Accuracy, Verisimilitude, and Scoring Rules

forthcoming in Australasian Journal of Philosophy

Jeffrey Dunn (DePauw University)

jeffreydunn@depauw.edu

Abstract

Suppose that beliefs come in degrees. How should we then measure the accuracy of these degrees of belief? Scoring rules are usually thought to be the mathematical tool appropriate for this job. But there are many scoring rules, which lead to different ordinal accuracy rankings. Recently, Fallis and Lewis (2016) have given an argument that, if sound, rules out many of the many popular scoring rules, including the Brier score, as genuine measures of accuracy. I respond to this argument, in part by noting that the argument fails to account for verisimilitudethat certain false hypotheses might be closer to the truth than other false hypotheses. Oddie (forthcoming), however, has argued that no member of a very wide class of scoring rules (the so-called proper scores) can appropriately handle verisimilitude. I explain how to respond to Oddies argument and recommend a class of weighted scoring rules that, I argue, genuinely measure accuracy while escaping the arguments of Fallis and Lewis as well as Oddie.

1 Introduction

Suppose that one is attracted to an epistemic picture according to which epistemic value comes first, and from this we derive our epistemic norms. An important part of this project is to say something about epistemic value. There are many ways to go here, but a particularly popular approach thinks that epistemic value is had by beliefs, and that the value of a belief increases as the belief increases in accuracy.[1] If we think of beliefs as coming in degrees, then we can represent belief states with credence functions: assignments of real numbers to propositions, where the greater the number the stronger the belief in that proposition. The question for such a theorist is to say how to rank credence functions in terms of their accuracy. Such a ranking is of course only part of the whole story—one would still need to explain how these facts about accuracy generate epistemic norms.[2] But getting the accuracy facts right is clearly necessary. This paper is about the right way to measure accuracy for credence functions. While I've motivated the project by appealing to a certain conception of epistemology, even those who don't take this approach have reason to be interested. For even if accuracy isn't *the only* thing that matters, it is certainly an important component of what matters.

Mathematicians and statisticians have devised mathematical tools—often called *scoring rules*—to rank the accuracy of probabilistic forecasts. This is a good place to start if one wants to rank credence functions for accuracy. For any of these rules to assist us in measuring *accuracy*, some minimal constraints must be placed on them. For instance, any measure of accuracy

[1] Goldman (1999) dubs this view 'veritism'.

[2] James Joyce (1998) gives the first indication how such a story might go. For other representative work in this vein see Greaves & Wallace (2006), Gibbard (2008), Joyce (2009), Leitgeb & Pettigrew (2010a,b), and Pettigrew (2016).

must be truth-directed in the sense that if two credence functions differ only in that the first gives greater credence to the true propositions, then the first is more accurate than the second. Another widely accepted constraint is that scoring rules must be *proper*.[3] Roughly, a proper score is one where any probability function expects itself to be at least as accurate as any other probability function according to that score.[4] There are several other plausible constraints on scoring rules.[5] But these widely-agreed upon constraints leave us with many possible ways to measure accuracy. So, it is an open question what is the right way to do it.

In this paper, I won't give a maximally specific answer to this question, but I will give an answer that rules out many approaches. I'll do this by first responding to arguments in Fallis & Lewis (2016). They argue against the very popular Brier score, and in favor of what I'll call the Partition-based Logarithmic Score. I argue that their arguments are flawed, which can be seen in an especially clear way when we focus on the phenomenon of verisimilitude: the fact that certain hypotheses while false, may be closer to the truth than other hypotheses. This motivates a search for scoring rules that can properly handle this phenomenon. This search, however, is challenged by a recent argument from Graham Oddie (forthcoming) that purports to show that no proper scoring rule can handle the phenomenon of verisimilitude. I respond to Oddie's argument and defend a particular class of scoring rules as those that genuinely measure accuracy. One feature of this class is worth noting: scores in this class maintain that the way that credence is distributed to false hypotheses can affect the accuracy of a

[3] Though see Blackwell Drucker [forthcoming] for a dissenting view.

[4] More formally, and using terminology to be introduced below, a local score, $\mathfrak{s}$, is proper just in case for $0 \leq p \leq 1$, $p\mathfrak{s}(1, x) + (1 - p)\mathfrak{s}(0, x)$ is minimized at $x = p$ (a score is said to be *strictly proper* if it is uniquely minimized at $x = p$).

[5] See, for instance, Joyce (2009) or Pettigrew (2016), chs. 3-4.

credence function. This is especially plausible in cases where more credence is given to false hypotheses that are closest to the truth, but I argue that it is defensible independent of considerations of verisimilitude (and, further, that this feature sometimes is in conflict with considerations of verisimilitude). The paper closes by considering how my proposal compares with a recent argument due to Richard Pettigrew (2016) that the Brier score is the unique measure of accuracy, as well as how extant accuracy-based arguments fare on my proposal.

But before getting to all this, some preliminary distinctions will (hopefully) make things easier to follow.

2 Global and Local, Partitions and Algebras

Distinguish first between a local scoring rule and a global scoring rule. A *local scoring rule* takes as inputs a prediction about a proposition—in our case, a credence in a proposition—and the truth value of that proposition, and gives you an accuracy score. So, for instance, we'd use a local scoring rule if we wanted to score for accuracy a credence of 0.75 that it will rain on a day when it does rain. More formally, a local scoring rule is a function $\mathfrak{s} : \{0, 1\} \times [0, 1] \to [0, \infty]$. One example is the **Local Brier Score** (Brier, 1950):

$$\mathfrak{b}(1, c(X)) = (1 - c(X))^2$$

$$\mathfrak{b}(0, c(X)) = (0 - c(X))^2$$

This takes the score of a credence, $c(X)$, in proposition X to be the square of the difference between the credence and the truth (with 1 representing truth and 0 representing falsity).[6]

[6] Note that a score of 0 obtains when one is fully confident in X and X is true or when one has no confidence in X and it is false, so the Local Brier Score is best thought of as a

A *global scoring rule*, in contrast, takes as input a set of predictions—in our case, an entire credence function—and the true state of the world, and delivers the accuracy of that entire set of predictions. There are different ways to build global scoring rules. One way to do this is to think of credences as assigned to a *partition* (rather than an set of propositions) of the possible outcomes. We then score the way that probability is distributed to the partition relative to which cell of the partition is actual. One can take this route and define the **Partition-based (Global) Brier Score**. Let $H_1, H_2, \ldots, H_n$ be a set of n mutually exclusive and exhaustive hypotheses, H_t be the true hypothesis, and $v_{H_t}(X)$ be the function that takes value 1 when $X = H_t$ and 0 otherwise. We have:

$$\mathfrak{B}(H_t, c) = \sum_{i=1}^{n}(v_{H_t}(H_i) - c(H_i))^2$$

This is just the sum of the Local Brier Score applied to the elements of the partition. But not all partition-based global scoring rules are like this. For instance, the **Partition-based (Global) Logarithmic Score**[7] (which we'll see later in this paper) is:

$$\mathfrak{L}(H_t, c) = -ln(c(H_t))$$

Notice here we don't sum a local score over all the elements of a partition. Instead, the score is determined by the natural logarithm of the *true* hypothesis.

A related, but importantly different way to go, is to build a global scoring rule by summing the local score for some set of *propositions* over which a credence function is defined.[8] We can do this to get the **Proposition-based**

measure of *inaccuracy.* In fact this is true of all the scores we will encounter in this paper: they are all measures of *inaccuracy.*

[7] Throughout this paper I use the natural logarithm in all statements of various logarithmic scores, but choosing any base for the logarithm will result in a score that yields the same ordering of credence functions in terms of accuracy.

[8] Pettigrew (2016) calls such scores 'additive inaccuracy measures'.

(Global) Brier Score. Let $\mathcal{F}$ be the set of propositions to which credence function c assigns credence, and let w be a possible world defined as finely grained as the propositions in $\mathcal{F}$ allow. Then we have:

$$\mathfrak{B}^*(w, c) = \textstyle\sum_{X \in \mathcal{F}} (v_w(X) - c(X))^2$$

Alternatively, there is the **Proposition-based (Global) Logarithmic Score**:

$$\mathfrak{L}^*(w, c) = \textstyle\sum_{X \in \mathcal{F}} \mathfrak{l}(v_w(X), c(X))$$

where

$$\mathfrak{l}(1, c(x)) = -ln(c(X))$$

$$\mathfrak{l}(0, c(x)) = -ln(1 - c(X))$$

In sum, then, we have local scoring rules, which score particular credences and we have global scoring rules, which can be either partition-based or proposition-based, that score entire credence functions. Since we'll be concerned primarily with global scores, I'll drop the word 'Global' for readability; if I'm discussing a local score, I'll note that explicitly.

3 Fallis & Lewis Against the Brier Score

Fallis & Lewis (2016) argue against the Partition-based Brier Score. Their argument turns on the fact that this score does not satisfy a plausible-looking constraint on scoring rules. The constraint is:

(M3) All other things being equal, if credence function c_2 assigns a lower probability to some false hypothesis than credence function c_1 does, then c_2 is epistemically better than c_1.

What (M3) says depends on what "all other things being equal" means. Fallis & Lewis are focusing solely on probabilistically coherent credence functions defined over the same partition, so there's no way for c_2 to assign a lower probability than c_1 to some false hypothesis and yet c_2 and c_1 be in every other way identical. What they mean by "all other things being equal" is that where H_j is a false hypothesis (and so $c_1(H_j) > c_2(H_j)$), for all $k \neq j$ there is some real number, α, such that $c_1(H_k) = \alpha c_2(H_k)$. Intuitively, we can think of (M3) as saying the following. Suppose that c_2 is generated from c_1 by taking credence from a false hypothesis that c_1 gives positive credence to and redistributing it to the rest of the hypotheses preserving the ratios between these hypotheses that c_1 encodes. In that case, (M3) says that c_2 is epistemically better than c_1.

Of note is the following. Suppose that, as above, H_j is the false hypothesis, and that $c_1(H_j) > c_2(H_j) = 0$. And suppose, too, that all else is equal between c_1 and c_2 in the sense of (M3). Then it follows that c_2 is the credence function that would result from conditionalizing[9] c_1 on the evidence $\overline{H_j}$.[10,11] So, (M3) entails that when one conditionalizes on the evidence $\overline{H_j}$ when H_j is false, the resulting credence function is epistemically better than the initial credence function, where in this context 'epistemically better than' is to be understood in terms of greater accuracy.

Fallis & Lewis give an example of how the Partition-based Brier Score

[9] If c_1 is my credence function now and I receive evidence E (and nothing else), I conditionalize iff my next credence function, c_2, satisfies: $c_2(X) = c_1(X|E)$, for all propositions X over which c_2 and c_1 are defined.

[10] Here and throughout I use $\overline{X}$ to refer to the truth-functional negation of X.

[11] Proof: Since the H_i are mutually exclusive and exhaustive, $\overline{H_j}$ is equivalent to $\bigvee_{i \neq j} H_i$. Thus, c_2 is the credence function that would result from c_1 upon learning $\overline{H_j}$ just in case $c_2(H_i) = c_1(H_i | \bigvee_{i \neq j} H_i)$. The right-hand side of this is equal to $c_1(H_i \wedge (\bigvee_{i \neq j} H_i))/c_1(\bigvee_{i \neq j} H_i)$, which is equal to $c_1(H_i)/c_1(\bigvee_{i \neq j} H_i)$ for all $i \neq j$. Thus, for all $i \neq j$, $c_1(H_i) = \alpha c_2(H_i)$ with $\alpha = 1/c_1(\bigvee_{i \neq j} H_i)$. Further, if $c_2(H_j) = 0$, then this is the only such α that also ensures that $c_2(\bigvee_{i \neq j} H_i) = 1$ as it must if c_2 is to be a probability function.

violates (M3), which plays a role in why they think (M3) is a good constraint. Here's the example:

Case 1: Suppose there are three mutually exclusive and exhaustive hypotheses: H_1, H_2, and H_3. Suppose H_1 is true and consider the following two credence functions:

	$\boxed{H_1}$	H_2	H_3
c_1	1/4	1/2	1/4
c_2	1/3	2/3	0

Fallis & Lewis note that according to the Partition-based Brier Score, c_1 is more accurate than c_2, which is a violation of (M3). They further argue that we can see that (M3) is correct here. For notice that c_2 is exactly the credence function you would arrive at if you initially had credence function c_1, learned that H_3 was false, and conditionalized. This transition, they say, corresponds to a scientifically-respectable kind of inference: an "elimination experiment". An elimination experiment occurs when one gets evidence that definitively rules out a false hypotheses and provides no information about the other hypotheses. According to Fallis & Lewis, when one updates one's credence function in response to an elimination experiment, one has made an epistemic improvement. But the partition-based Brier score denies this, since it says that the accuracy of c_2 is less than that of c_1. They write:

> . . . defenders of the Brier rule would need to explain why simply eliminating a false hypothesis would decrease actual epistemic utility. In other words, they need to have a story about what is epistemically bad about 'elimination experiments'. (p. 9)

This, then, is their argument: the partition-based Brier Score violates (M3), and we can see this is mistaken by considering elimination experiments.

4 Response to Fallis & Lewis

One might claim that Fallis & Lewis's argument is flawed because it asks us to judge the transition from c_1 to c_2 by looking at the *actual* accuracy scores of these credence functions. But, goes the objection, this is unfair, since the agent making the transition doesn't have the information about which hypothesis is true. Thus, judgments about the appropriateness of a transition from one credence function to another cannot dictate whether the one credence function is *in fact* more accurate than the other.

It is important to see that this doesn't undermine Fallis & Lewis's argument. For if the Partition-based Brier Score really is the measure of accuracy, then there is a question we can ask from a third-personal view: given that H_1 is in fact true, and given that an agent changes from c_1 to c_2, has the agent in fact made an accuracy improvement? We can ask this question while acknowledging that it is a different question from the question of whether the agent herself is reasonable in making the transition. Our measure of accuracy needs to give the right answer to such questions.

That said, I think there is still a flaw in Fallis & Lewis's argument. As noted, (M3) entails that when one conditionalizes on the evidence $\overline{H_j}$ when H_j is false, the resulting credence function is more accurate than the initial credence function. But this is not always the case. To see the basic worry, consider the following example that doesn't involve credences but just all-or-nothing beliefs. Suppose that it actually is Monday and I believe truly: "It is Monday or Tuesday". In addition, I hold many false beliefs of the form "If it is Monday then X" where X is false. I subsequently learn the true proposition "It is not Tuesday". This leads me to revise my beliefs so that I believe it is Monday and believe the various Xs. I claim that in such a situation I haven't made an epistemic improvement, even though

I've definitively ruled out a false hypothesis and then updated. Why have I not made an improvement? Because the proposition "It is not Tuesday" functions as misleading evidence for me: it is a true proposition that when learned, leads me to infer and believe many false propositions.

We can now adapt a case of misleading evidence to a credal context to see that (M3) is not in general true:

Case 2: Suppose there are three propositions an agent cares about—A, B, and C—and thus eight mutually exclusive and exhaustive hypotheses. Suppose that, in fact, all of A, B, and C are true. Thus, what I have called H_1 in the diagram below is in fact the true hypothesis. Consider an agent with the following initial assignment of credences to hypotheses:

$\boxed{H_1}$	H_2	H_3	H_4	H_5	H_6	H_7	H_8
ABC	$A\overline{B}C$	$AB\overline{C}$	$A\overline{BC}$	$\overline{A}BC$	$\overline{AB}C$	$\overline{A}B\overline{C}$	$\overline{ABC}$
$0.00\overline{3}$	$0.00\overline{3}$	$0.00\overline{3}$	0.19	0.8	0	0	0

If we let c_1 be the credence function encoding this distribution, here are the credences in the true propositions that the agent has:

$$c_1(A) = 0.2 \qquad c_1(B) = 0.80\overline{6} \qquad c_1(C) = 0.80\overline{6}$$

Now, suppose that the agent learns that A is true. If the agent conditionalizes, then he removes all credence from H_5 and assigns it to H_1-H_4 in such a way that the ratios between the credences assigned to H_1-H_4 remain the same. This results in the following distribution of credence to hypotheses:

$\boxed{H_1}$	H_2	H_3	H_4	H_5	H_6	H_7	H_8
ABC	$A\overline{B}C$	$AB\overline{C}$	$A\overline{BC}$	$\overline{A}BC$	$\overline{AB}C$	$\overline{A}B\overline{C}$	$\overline{ABC}$
$0.01\overline{6}$	$0.01\overline{6}$	$0.01\overline{6}$	0.95	0	0	0	0

If we let c_2 be the agent's new credence function encoding this distribution, here are the credences in the true propositions:

$$c_2(A) = 1 \qquad c_2(B) = 0.0\overline{3} \qquad c_2(C) = 0.0\overline{3}$$

Notice that c_2 assigns the false hypothesis, H_5, a lower credence than does c_1, but other than this all else is equal, in the sense of (M3). Thus, Fallis & Lewis have to maintain, in line with (M3), that c_2 is epistemically better than c_1.

But this is mistaken. The agent was initially very inaccurate about A but very accurate about B and C. Upon changing credences, the agent is perfectly accurate about A but almost perfectly inaccurate about both B and C. That is a bad trade. A in this scenario is misleading evidence: evidence that, while true, leads one to greater inaccuracy about other matters. Looked at in this way, the epistemic value of the initial credence function, c_1, should be greater than the epistemic value of the updated credence function, c_2. So, (M3) is false and in this situation the Partition-based Brier Score gives the correct verdict.

Fallis & Lewis do consider the objection that when one gets true but misleading evidence one doesn't thereby improve one's credence function:

> Admittedly, the results of experiments are sometimes misleading, such that actual epistemic utility goes down even though

> expected epistemic utility goes up. But there is nothing at all misleading about a result that definitively eliminates a false hypothesis (and that has no other effect on one's cognitive state). (p. 584)

But what we've seen here is that the evidence itself can be accurate (H_5 really is false), but it can still point in a misleading direction. By focusing on generic mutually exclusive and exhaustive hypotheses, one can miss ways that certain hypotheses, while false, are closer to the truth than others. That is, one can miss the phenomenon of verisimilitude. In the case just given, although H_5 is false, it gets things right with respect to propositions B and C. Given the remaining distribution of credences, it looks like the epistemic state that retains credence in the "mostly true" H_5 is better than the epistemic state that shifts the bulk of this credence to the "mostly false" H_4.

Despite this, one might think (M3) still seems true. Note, however, that there are principles similar to (M3) that are true. For instance, if c_2 is the credence function you get from removing credence from a false hypothesis that c_1 gives credence to and distributing that credence to the *true* hypothesis, then c_2 is more accurate than c_1. *That* principle is true (it also, however, doesn't distinguish between the Brier Score and any other proper score). So, we can reject (M3) while recognizing why it seems plausible.

Case 2 does more than neutralize the (M3)-based argument against the Partition-based Brier Score, however. It also demonstrates a problem with the scoring rule that Fallis & Lewis champion: the Partition-based Logarithmic Score, $\mathfrak{L}$. As Fallis & Lewis note (and which is easily proved), if c_2 gives the true hypothesis a greater credence than does c_1, then the Partition-based Logarithmic Score assigns c_2 greater epistemic value than it does c_1.

That is, if the Partition-based Logarithmic Score is the correct measure of accuracy, then (M3) is true. So, Case 2 shows that (M3) is false and also rules out the Partition-based Logarithmic Score.

5 Problems for the Partition-based Brier Score

Despite this, it is not smooth sailing for the Partition-based Brier Score. Even though it seems to get the correct verdict in Case 2, it does not get the right result in a related case:

Case 3: Keep the formal details of Case 2 the same, including that the H_i are generated from propositions A, B, and C, but suppose that H_2 is now the true state of the world. The Partition-based Brier Score gives both c_1 and c_2 the same scores as it did in Case 2, so we must say that c_1 is more accurate than c_2. But in this case, that verdict is mistaken. This can be seen by looking at the credences in the true propositions (which are, in this case, $A, \overline{B}$, and C):

$$c_1(A) = 0.2 \qquad c_1(\overline{B}) = 0.19\overline{3} \qquad c_1(C) = 0.80\overline{6}$$

$$c_2(A) = 1 \qquad c_2(\overline{B}) = 0.9\overline{6} \qquad c_2(C) = 0.0\overline{3}$$

To the extent that one agrees that c_1 is more accurate than c_2 in Case 2, one seems compelled to say that c_2 is more accurate than c_1 in this case. But then the Partition-based Brier Score cannot be correct since it gives the opposite verdict.

6 Solution: Weighted Proposition-based Scores

Cases 2 and 3, I believe, show problems for (M3) as well as the two partition-based scores we've considered. But they also suggests a remedy: if partition-based scores are missing facts about hypotheses that are closer to the truth than others, then we need to make our scores sensitive to verisimilitude. The natural way to do so is to shift to a proposition-based score. For, the idea goes, a proposition-based score is going to be able to "see" that H_5 is closer to H_1 than H_4 in Case 2.

The natural way to pursue this is to turn to a proposition-based score that scores each proposition in an entire algebra of propositions. That is, we score the credence the agent has in each proposition in a set of propositions closed under negation and disjunction. However, as Graham Oddie (forthcoming) points out, a complete algebra of propositions doesn't allow us to track verisimilitude as we'd like. This is rather surprising, but can be seen with a simple example consisting of two atomic propositions, A and B. Intuitively, $A\overline{B}$ is closer than $\overline{AB}$ to AB. But if we look at the entire algebra of propositions generated by A and B, we get that they agree with AB about the same number of propositions. For $A\overline{B}$ and AB agree with each other about the truth of three propositions: A, $A \vee \overline{B}$, $A \vee B$. But $\overline{AB}$ and AB agree with each other about the truth of three propositions, too: $\overline{A} \vee B$, $A \vee \overline{B}$, $A \leftrightarrow B$.

If we want to be sensitive to verisimilitude, then, we will have to privilege certain propositions. Our intuitive sense that $A\overline{B}$ is closer than $\overline{AB}$ to AB is plausibly the result of us caring in particular about the atomic propositions A, B, and their negations. If we cared especially about getting biconditionals correct, for instance, $\overline{AB}$ may appear closer than $A\overline{B}$ to AB.

In light of this, I propose that the correct way to measure accuracy is with a proposition-based proper score, where the score for certain propositions are weighted more heavily than others. More carefully, let $\mathfrak{s}(v_w(X), c(X))$ be a local, strictly proper score. Let $\mathcal{F}$ be the set of propositions to which the credence function c assigns credence. Let w be a possible world defined as finely grained as the propositions in $\mathcal{F}$ allow. Finally, let $\lambda(X)$ be a weighting function that assigns propositions in $\mathcal{F}$ a weight.[12] Then, the appropriate score to use is:

$$\mathfrak{S}(w, c) = \sum_{X \in \mathcal{F}} \lambda(X)\mathfrak{s}(v_w(X), c(X))$$

Notice that this is a schema, not a single score. You get a definite score only after supplying a local score, $\mathfrak{s}$, and a weighting function, λ. I'll refer to this proposal as the *weighted score proposal* and I'll refer to, *e.g.,* the specific instance of such a score that uses the local Brier score as the *Weighted Brier Score.*

Suppose, then, that in Cases 2 and 3 we use a λ that gives most of the weight to A, B, C, and their negations. Then, both the Weighted Brier and Weighted Logarithmic Scores give the correct verdict: c_1 is better than c_2 in Case 2, and the opposite in Case 3.

How, on the weighted score proposal, do we determine which propositions are most heavily weighted? In many cases it is natural to focus on the atomic propositions and their negations. But it is certainly possible for there to be situations where either the agent or those who are scoring the agent have other propositions that are of interest. I propose that the weighting of propositions is fixed by these interests.[13] Much more could be said about how propositions are appropriately weighted, but if we hope to capture the

[12] More on how to assign these weights shortly.

[13] If one thinks that there are objective facts about which propositions are most important in inquiry—perhaps in light of explanatory value—one could build this into the weighting function, though I don't assume that here. See Pérez Carballo (forthcoming) for some thoughts on this.

phenomenon of verisimilitude, some such weightings are needed.

To get a feel for this proposal, consider an example. Suppose that A is the proposition that it will rain this evening and B is the proposition that it will be windy this evening, and that in fact it will be both rainy and windy. Initially, a forecaster assigns equal probabilities (0.3) to the false hypotheses. The remainder of her credence is given to the (true) hypothesis that it will be both rainy and windy. She subsequently revises her opinion to eliminate her credence in no rain and no wind. That is, we have the following case:

Case 4:

	$\boxed{AB}$	$A\overline{B}$	$\overline{A}B$	$\overline{AB}$
c_1	0.1	0.3	0.3	0.3
c_2	0.1429	0.4286	0.4286	0

Has the forecaster improved her accuracy? If what we care about are the propositions A, B, and their negations, then it looks like she has increased in accuracy. She's gone from a credence of 0.4 in each true proposition to a credence of approximately 0.57. And, when A, B, and their negations are weighted most heavily by λ, both the Weighted Brier and Weighted Logarithmic scores agree. And such a verdict makes sense. Credence is taken from the wholly false $\overline{AB}$ and distributed to the two other false hypotheses that are both partially accurate.

But suppose the forecaster really cares about not wrongly claiming that only one of rain or wind will occur, when in fact both will. Then, we may not see the transition from c_1 to c_2 as an improvement. For now the forecaster has become very confident that $A \leftrightarrow B$ is false. If we give heaviest weight to this biconditional and its negation, we get the opposite result: c_2 is no longer an accuracy improvement over c_1. Again, this makes sense. In such a scenario, $\overline{AB}$ is closer to AB than the other hypotheses.[14]

7 Proximity and Verisimilitude

Graham Oddie (forthcoming), however, attempts to show that *no* proper score—whether weighted or not—can account for verisimilitude. If he's right, the weighted score proposal is mistaken. Oddie's argument relies on a proof that no proper score can respect a constraint he calls *Proximity*. Proximity says that if a credence function distributes some credence to a set of false mutually exclusive hypotheses, then if you redistribute the credence in these false hypotheses so that it is all concentrated on a false hypothesis that is closest to the truth, you do not increase your inaccuracy.

Suppose, for the sake of argument, that both $\overline{A}B$ and $A\overline{B}$ are closer to AB than is $\overline{AB}$. Proximity would say that the following transition must be an improvement in accuracy.

Case 5:

	$\boxed{AB}$	$A\overline{B}$	$\overline{A}B$	$\overline{AB}$
c_1	0	$0.\overline{3}$	$0.\overline{3}$	$0.\overline{3}$
c_2	0	1	0	0

This is because we take the credence in false hypotheses, and put it all into $A\overline{B}$, which is one of the false hypotheses closest to the truth. However, the Weighted Brier and Weighted Logarithmic scores say that c_1 is more accurate than c_2 even when the atomic propositions are the only propositions

[14] One might wonder what happens when the forecaster cares about all propositions equally. This is addressed in Section 8.

with positive weight. So, they violate Proximity. Oddie takes this to show that no proper score can respect insights concerning verisimilitude and hence that no proper score can genuinely measure accuracy.[15]

In response, I claim that Proximity is false: not every way of moving credence from false hypotheses to a false hypothesis that is closest to the truth constitutes an accuracy improvement. For example, in Case 5 I claim that Proximity is false because c_2 is not more accurate than c_1. To be clear: this is to maintain that while verisimilitude is one aspect of accuracy, it is not the sole aspect.

What can be said in favor of this response? Note first that Case 5 is importantly different than Case 4. In Case 4, accuracy is improved with respect to both A and B when transitioning from c_1 to c_2. In Case 5, in contrast, accuracy is increased with respect to A, but decreased with respect to B.

More important than this, however, is the fact that c_2 expresses extreme confidence in one particular false hypothesis, while c_1 does not. Suppose, again, that A is the proposition that it will rain and B the proposition that it will be windy. In Case 5, our forecaster initially assigns equal credence (1/3) to it being rainy but not windy, to it being windy but not rainy, and to it being neither windy nor rainy. She subsequently revises her opinion to be completely confident that it will be rainy and not windy. She has not improved in accuracy because she has overcommitted to a specific, false, hypothesis. Better to spread one's credence out among the false hypotheses, or at least among $A\overline{B}$ and $\overline{A}B$.[16] This mitigates the fact that she has, in a

[15] Note, too, that this can be seen as simply an extreme version of an elimination experiment, the only difference being that in this case, the true hypothesis is given no credence. And (M3)—which I've already argued is false—says that c_2 is more accurate than c_1.

[16] And note, that if credence is removed from $\overline{AB}$ and distributed equally to $A\overline{B}$ and $\overline{A}B$, the Weighted Brier and Weighted Logarithmic scores agree that the transition is an accuracy improvement.

sense, moved closer to the truth.

One way to motivate this picture of accuracy is to draw an analogy with all-or-nothing belief. While we must be cautious in identifying high credence with all-or-nothing belief, c_2 is analogous the all-or-nothing belief state of someone who fully believes that it will rain and not be windy. On the other hand, c_1 is analogous to the more cautious all-or-nothing belief state of someone who withholds judgment between the three false hypotheses. Both make a mistake, of course, but it is plausible that c_2's mistake is worse.

And there is a further argument in favor of the claim that distributing too much credence to one false hypothesis is bad for overall accuracy.[17] This argument appeals to two ideals: it is better for credence functions to (1) assign more credence to truths and (2) less credence to falsehoods. Suppose we're only dealing with a set of mutually exclusive and exhaustive hypotheses. The credence function that assigns credence 1 to the true hypothesis is perfectly satisfying both these ideals. Now, consider two different credence functions. Both assign credence 0 to the true hypothesis, but the first assigns credence 1 to some particular false hypothesis, while the second distributes this credence evenly among all the false hypotheses. The first credence function is doing poorly with respect to ideal 1 (assigning lots of credence to the truth) *and* it is doing poorly with respect to ideal 2 (assigning less credence to falsehoods). For it assigns credence 1 to a falsehood. The second credence function is, like the first, doing poorly with respect to ideal 1. But it is doing better with respect to ideal 2 (assigning less credence to falsehoods). Each

[17] In a blog post on M-Phi Knab & Schoenfield (2015, March 12) defend a principle they call **Falsity Distributions Don't Matter:** "For any partition of theories: $t_1, \ldots, t_n$, a probabilistic agent's accuracy with respect to this partition at world w should be determined solely by the amount of credence she invests in the true theory at w, and the amount of credence she invests in false theories at w. The way she distributes her credences amongst the false theories at w shouldn't affect her accuracy." What I say here can be seen as a reason to doubt this principle.

falsehood is assigned very little credence. So, it is perfectly reasonable to see the second credence function as more accurate than the first.

Thus, I maintain that the weighted accuracy proposal can capture facts about verisimilitude when it should, but the fact that it violates Proximity is no strike against it as a proposal about how to properly measure accuracy.

8 When Only the Partition Matters

So far we've seen that (M3) fails in scenarios where some hypotheses are closer to the truth than others. But what about contexts where verisimilitude is *not* a factor? These will be cases where the only propositions that matter are the mutually exclusive and exhaustive hypotheses themselves. One might think that while (M3) has counterexamples when verisimilitude is a factor, violations of it are still objectionable in cases where verisimilitude is not a factor.

One way to make this objection vivid is to look back at Fallis & Lewis's initial case against the Brier Score (Case 1). When verisimilitude *is* relevant, the defender of the weighted score proposal has a clear thing to say: the case is under-described, since we don't have any information about the propositions that generate the three hypotheses. When we fill in such details, we get plausible results.[18]

[18] For example, here are two different ways we could fill in the details:

Case 1.1:

	$\boxed{AB}$	$A\overline{B}$	$\overline{A}B$	$\overline{AB}$
c_1	1/4	1/2	0	1/4
c_2	1/3	2/3	0	0

Case 1.2:

	$\boxed{AB}$	$A\overline{B}$	$\overline{A}B$	$\overline{AB}$
c_1	1/4	1/2	1/4	0
c_2	1/3	2/3	0	0

But the objection now is what to say about Case 1 when only the mutually exclusive and exhaustive hypotheses are given positive weight by λ. We saw earlier that the partition-based Brier score says that c_2 is less accurate than c_1 in Case 1 and so the Weighted Brier score is going to say the same thing when we have such a λ. And though the Weighted Logarithmic score doesn't give that result in Case 1, it does give a similar result in a related case.

Case 1*:

	$\boxed{H_1}$	H_2	H_3
c_1	0.2	0.6	0.2
c_2	0.25	0.75	0

The objection, then, is that though (M3) is false in general, it is nevertheless true when none of the (false) mutually exclusive and exhaustive hypotheses are closer to the truth than any others. But the weighted score proposal says that (M3) is false even in some such cases.[19]

The response to this objection builds off the response to Oddie's Proximity objection. I argued that a credence function can decrease in accuracy by becoming *extremely* confident in one particular false hypothesis. This, I claim, is what happens in the situations depicted in Cases 1 and 1*: too much credence is given to one particular false hypothesis. So even though c_2 assigns less overall credence to the *set* of false hypotheses, it is worse

If λ assigns the atomic propositions (and their negations) heavy weights, then, in Case 1.1, we get that c_2 is an improvement over c_1 according to both the Weighted Brier and Weighted Logarithmic scores. And in Case 1.2, we get that c_2 is not an improvement over c_1. These are plausible verdicts when the atomic propositions are what matter.

[19] In an unpublished paper, (Lewis & Fallis, 2016, pp. 13-4), prove that any proposition-based proper score that assigns weights only to the mutually exclusive and exhaustive hypotheses, will face these sorts of (M3) violations.

than c_1 because it invests a large amount of credence in one particular false hypothesis.[20]

Importantly, such a view sits comfortably with the idea that accuracy is the sole epistemic value. Accuracy, on this picture, depends not only on how much total credence is given to the true and the false, but also on how credence is distributed amongst the false hypotheses. But it is still accuracy alone that is of epistemic value. Given this picture of accuracy, we should expect there to be (M3) violations, even in cases where verisimilitude is not a factor.

I have just claimed that violations of (M3) are not objectionable when too much credence is given to one particular false hypothesis. Different proper scores will differ over how much is "too much". Cases 1 and Case 1^* show us this. The Weighted Logarithmic score says that c_2 is more accurate than c_1 in Case 1, but not in Case 1^*. The Weighted Brier score says that c_2 is less accurate than c_1 in both cases. So, different scores disagree about how extreme the credence must be in the particular false hypothesis before we get acceptable violations of (M3). But that disagreement is one of degree and not of kind. Once we think there are acceptable (M3) violations—even without verisimilitude—then where exactly to draw the line is a subtle matter. That is why I advocate for some weighted proper score, but offer no argument for one particular proper score such as the Weighted Brier score or the Weighted Logarithmic score.

[20] For instance, in Case 1^* c_1 assigns to false hypotheses a total of 0.8 credence and c_2 assigns to false hypotheses a total of 0.75 credence. But since c_2 gives all of this to one particular hypothesis (H_2), it is worse than c_1.

9 Against Pettigrew's argument for the Brier score

In recent work, Richard Pettigrew (2016) has gone further and argued in favor of the Proposition-based Brier Score over any alternative score.[21] In this section I explain why I don't think this argument works. We have a reason to go for a weighted score, but not necessarily the Brier score. To see Pettigrew's argument in favor of the Proposition-based Brier Score we need to first see the relationship between scoring rules and *additive Bregman divergences*. In general, a divergence, $\mathfrak{D}$, is a function on two vectors in $[0,1]^n$, $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ such that $\mathfrak{D}(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in [0,1]^n$. More intuitively, in the context of credence functions, a divergence is a measure of the difference between two credence functions.

A certain subclass of divergences are the additive Bregman divergences.[22] As Pettigrew shows, drawing on work by Predd *et al.* (2009), for every additive scoring rule, $\mathfrak{S}$, that meets several conditions[23] there is an associated additive Bregman divergence, $\mathfrak{D}$ such that $\mathfrak{S}(w, c) = \mathfrak{D}(v_w, c)$ where v_w is the omniscient credence function at w.[24] Additive Bregman divergences are hence more general than scoring rules, since the scoring rule gives you a measure of the difference between a credence function and the truth-values at a world whereas a divergence gives you a measure of the difference between two arbitrary credence functions.

Pettigrew's argument in favor of the Brier score, depends on an important feature of the divergence associated with the Brier score: it is the sole

[21] Pettigrew (2016) also does not consider weighted scores of the sort I advocate for, but as I note in the section below, most of his arguments are compatible with such weightings.

[22] For details, see Pettigrew (2016, pp. 84-5).

[23] These conditions are: Alethic Vindication, Perfectionism, Divergence Additivity, Divergence Continuity, and Decomposition. For the definitions, see chapter 4 of Pettigrew (2016).

[24] The omniscient credence function at w is just the credence function that assigns all truths at w credence 1 and all falsehoods at w credence 0.

symmetric divergence. That is, for the Brier score and only the Brier score, it is true that the divergence from c_2 to c_1 is the same as the divergence from c_1 to c_2. Here is what Pettigrew says:

> We have a strong intuition that the inaccuracy of an agent's credence function at a world is the distance between that credence function and the ideal credence function at that world. But we have no strong intuition that this distance must be the distance from the ideal credence function to the agent's credence function rather than the distance to the ideal credence function from the agent's credence function; nor have we a strong intuition that it is the latter rather than the former. But if there were non-symmetric divergences that gave rise to measures of inaccuracy, we would expect that we would have intuitions about this latter question, since, for at least some accounts of the ideal credence function at a world and for some agents, this would make a difference to the inaccuracies to which such a divergence gives rise. (Pettigrew, 2016, p. 67)

Pettigrew maintains that we don't have the intuition that it is distance *to* the omniscient credence function or distance *from* the omniscient credence function that matters for accuracy. But if the true accuracy score corresponded to a non-symmetric divergence, we would have such intuitions. Hence, the true score does not correspond to a non-symmetric divergence.

If sound, this argument rules out every score except the Proposition-based Brier Score. But I believe it is unsound, because there is no reason to think the conditional premise is true. If we had strong intuitions that divergences must be symmetric, this might tell against scores like the Logarithmic Score, but the mere lack of intuitions about this is not probative.

What's more, I think a positive story can be given for why a non-symmetric divergence may be appropriate. This story doesn't give a reason to favor a non-symmetric divergence, but I think it does undermine any advantage a symmetric divergence has. The key is to note that when we score the accuracy of a credence function, we are really comparing two different kinds of things: a doxastic state (a credence function) and the truth (a set of truth-values). This difference is obscured if we think of scoring rules as comparing a credence function with an omniscient credence function (as Pettigrew does). Of course, formally, things come out the same no matter which way one goes since $v_w(x)$ takes the same value as the omniscient credence function at w. But if the things being compared are different in kind, this makes it unsurprising that a non-symmetric divergence underwrites our inaccuracy measure. For to ask how far a credence function is to the truth is different than asking how far an omniscient credence function is to a world with partial truths.

An example may help to illustrate this difference. Suppose we have a credence function, c, defined on $\mathcal{F} = \{P, \overline{P}\}$, and suppose $c(P) = 0.7$ and $c(\overline{P}) = 0.3$. In a world, $w_\star$ where P is true we can ask:

Q1: How far is c from $w_\star$?

But here's a different question. Suppose $c_\star$ is the omniscient credence function at $w_\star$ and so $c_\star(P) = 1$ and $c_\star(\overline{P}) = 0$. Imagine now a different world, w where there are partial truths and where the truth-value of P is 0.7 and the truth-value of $\overline{P}$ is 0.3. We could ask:

Q2: How far is $c_\star$ from w?

Symmetric divergences give the same answer to Q1 and Q2. But I don't think the questions are the same, nor is there reason to think they should receive the same answer. Q1 is asking how accurate a degreed belief state is that is trying to match a world where propositions are either fully true or fully false. Q2 is asking how accurate an all-or-nothing belief state is that is trying to match a world where propositions are partially true and partially false.

10 Weighted Scores and Accuracy-Based Arguments

The weighted score proposal is that accuracy is measured by a score that satisfies the schema below:

$$\mathfrak{S}(w, c) = \textstyle\sum_{X \in \mathcal{F}} \lambda(X)\mathfrak{s}(v_w(X), c(X))$$

One natural question about this proposal is whether the scores that fit this schema will still underwrite the accuracy-based arguments for epistemic norms such as probabilism[25] and conditionalization.[26] To answer this question, first suppose that the weighting function, $\lambda(X)$, is such that for some $X \in \mathcal{F}$, $\lambda(X) = 0$. In that case, certain propositions are completely ignored by our scoring rule. Thus, credence in those propositions could be set to absolutely anything—values above 1 or below 0, for instance—without affecting the accuracy score of the credence function. So, if we allow $\lambda(X)$ to assign certain propositions zero weight, then we won't get the accuracy-based results that rational credence functions must satisfy various epistemic norms.[27]

[25] Probabilism is the claim that every rational credence function is a probability function.

[26] For representative work in this area, see the citations in footnote 2.

[27] Here's a simple example, which shows that we *won't* be able to argue for Probabilism by showing that for every probabilistically incoherent credence function there is a coherent credence function more accurate than it. Suppose $\mathcal{F}$ contains only P and $\overline{P}$. Let $c_1(P) = c_1(\overline{P}) = 1$. This function is incoherent, and it is dominated by (among other functions) $c_2(P) = c_2(\overline{P}) = 0.5$ if all propositions are weighted equally. In a world where P is true, c_1 gets a score of 1 and c_2 gets a score of 0.5, and likewise in a world where $\overline{P}$ is true. But now suppose $\lambda(\overline{P}) = 0$. c_2 no longer dominates c_1, since in a world where P is true, c_1 gets a score of 0 and c_2, a score of 0.25.

However, we do get something weaker. Suppose that our weighting function, $\lambda(X)$ assigns no propositions a weight of zero. Instead, it assigns the propositions that we do not care about a very, very low, but still positive weight. As Pérez Carballo (forthcoming) proves, if $\mathfrak{s}(v_w(X), c(X))$ is a strictly proper local score, then $\sum_{X \in \mathcal{F}} \lambda(X)\mathfrak{s}(v_w(X), c(X))$ is a strictly proper global score so long as $\lambda(X) > 0$ for all X. Hence, if we weight propositions in such a way that none are assigned a weight of zero, we get accuracy-based results for Probabilism (Pettigrew, 2016, Theorem 4.3.4), the Principal Principle (Pettigrew, 2016, Theorem 10.0.1), and for Plan Conditionalization (Pettigrew, 2016, Theorem 14.1.1, due to Greaves and Wallace). The only accuracy-based argument that fails to go through is the accuracy-based argument for the Principle of Indifference (Pettigrew, 2016, Theorem 13.1.1). But if certain propositions are more important than others, we wouldn't expect the Principle of Indifference to be true, since it is worse to be wrong about the more heavily weighted propositions than the less heavily weighted. Further, when we have a weighting such that $\mathcal{G} \subseteq \mathcal{F}$, all $X \in \mathcal{G}$, $\lambda(X) = n > 0$, and where for any $Y \in \mathcal{F} - \mathcal{G}$, $\lambda(Y) = 0$, we do get the Principle of Indifference argument going through for the propositions in $\mathcal{G}$. So this is no objection to using weighted accuracy scores.

11 Conclusion

Fallis & Lewis argue against the Brier score on the basis of principle (M3). I've argued that this principle is false: sometimes conditionalizing on true information can lead to a less accurate credal state. This is especially clear in cases of verisimilitude, where you remove all credence from a mostly true (but false) hypothesis and redistribute it so the bulk of it goes to a mostly false hypothesis. However, I've argued that even in cases where verisimilitude is not a factor (M3) is still mistaken. This is because it is sometimes an overall loss in accuracy when one assigns an extreme amount of credence to one particular false hypothesis. This helps us respond to Oddie's recent objection to proper scoring rules on the basis of their violation of what he calls Proximity. The kinds of accuracy scores that account for this are the proposition-based proper scores, among them the proposition-based Brier and Logarithmic scores. Such accuracy scores are sensitive to verisimilitude when it is relevant and still underwrite the most interesting formal results in the accuracy-first program.

References

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Fallis, D. & Lewis, P. J. (2016). The brier rule is not a good measure of epistemic utility (and other useful facts about epistemic betterness). *Australasian Journal of Philosophy*, *94*(3), 576–590.

Gibbard, A. (2008). Rational credence and the value of truth. In T. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Oxford University Press, vol. 2.

Goldman, A. (1999). *Knowledge in a Social World*. Oxford University Press.

Greaves, H. & Wallace, D. (2006). Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, *115*(459), 607–632.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*(4), 575–603.

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

Knab, B. & Schoenfield, M. (2015, March 12). A strange thing about the brier score. *[Web log post]*, *Retrieved from* `http://m-phi.blogspot.com/2015/03/a-strange-thing-about-brier-score.html`.

Leitgeb, H. & Pettigrew, R. (2010a). An objective justification of Bayesianism I: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201–235.

Leitgeb, H. & Pettigrew, R. (2010b). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science*, *77*(2), 236–272.

Lewis, P. J. & Fallis, D. (2016). Accuracy, conditionalization, and probabilism. `http://philsci-archive.pitt.edu/12517/`.

Oddie, G. (forthcoming). What accuracy could not be. *British Journal for the Philosophy of Science*.

Pérez Carballo, A. (forthcoming). Good questions. In K. Ahlstrom-Vij & J. Dunn (Eds.), *Epistemic Consequentialism*, Oxford University Press.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.

Predd, J. B., Seiringer, R., Lieb, E. H., Osherson, D. N., Poor, H. V., & Kulkarni, S. R. (2009). Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory*, *55*(10), 4786–4792.