. I do not have a general answer to the question of how finely agents need to partition the set of states of the world for the representative scoring rules to reflect their expected losses from their credence function in the best way. However, I lean toward a subjectivist solution—your global score should reflect the finest partition that is cognitively available.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

4.1. Experiments. Suppose Alice is a scientist interested in whether X. She decides to spend some resources to perform an experiment that will provide her with new evidence. She cannot perform every potential experiment, so she will have to make some choices on the basis of the expected informational value of the results. How might she go about deciding?

Following Greaves and Wallace (2006), we call a partition $\mathcal{E}$ of Ω an *experiment* if the agent will learn some element of that partition is true. We say that an experiment $\mathcal{E}$ is performed once the agent learns which element of $\mathcal{E}$ in fact obtains.

For example, suppose Alice is wondering whether her favorite team will win tonight. She knows that if Jake is pitching, there is a .75 chance they will win. Otherwise there is a .25 chance. If she looks at the team's website, she can find out who is pitching. So, in this case, there are four relevant states $\{W \wedge P, W \wedge \neg P, \neg W \wedge P, \neg W \wedge \neg P\}$, where W stands for the proposition that her team wins, and P stands for the proposition that Jake is pitching. She can perform the experiment $\mathcal{E}$ of finding out who is pitching and thereby partition Ω into $P = \{W \wedge P, \neg W \wedge P\}$ and $\neg P = \{W \wedge \neg P, \neg W \wedge \neg P\}$.

If Alice performs $\mathcal{E}$, she will learn either P or $\neg P$. So, Alice's posterior credence in W, denoted $b_{\mathcal{E}}(W)$, will either be $b(W|P) = .75$ or $b(W|\neg P) = .25$, although she does not yet know which. Suppose Alice's current credence in W is .5. How valuable is the experiment $\mathcal{E}$ to her in expectation?

Before performing $\mathcal{E}$, Alice assigns her credence an expected loss of $\mathrm{EG}(.5) = .5(g_1(.5) + g_0(.5))$. If she learns P, she will either get a score of $g_1(.75)$ or $g_0(.75)$. So, her expected score given that she learns P is

$$b(W|P)g_1(.75) + b(\neg W|P)g_0(.75) = .75g_1(.75) + .25g_0(.75)\,.$$

Likewise, her expected score given that she learns $\neg P$ is $.25g_1(.25) + .75g_0(.25)$. Given the constraints we have established, she must also have a .5 credence in P. So, the expected score she currently assigns to her as-yet-unknown posterior credence after performing $\mathcal{E}$ is

$$\mathrm{EG}(b^{\mathcal{E}}(W)) = .375(g_1(.75) + g_0(.25)) + .125(g_0(.75) + g_1(.25)). \tag{10}$$

Greaves and Wallace show that updating by conditionalization is the policy that minimizes expected loss for all proper scoring rules. So, (10) is less than or equal to Alice's current score of EG(.5), and the inequality is strict if G is strictly proper.

From a practical point of view, this result follows from Good (1967), who shows that in any decision context, free information never has negative value in expectation.[21] Indeed, we can quantify the expected value of performing $\mathcal{E}$

21. See Myrvold (2012) for an in-depth discussion of Good's theorem and scoring rules.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

as the difference between Alice's current expected loss and her expected posterior loss, that is, as $\mathrm{Val}(\mathcal{E}) = \mathrm{EG}(.5) - \mathrm{EG}(b^{\mathcal{E}}(W))$. If we identify utility with dollars, Alice should pay up to $\$\mathrm{Val}(\mathcal{E})$ to learn the result of experiment $\mathcal{E}$.[22]

Note that the precise value of an experiment depends on which scoring rule Alice uses for the proposition under investigation. Suppose, for instance, that Alice had the option of performing either $\mathcal{E}$ or an alternative experiment $\mathcal{E}'$ that would reveal who won the game with 20% probability and would reveal no relevant information otherwise. That is, if Alice performs $\mathcal{E}'$ instead of $\mathcal{E}$, then there is a 20% chance she will learn either that W or that $\neg W$ for sure and an 80% chance her credence in W will remain at .5. If Alice can only perform one of the two experiments, she will prefer to perform $\mathcal{E}$ if she uses the Brier Score, but she will prefer to perform $\mathcal{E}'$ if she uses the Log Score.[23]

Put slightly differently, the evidence an agent chooses to gather will depend on her scoring rule, which in turn depends on what sorts of decisions she expects she will make on the basis of her credence in the proposition in question. The decision of which evidence to collect is, what we might call, practico-epistemic. The evidence itself will determine her credal state, but her process of investigation is a practical one.

Finally, we observe that some experiments are absolutely preferable to others. That is, sometimes Alice will prefer to perform $\mathcal{E}_1$ to $\mathcal{E}_2$ regardless of her scoring rule. For instance, suppose Alice is interested in whether Bob or Carol will win the upcoming election. One polling company asks 5,000 people whom they plan to vote for. A second company asks only 50 (distinct) people. If we stipulate that each company uses a reasonable method of selecting participants, Alice will always prefer to learn the results of the first poll to the results of the second, although she would prefer to learn the results of both polls to the results of either one.

When performing $\mathcal{E}_1$ is preferable to performing $\mathcal{E}_2$ according to every strictly proper rule, $\mathcal{E}_1$ *first-order stochastically dominates* $\mathcal{E}_2$. That is, re-

22. For simplicity, we restrict attention here just to the gain in value with respect to the proposition W instead of the whole credence function.

23. From the setup, if Alice performs $\mathcal{E}'$, she has credence .1 she will learn W and credence .1 she will learn $\neg W$. In either case, she is guaranteed a perfect score of 0, since she will end up with credence 1 (0) in W just if W is true (false). Conditional on learning nothing relevant, she has credence .5 she will receive a score of $g_1(.5)$ and credence .5 she will receive a score of $g_0(.5)$. She assigns credence .8 to learning nothing relevant, so her total expected score should she perform experiment $\mathcal{E}'$ is $\mathrm{EG}(b^{\mathcal{E}'}(W)) = .4(g_1(.5) + g_0(.5))$. According to the Brier Score, she expects to have disutility .1875 after performing $\mathcal{E}$ and disutility .2 after performing $\mathcal{E}'$. On the Log Score, she expects disutility of approximately .56 after performing $\mathcal{E}$ and disutility of approximately .55 after performing $\mathcal{E}'$.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

gardless of the kinds of bets Alice expects to make, she is in expectation better off if she performs $\mathcal{E}_1$.[24]

4.2. Evaluation of Agents. One special kind of experiment we often conduct is to ask other agents what their credence is in some proposition. Instead of going to her favorite team's website, for instance, Alice might ask Bob how confident he is that the team won.

Ex ante, Alice does not know what Bob thinks. We can determine Alice's assessment of the value of asking Bob for his credence the same way we determined the value of other experiments above. Let $\mathcal{B}$ be the experiment of asking Bob. Then $\text{Val}(\mathcal{B}) = \text{EG}(.5) - \text{EG}(b^{\mathcal{B}}(W))$. That is, the value of asking Bob is the expected difference in inaccuracy (according to Alice's scoring rule) of her own credence after asking Bob and her current credence.

Note that $\text{EG}(b^{\mathcal{B}}(W))$ is not Alice's assessment of Bob's inaccuracy but is instead her assessment of what her own inaccuracy will be after talking to Bob. For instance, suppose Alice knows that Bob always has credence 0 in truths and credence 1 in falsehoods. If Bob tells her he is certain that her team lost, then she will become certain that they won and vice versa. So, she is ex ante certain that Bob will be perfectly inaccurate and also certain that she herself will be perfectly accurate after talking to him.

More formally, we can capture this distinction between Alice's assessment of her own expected posterior inaccuracy and her assessment of Bob's expected inaccuracy as follows. Let $B = x$ refer to the proposition that Bob's credence in W is x, and let $b_x(W) := b(W|B = x)$. We then have[25]

$$\begin{aligned}\text{EG}\left(b^{\mathcal{B}}(W)\right) = \sum_x b(B = x)\big[\, & b_x(W)g_1(b_x(W)) \\ & + (1 - b_x(W))g_0(b_x(W))\,\big],\end{aligned} \tag{11}$$

$$\text{EG(B)} = \sum_x b(B = x)(b_x(W)g_1(x) + (1 - b_x(W))g_0(x)). \tag{12}$$

As before, then, the value of the experiment of asking Bob (i.e., eq. [11]) is determined by what Alice expects of the disutility of her own future credence, which of course depends on her expectations about what she will use her credence for. That is why we here look at $G(b_x(W))$—that is, the inaccuracy of Alice's credences after talking to Bob.

Equation (12) measures Alice's assessment of Bob's inaccuracy before she learns what he actually thinks. Practically, her view of Bob's inaccuracy

24. See DeGroot and Fienberg (1982, 1983) and Schervish (1989) for more on the value of experiments and first-order dominance.

25. More generally, if Bob's credence could take on any value in [0,1], we can straightforwardly replace the sums in both eqq. (11) and (12) with integrals.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

measures how much Alice expects to lose if she were to switch over from her own credences to Bob's to make decisions while retaining her preferences over outcomes. So, if she expects that Bob is more accurate than she is, she would prefer (ex ante) to use his credences to hers.[26] Note that, if Bob is an epistemic expert for Alice, then equations (11) and (12) coincide. That is, if for any x, $b(W|B = x) = x$, then $g_i(b_x(W)) = g_i(x)$.

As with experiments, which agents are expected to be more accurate than which others is scoring-rule dependent. For instance, Alice might expect Bob to be more accurate with respect to the Brier Score but less accurate with respect to the Log Score than Carol is.[27] If Alice uses the Brier Score, then she would prefer Bob's credences to her own, given the decisions she actually expects to make.

Alice may also sometimes expect one agent to be more accurate than another regardless of which scoring rule she uses. For example, suppose Alice treats Bob and Carol both as epistemic experts, but she knows that Carol will have either credence .8 or .2 in W, while Bob will have either credence .6 or .4 in W. It is easy to check, by equation (12), that for any G that is strictly proper, Carol is in expectation more accurate than Bob. So, regardless of the bets Alice expects she will make, she thinks she would be better off using Carol's credences than Bob's credences.[28] In other words, one agent is judged absolutely more accurate than another if she is thought to be doing better regardless of the purposes of inquiry. One agent is judged better relative to a particular rule if she is thought to be doing better relative to the particular purposes of inquiry for the proposition(s) in question.

5. Conclusion. We began with two questions: (1) Which measures of inaccuracy are legitimate? and (2) When should we use a particular measure over another? Schervish's theorem provided an answer to both. Strictly proper scoring rules are the right tools for measuring an agent's inaccuracy when she is broadly uncertain what sorts of bets she might face. The exact nature of her expectation of future decision problems will determine which proper scoring rule in particular is right for her. Although Schervish's theorem identifies inaccuracy with expected practical loss, it retains some claim to being a

26. Because Alice does not yet know what Bob's credences are, G can be strictly proper even though she still expects Bob to be more accurate than she is.

27. To see why, imagine that Alice knows Bob has either credence .25 or .75 in W and that Carol has credence 1, 0, or .5. If she treats both agents as experts with respect to W, then—with the right numbers filled in—asking Bob for his credence can be made equivalent to experiment $\mathcal{E}$, and asking Carol can be made equivalent to $\mathcal{E}'$ above.

28. See DeGroot and Fienberg (1982, 1983) and DeGroot and Eriksson (1985) for a detailed characterization of when one credence function is more accurate than another on every strictly proper rule.

https://doi.org/10.1086/693444 Published online by Cambridge University Press

measure of epistemic utility. After all, it measures the value of a doxastic state and the value of truth.

REFERENCES

Bronfman, A. 2009. "A Gap in Joyce's Argument for Probabilism." Unpublished manuscript, University of Michigan.

D'Agostino, M., and C. Sinigaglia. 2010. "Epistemic Accuracy and Subjective Probability." In *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, ed. M. Suárez, M. Dorato, and M. Rédei, 95–105. Dordrecht: Springer.

DeGroot, M. H., and E. Eriksson. 1985. "Probability Forecasting, Stochastic Dominance, and the Lorenz Curve." In *Bayesian Statistics 2: Proceedings of the Second Valencia International Meeting*, ed. J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, 99–118. Amsterdam: Elsevier.

DeGroot, M. H., and S. E. Fienberg. 1982. "Assessing Probability Assessors: Calibration and Refinement." In *Statistical Decision Theory and Related Topics III*, vol. 1, ed. S. S. Gupta and J. O. Berger. New York: Academic Press.

———. 1983. "The Comparison and Evaluation of Forecasters." In "Proceedings of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics," special issue, *Statistician* 32 (1–2): 12–22.

Gibbard, A. 2007. "Rational Credence and the Value of Truth." In *Oxford Studies in Epistemology*, vol. 2, ed. T. Gendler and J. Hawthorne, 143–64. Oxford: Oxford University Press.

Good, I. 1967. "On the Principle of Total Evidence." *British Journal for the Philosophy of Science* 17:319–22.

Greaves, H., and D. Wallace. 2006. "Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility." *Mind* 115 (632): 607–32.

Jose, V. R. 2007. "A Characterization for the Spherical Scoring Rule." *Theory and Decision* 66:263–81.

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

———. 1999. *The Foundations of Causal Decision Theory*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge: Cambridge University Press.

———. 2009. "Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief." In *Degrees of Belief*, vol. 342, ed. F. Huber and C. Schmidt-Petri, 263–97. Dordrecht: Springer.

Konek, J., and B. A. Levinstein. 2017. "The Foundations of Epistemic Decision Theory." *Mind*, forthcoming.

Leitgeb, H., and R. Pettigrew. 2010a. "An Objective Justification of Bayesianism I: Measuring Inaccuracy." *Philosophy of Science* 77:201–35.

———. 2010b. "An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy." *Philosophy of Science* 77:236–72.

Myrvold, W. C. 2012. "Epistemic Value and the Value of Learning." *Synthese* 187 (2): 547–68.

Pettigrew, R. 2013. "A New Epistemic Utility Argument for the Principal Principle." *Episteme* 10 (1): 19–35.

———. 2016a. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

———. 2016b. "Accuracy, Risk, and the Principle of Indifference." *Philosophy and Phenomenological Research* 92 (1): 35–59.

———. 2016c. "Jamesian Epistemology Formalised: An Explication of 'The Will to Believe.'" *Episteme* 13 (3): 253–68.

Popper, K. 1959. *The Logic of Scientific Discovery*. New York: Basic.

Rényi, A. 1955. "On a New Axiomatic Theory of Probability." *Acta Mathematica Academiae Hungarica* 6 (3): 286–335.

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

Schervish, M. 1989. "A General Method for Comparing Probability Assessors." *Annals of Statistics* 17:1856–79.

https://doi.org/10.1086/693444 Published online by Cambridge University Press