Consequences of Calibration

J Robert G Williams and Richard Pettigrew

March 23, 2023

Abstract

Drawing on a passage from Ramsey's Truth and Probability, we formulate a simple, plausible constraint on evaluating the accuracy of credences: the Calibration Test. We show that any additive, continuous accuracy measure that passes the Calibration Test will be strictly proper. Strictly proper accuracy measures are known to support the touchstone results of accuracy-first epistemology, for example vindications of probabilism and conditionalization. We show that our use of Calibration is an improvement on previous such appeals by showing how it answers or sidesteps problems that have been raised for previous work in this area.

In this paper, we offer a new set of axioms that characterise the epistemic utility functions that are most often used in arguments in favour of Bayesian norms such as Probabilism and Conditionalization.[1] These are the additive and continuous strictly proper epistemic utility functions. They purport to provide different ways of measuring how good an individual's credence function is from a purely epistemic point of view, given a particular way the world is. Providing such a characterization is important because of mathematical results like the following: for any additive and continuous strictly proper epistemic utility function, if a credence function does not satisfy Probabilism, then there is another that does satisfy it that is guaranteed to better by the lights of that epistemic utility function. So, if each of the axioms in our characterization is a plausible requirement on a measure of epistemic utility, then, in conjunction with this result, we can offer a strong argument for Probabilism. And there are similar mathematical results that underpin epistemic utility arguments for Conditionalization as well.

Our characterization is based on a suggestion by Frank P. Ramsey and appeals to the virtue of calibration. We begin, in Sections 1 and 2, by describing Ramsey's proposal and making our characterization precise; then, in Section 3, we answer objections inspired by other treatments of calibration in epistemic utility theory.

1 The Calibration Test

In 'Truth and Probability', Frank Ramsey gives the following intriguing account of when a credence is the best one to assign to each proposition in a given set:

> Granting that [an agent] is going to think always in the same way about all yellow toadstools, we can ask what degree of confidence it would be best for him to have that they are unwholesome. And the answer is that it will in general be best for his degree of belief that a yellow toadstool is unwholesome to be equal to the proportion of yellow toadstools that are unwholesome. (Ramsey 1926 [1931], 195)

Let's see precisely what this says. Suppose Sally is contemplating m propositions of the form: Yellow Toadstool i is wholesome. These m propositions constitute her agenda. Now, either because of limits on her time or patience, or because she has no reason to treat them differently, Sally is committed to taking the same doxastic attitude to each member of this set—as we will say, she is committed to adopting a 'homogeneous' credence function on her agenda. Ramsey proposes that, among all homogeneous credence functions over this agenda, the best is the one that assigns the credence $\frac{k}{m}$ to each proposition on it, where m is the total number of yellow toadstools (= total number of propositions on that agenda), and k is the number of those that are in fact wholesome (= total number of true propositions on that agenda). Call this the 'perfectly calibrated' credence function, relative to that set of propositions and those facts about what is true and what is false. Restating Ramsey: given a fixed agenda of propositions and a truth value distribution, the best homogeneous credence function is the perfectly calibrated one. Let's say that a measure of epistemic utility

[1] Arguments for Probabilism: (Savage 1971, Joyce 2009, Predd *et al.* 2009, Pettigrew 2016). Arguments for Conditionalization: (Oddie 1997, Greaves and Wallace 2006, Briggs and Pettigrew 2020, Nielsen 2021).

'passes the Calibration Test' if it upholds Ramsey's judgment here; that is, among the homogeneous credence functions on a given algebra, the one that has greatest epistemic utility at a given world is the one that is perfectly calibrated at that world.

The main result we prove is the following: any additive and continuous epistemic utility function that passes the Calibration Test is strictly proper. We define these terms, and prove the result, in the next section.

2 The result

We start by defining our terms:

- If $\mathcal{F}$ is a finite set of propositions, then:
  - A 'possible world relative to $\mathcal{F}$' is a classical assignment of truth values to the propositions in $\mathcal{F}$. Denote the set of these $\mathcal{W}_\mathcal{F}$.
  - Given a possible world w in $\mathcal{W}_\mathcal{F}$ and a proposition A in $\mathcal{F}$, let
$$w(A) =_{\text{df.}} \begin{cases} 1 & \text{if } A \text{ is true at } w \\ 0 & \text{if } A \text{ is false at } w \end{cases}$$
  - A 'credence function on $\mathcal{F}$' is a function $c : \mathcal{F} \to [0,1]$. Denote the set of them $\mathcal{C}_\mathcal{F}$.
  - A credence function on $\mathcal{F}$ is 'homogeneous' if it assigns the same credence to all members of $\mathcal{F}$. That is, $c(A) = c(B)$ for all A, B in $\mathcal{F}$.
  - Given a world w in $\mathcal{W}_\mathcal{F}$, the 'perfectly calibrated credence function on $\mathcal{F}$' is the one that assigns to each proposition in $\mathcal{F}$ the frequency of truths among all propositions in $\mathcal{F}$. We denote it $b^w_\mathcal{F}$. That is, for all A in $\mathcal{F}$
$$b^w_\mathcal{F}(A) = \frac{|\{X \in \mathcal{F} : X \text{ is true at } w\}|}{|\{X \in \mathcal{F}\}|}$$
- An 'epistemic utility function' is a class of functions
$$\{\mathfrak{A}_\mathcal{F} : \mathcal{C}_\mathcal{F} \times \mathcal{W}_\mathcal{F} \to [-\infty, 0] \mid \mathcal{F} \text{ is a finite set of propositions}\}$$
  A couple of examples:
  - Absolute value measure: given a finite set of propositions $\mathcal{F}$,
$$AS_\mathcal{F}(c, w) = -\sum_{A \in \mathcal{F}} |c(A) - w(A)|$$
  - Brier score: given a finite set of propositions $\mathcal{F}$,
$$BS_\mathcal{F}(c, w) = -\sum_{A \in \mathcal{F}} |c(A) - w(A)|^2$$

  (Many discussions of epistemic utility theory do not make explicit that each epistemic utility function they discuss is really a class of functions, one for each finite set of propositions, but it is always implicit.)

Additivity An epistemic utility measure $\mathfrak{A}$ is 'additive' if there is a scoring rule $\mathfrak{s} : \{0,1\} \times [0,1] \to [-\infty, 0]$ such that, for any finite set of propositions $\mathcal{F}$,

$$\mathfrak{A}_{\mathcal{F}}(c, w) = \sum_{A \in \mathcal{F}} \mathfrak{s}(w(A), c(A))$$

We say that '$\mathfrak{s}$ generates $\mathfrak{A}$'.

Continuity An additive epistemic utility function $\mathfrak{A}$ generated by scoring rule $\mathfrak{s}$ is 'continuous' if $\mathfrak{s}(1, x)$ and $\mathfrak{s}(0, x)$ are both continuous functions of x on $[0, 1]$.

Calibration Test An epistemic utility function $\mathfrak{A}$ 'passes the calibration test' if, for any finite set of propositions $\mathcal{F}$ and any world w in $\mathcal{W}_{\mathcal{F}}$, the perfectly calibrated credence function $b^w_{\mathcal{F}}$ is the best of the homogeneous credence functions on $\mathcal{F}$. That is, $\mathfrak{A}_{\mathcal{F}}(b^w_{\mathcal{F}}, w) > \mathfrak{A}_{\mathcal{F}}(c, w)$ for all homogeneous $c \neq b^w_{\mathcal{F}}$ on $\mathcal{F}$.

Strict Propriety An additive epistemic utility function $\mathfrak{A}$ generated by $\mathfrak{s}$ is 'strictly proper' if, for any $p \neq q$ in $[0, 1]$,

$$p\mathfrak{s}(1, p) + (1 - p)\mathfrak{s}(0, p) > p\mathfrak{s}(1, q) + (1 - p)\mathfrak{s}(0, q)$$

The absolute value measure is additive and continuous, but it does not pass the Calibration Test and it is not strictly proper. The Brier score is additive and continuous, and it passes the Calibration Test and it is strictly proper.

We can now state and prove our main result:

Theorem 1. *Additivity + Continuity ⇒ (Calibration Test ⇔ Strict Propriety)*

Proof. Suppose $\mathfrak{A}$ is an additive and continuous accuracy measure generated by $\mathfrak{s}$.

First, ⇒. Suppose p is a rational number in $[0, 1]$. So there is a positive integer m and a non-negative integer k such that $k < m$ and $p = \frac{k}{m}$. Then it is possible to construct a set of propositions $\mathcal{F}$ and a possible world w such that (i) $\mathcal{F}$ contains m propositions and (ii) k of those propositions are true at w. So $b^w_{\mathcal{F}}(A) = \frac{k}{m}$, for all A in $\mathcal{F}$. Then, by Calibration Test, $\mathfrak{A}_{\mathcal{F}}(b^w_{\mathcal{F}}, w) > \mathfrak{A}_{\mathcal{F}}(c, w)$ for any homogenous credence function $c \neq b^w_{\mathcal{F}}$. But,

$$\begin{aligned}
\mathfrak{A}_{\mathcal{F}}(b^w_{\mathcal{F}}, w) &= \sum_{A \in \mathcal{F}} \mathfrak{s}(w(A), b^w_{\mathcal{F}}(A)) \\
&= \sum_{\substack{A \in \mathcal{F} \\ w(A)=1}} \mathfrak{s}(1, p) + \sum_{\substack{A \in \mathcal{F} \\ w(A)=0}} \mathfrak{s}(0, p) \\
&= k\mathfrak{s}(1, p) + (m - k)\mathfrak{s}(0, p)
\end{aligned}$$

And, if $c(A) = q$ for all A in $\mathcal{F}$,

$$\begin{aligned}
\mathfrak{A}_{\mathcal{F}}(c, w) &= \sum_{A \in \mathcal{F}} \mathfrak{s}(w(A), c(A)) \\
&= \sum_{\substack{A \in \mathcal{F} \\ w(A)=1}} \mathfrak{s}(1, q) + \sum_{\substack{A \in \mathcal{F} \\ w(A)=0}} \mathfrak{s}(0, q) \\
&= k\mathfrak{s}(1, q) + (m - k)\mathfrak{s}(0, q)
\end{aligned}$$

So,

$$\frac{k}{m}\mathfrak{s}(1,p) + \frac{m-k}{m}\mathfrak{s}(0,p) > \frac{k}{m}\mathfrak{s}(1,q) + \frac{m-k}{m}\mathfrak{s}(0,q)$$

That is,

$$p\mathfrak{s}(1,p) + (1-p)\mathfrak{s}(0,p) > p\mathfrak{s}(1,q) + (1-p)\mathfrak{s}(0,q)$$

So, for all rationals p in $[0,1]$ and any q in $[0,1]$,

$$p\mathfrak{s}(1,p) + (1-p)\mathfrak{s}(0,p) > p\mathfrak{s}(1,q) + (1-p)\mathfrak{s}(0,q)$$

Now suppose, for the sake of deriving a contradiction, that there are $r \neq s$ in $[0,1]$ such that

$$r\mathfrak{s}(1,r) + (1-r)\mathfrak{s}(0,r) < r\mathfrak{s}(1,s) + (1-r)\mathfrak{s}(0,s)$$

By Continuity, $x\mathfrak{s}(1,x) + (1-x)\mathfrak{s}(0,x)$ and $x\mathfrak{s}(1,s) + (1-x)\mathfrak{s}(0,s)$ are continuous functions of x. So there is a neighbourhood around r such that, for any x within that neighbourhood,

$$x\mathfrak{s}(1,x) + (1-x)\mathfrak{s}(0,x) < x\mathfrak{s}(1,s) + (1-x)\mathfrak{s}(0,s)$$

But any such neighbourhood will contain rational numbers, since the rationals are dense in the reals. So there is rational p in the neighbourhood such that

$$p\mathfrak{s}(1,p) + (1-p)\mathfrak{s}(0,p) < p\mathfrak{s}(1,q) + (1-p)\mathfrak{s}(0,q)$$

But that contradicts our result above. So, for any real $p \neq q$ in $[0,1]$,

$$p\mathfrak{s}(1,p) + (1-p)\mathfrak{s}(0,p) > p\mathfrak{s}(1,q) + (1-p)\mathfrak{s}(0,q)$$

as Strict Propriety requires.

Second, $\Leftarrow$. Suppose $\mathfrak{s}$ is strictly proper; suppose $\mathcal{F}$ contains m propositions of which k are true at w. Then, if b_q is the homogeneous credence function defined on $\mathcal{F}$ that assigns credence q to every proposition in $\mathcal{F}$, then

$$\mathfrak{A}_{\mathcal{F}}(b_q, w) = \sum_{A \in \mathcal{F}} \mathfrak{s}(w(A), q) = k\mathfrak{s}(1,q) + (m-k)\mathfrak{s}(0,q)$$

And this is maximised by the same q that maximises

$$\frac{1}{m}\left(k\mathfrak{s}(1,q) + (m-k)\mathfrak{s}(0,q)\right) = \frac{k}{m}\mathfrak{s}(1,q) + (1-\frac{k}{m})\mathfrak{s}(0,q)$$

And since $\mathfrak{s}$ is strictly proper, this is maximised at $q = \frac{k}{m}$, as required by the Calibration Test.

□

3 Discussion

So much for the result. In this section, we argue for its significance. To recap: in the first paragraph of this paper, we noted that additive and continuous strictly proper epistemic utility functions support the mathematical 'dominance' results that have been

wielded in philosophical arguments for probabilism, conditionalization and the like.[2] The contribution of this paper, therefore, is a new characterization of that dominance-argument-supporting class of epistemic utility functions.

It would be nice if we could now argue that this paper's three axioms, Additivity, Continuity, and the Calibration Test, are more plausible than other axioms that have been proposed. But there are currently many, many attempts to characterise the class of epistemic utility functions. Some permit all additive and continuous strictly proper functions; others go wider than that, many go narrower, and some consider a different class altogether. We can't hope to survey them all here and show that each is lacking in some way that ours is not. So, instead, we will focus on the Calibration Test, dispel some possible misgivings about it, and explain how it improves on the other characterization of additive and continuous strictly proper epistemic utility functions that appeals to the notion of calibration, namely, Richard Pettigrew's 'new account' of accuracy (Pettigrew 2016, Chapter 4).[3]

First, the misgivings. In his original paper giving an epistemic utility argument for Probabilism, James M. Joyce (1998) discussed the possibility of appealing to calibration. He dismissed it on the grounds that being calibrated is not a virtue for which we should aim.[4] His point is simple. He asks us to consider someone with just two credences, one in a proposition A and the other in its negation $\overline{A}$. And he observes that, for such an individual, assigning a credence of 0.5 to each proposition is guaranteed to be perfectly calibrated; whether A is true or false, exactly half of the propositions about which she has an opinion will be true. Then he asks us to consider someone who assigns credence 0.6 to A and 0.4 to $\overline{A}$. Their credence function seems to be better, epistemically speaking, at the world in which A is true. After all, it assigns a higher credence to the true proposition, A, and lower probability to the false proposition, $\overline{A}$. And yet it is not perfectly calibrated, while the original one is. Put differently: an epistemic utility function that rewards calibration will not be *truth-directed*, where this just means that if credence function c always gives at least as high credence to truths as credence function c', and sometimes higher, and c always gives at most as high credence to falsehoods as c', and sometimes lower, then c is better from an epistemic point of view. Requiring that an epistemic utility function is truth-directed is the heart of Joyce's accuracy-based account of epistemic utility.

However, notice that Joyce's argument is no objection to Ramsey's suggestion and the Calibration Test we extracted from it. The Calibration Test makes no demands on how the epistemic utility function should compare the individual who assigns 0.5 to A and 0.5 to $\overline{A}$, on the one hand, and the individual who assigns 0.6 to A and 0.4 to $\overline{A}$, on the other. It speaks only of comparisons between homogeneous credence functions. So it says that assigning 0.5 to A and 0.5 to $\overline{A}$ is better than assigning 0.6 to A and 0.6 to $\overline{A}$, and similarly for 0.7, 0.972123, and so on. But there is no tension between that and Joyce's requirement that epistemic utility functions should be truth-

[2]For the specific results that suffice in the case of probabilism, see Pettigrew (2016, Thm.4.3.4-5, p.65-6), which relate additive continuous strictly proper scoring rules to additive Bregman divergences, and additive Bregman divergences to dominance.

[3]Our axioms overlap with Pettigrew (2016) on Additivity and Continuity, while replacing his other axioms with the Ramsey-inspired Calibration Test. With only a little tweaking, Pettigrew's stated motivations for Additivity and Continuity (§§4.1,4.2) can be restated in terms of epistemic utility and are no less (or more) persuasive than in the original form. So the main focus in the comparative evaluation below will be on the non-overlapping axioms.

[4]His discussion draws on Teddy Seidenfeld's (1985).

directed. Indeed, it is a remarkable corollary of our characterization of the additive and continuous strictly proper epistemic utility functions that an additive and continuous epistemic utility function that passes the Calibration Test will be truth-directed. That is because, as Mark Schervish (1989, Lemma A.1) shows, any additive and continuous strictly proper epistemic utility function is truth-directed. So Joyce's worries about calibration as a source of epistemic value have no purchase against Ramsey's suggestion.

Second, Pettigrew's 'new account' of accuracy (Pettigrew 2016, Chapter 4). We'll begin by describing it, and then we'll explain how our characterization improves on it by weakening its assumptions and answering objections that have been raised against it.

On Pettigrew's account, accuracy is the sole fundamental source of epistemic value. So epistemic utility functions measure accuracy. Because of this, he assumes that the best credence function at a given world w is the omniscient one v_w, which assigns credence 1 to all truths and credence 0 to all falsehoods. The accuracy of a credence function is then its proximity to the omniscient credence function. This proximity is measured using a divergence function $\mathfrak{D}$, which measures something like the distance between credence functions. So the inaccuracy of a credence function c at a world w is its divergence from the omniscient credence function at that world v_w—that is, it is $\mathfrak{D}(v_w, c)$. Pettigrew (2016, p.63) then assumes that this divergence decomposes into a (weighted) sum of two quantities: first, a measure of how far the credence function c is from c^w, the well-calibrated counterpart of c at w, defined below; and second, a measure of the inaccuracy of that well-calibrated counterpart itself. In symbols, he assumes that there are positive real numbers α, β such that the divergence that generates a legitimate epistemic utility function satisfies:

$$\mathfrak{D}(v_w, c) = \alpha\mathfrak{D}(c^w, c) + \beta\mathfrak{D}(v_w, c^w)$$

And he shows that any epistemic utility function generated from an additive and continuous divergence that decomposes in this way is additive, continuous, and strictly proper.[5]

So what's the well-calibrated counterpart of c at w? If c is defined on $\mathcal{F}$, then c^w assigns to a proposition A the frequency of truths among all propositions in $\mathcal{F}$ to which c assigns the same credence as it assigns to A. In symbols:

$$c^w(A) = \frac{|\{X \in \mathcal{F} : c(X) = c(A) \;\&\; X \text{ is true at } w\}|}{|\{X \in \mathcal{F} : c(X) = c(A)\}|}$$

Pettigrew's motivation for the Decomposition axiom starts from the assumption that all else equal, credence functions are more accurate the nearer they are to their well-calibrated counterparts. Further, other things are only not equal when two credence functions have different well-calibrated counterparts. This goes beyond anything in the passage from Ramsey quoted earlier, but certainly if you agree with Pettigrew's starting point you will like the Calibration Test. After all, any two homogeneous credence functions over the same agenda will have the same well-calibrated counterpart, which will just be the credence function that is perfectly calibrated over the relevant

[5] As Pettigrew mentions, this is a converse to DeGroot and Fienberg's result that any additive, continuous, and strictly proper epistemic utility function decomposes into a calibration component and what they call a refinement component (DeGroot and Fienberg 1983).

set of propositions—and since this perfectly calibrated homogeneous credence is maximally near itself, it will be the most accurate, just as the Calibration Test asserts.

Unsurprisingly, therefore, the Calibration Test follows from Pettigrew's Decomposition axiom. Given the second summand of the decomposition equation is constant across all homogeneous credence functions, we minimize $\mathfrak{D}(v_w, c)$ by minimizing the first summand, and this is uniquely achieved at $c = c^w$.

There are two senses in which the Calibration Test is weaker (and so easier to motivate) than the Decomposition axiom that Pettigrew imposes. The first is that, at least on its face, it is silent about the relative accuracy of homogeneous credence functions other than the perfectly calibrated one that it says is best. For all it says, proximity to the perfectly calibrated one may not be a good guide to accuracy. It only captures a limit case of Pettigrew's stated motivation—just the bit that Ramsey articulates in the quoted passage. The second way in which the Calibration Test is weaker than Decomposition is that it is, again on its face, altogether silent on the relative accuracy of credence functions that do not share the same well-calibrated counterpart—about those situations where, in Pettigrew's terms, all else is not equal. Pettigrew's axiom is highly committal on how things work out in such cases. Decomposition asserts the existence of a specific additive 'fudge factor' which, when combined with the proximity of the credence function to different well-calibrated counterparts, matches their overall accuracy. That particular hypothesis about how accuracy relates to well-calibrated counterparts goes well beyond anything present in the informal motivations. Put more formally, any epistemic utility function that satisfies Decomposition will also satisfy the Calibration Test, but there are epistemic utility functions that satisfy the Calibration Test while not satisfying Decomposition.[6]

It is a remarkable fact that Additivity, Continuity, and Calibration Test alone give everything that Pettigrew's much stronger set of axioms gives; they characterize exactly the same set of epistemic utility functions. For instance, using a trick due to Savage (1971) and developed by Predd *et al.* (2009), you can take an additive and continuous strictly proper epistemic utility function and define a divergence such that epistemic utility is proximity to the omniscient credence function relative to that divergence.[7] And then, by a theorem due to DeGroot and Fienberg (1983), you can show that Pettigrew's Decomposition axiom is true for any such epistemic utility function. So, while Ramsey's Calibration Test seems so much weaker than Decomposition on its face, it can be used to derive that axiom.

So our axiomatization is, on its face, weaker than Pettigrew's, and to that extent more plausible. But it also solves some problems that Pettigrew's axiomatization faces. We'll conclude our discussion by considering two objections raised by Ben Levinstein

[6]We explained above why Decomposition entails the Calibration Test. Now let's see why the Calibration Test does not entail Decomposition. Define an epistemic utility function so that the best credence function at a world is the one that assigns to every proposition the proportion of propositions that are true at that world. Then it will pass the Calibration Test but it will not satisfy Decomposition, for there can be no divergence that satisfies the equality in Decomposition such that the epistemic utility of a credence function at a world is the proximity of the credence function to the omniscient credence function at the world relative to that divergence. After all, the closest credence function to the omniscient one is the omniscient one itself, not the one that assigns the proportion of true propositions.

[7]Pettigrew (2016) describes the trick in Theorem I.B.4. For this reason, Additivity, Continuity, and the Calibration Test together entail the axiom that Pettigrew calls Perfectionism, which says that the epistemic utility of a credence function is its proximity to the perfect one, and the axiom that Pettigrew calls Alethic Vindication, which says that the perfect credence function at a world is the omniscient one that assigns credence 1 to all truths and credence 0 to all falsehoods.

(2017).

Levinstein's Globalism Objection notes an apparent tension between Additivity and Decomposition. It has two parts. The first turns on a putative conflict between Pettigrew's motivation for Additivity and the requirement of Decomposition. Paraphrasing the former, Levinstein writes that Pettigrew claims that 'credence functions are not holistic entities, but simply a way of listing out individual doxastic attitudes which are to be assessed without regard to one another.' Based on this, he objects: 'By appealing to calibration, however, Pettigrew requires precisely the opposite. How we assess what you think about one proposition depends on what you think about other propositions' (Levinstein 2017, Section 2.2).

Note, however, that the same criticism cannot be levelled at the Calibration Test. In Decomposition, the well-calibrated counterpart c^w of c is determined by the credence function c itself, together with the truth values at the world w. And so proximity to its well-calibrated counterpart is a genuinely global feature of a credence function. To determine this feature of a credence function, you can't look only at local features of each credence it assigns; you must look at relationships between them. But in the Calibration Test, whether something is the best homogeneous credence function on an agenda and at a world is determined only by the agenda and the world. Being perfectly calibrated in the relevant sense is not a genuinely global feature of a credence function of the sort to which Levinstein objects. To determine this feature of a credence function, you need only look at local features of each credence it assigns. At a world at which $\frac{k}{m}$ propositions in the agenda are true, we need only ask of each credence whether it is equal to $\frac{k}{m}$.

The same considerations suggest that the Calibration Test is immune to the second part of Levinstein's Globalism criticism. There, he says:

> Additionally, regardless of Pettigrew's motivation for other axioms, invoking calibration in Decomposition undercuts the achievements of accuracy-first epistemology. One goal of AFE is to justify global rational constraints on credence functions that aren't themselves explicitly alethic by appeal to local, alethic evaluations. (Levinstein 2017, Section 2.3)

But, as we have just seen, the Calibration Test does not appeal to a global feature of a credence function. That is, the property of a credence function to which it appeals makes no reference to relationships between the credences that it assigns in the way that the property to which Decomposition appeals does.

Levinstein's Monism Objection points out that Pettigrew is an epistemic value monist, who thinks that the only fundamental source of epistemic value is accuracy, or proximity to the omniscient credence function. But Decomposition seems to appeal to two different sources of epistemic value: proximity to being calibrated, and proximity to omniscience.

Note, however, that the Calibration Test is not subject to this complaint. It appeals only to one feature to characterize epistemic value, calibration, and so in that sense our account of epistemic value is monistic.[8] By appealing to calibration, we can characterise the additive and continuous epistemic utility functions. As we noted above,

[8]Characterizing epistemic value/utility in terms of a single property might not be sufficient for value monism, if we understand value monism to be a thesis restricted to teleological accounts of epistemic value only. So understood, the value monist accounts must give a one-property characterization of ideal credences, and then exclusively appeal to that ideal in motivating the rest of their account. That

these epistemic utility functions are truth-directed. What's more, they can be represented as measuring proximity to the omniscient credence function, so they can be viewed as measures of accuracy. On this way of construing it, our characterization depicts the epistemic value of accuracy as a derived, not fundamental, epistemic value. Calibration first, accuracy second.

Is a monistic calibration-first account of epistemic value attractive? A worry: calibration does not look suited to be the sole source of all epistemic value. Any credal state could be assessed for perfect accuracy, but the Calibration Test covers only a very restricted class of states, the ones that are homogeneous over their agenda. Before we tease out its implications, we have nothing to say about what the best non-homogeneous credences would be like. It seems a peculiar thesis that calibration is a 'source' of epistemic value for non-homogeneous credences to which it has no direct application.

This is fixable: we can consider the general case of an agent who has partitioned their agenda into cells, resolving to be homogeneous across each cell. A Generalized Calibration Test would then tell us that the best credal state consistent with these resolutions is one which the attitude assigned to propositions in a given cell matches the frequency of truth within that cell. There are two limit cases of Generalized Calibration: (i) the agent resolves to treat every proposition on the agenda the same way, so that the whole agenda forms a single cell; (ii) the agent does not resolve to assign any pair of distinct propositions the same attitude, and so every proposition forms its own cell. From type-i cases, we get the Calibration Test; from type-ii cases, we get the thesis that the best credal state simpliciter is the one which matches the omniscient credence function (all relative to the original agenda).[9] Perfect accuracy, à la Joyce and Pettigrew, covers directly only type ii cases, but perfect calibration covers all those cases and many more besides. Calibration, fully spelled out, is a generalization of accuracy. It is just that we did not need to spell it out fully for our formal result—the type-i limit sufficed.

In the light of these last observations, it's not too much of a stretch to conceive of calibration itself as a form of accuracy, just a different and more general conception of accuracy from the one to which Joyce and Pettigrew appeal. It seems to us that Ramsey may be read as explicating one way in which your credences can accurately match the world, namely, by matching the frequency of the truths among the propositions which you have resolved to treat the same. So construed, our characterization is a monistic

is not the shape of our account. In particular (and we bang the table at this point for emphasis) we are definitely not saying that the ideal credences are the well-calibrated ones.

So how should we define value monism? Eyes on the prize: noting that a theory is not value monist would only be an objection to a theory only if there's some reason why theories should be value monist. There's definitely an Ockhamite appeal to doing more with less, and so favouring single-property characterizations of epistemic value over multi-property accounts. But we don't see why it would be an objection to any characterization of epistemic value that it didn't take a teleological form. So in the main text, we'll stick with the looser sense of 'value monism' that isn't analytically teleological. (We thank two referees for pressing us to address this worry).

[9] As a referee pointed out to us, arbitrary resolutions like this can limit an agent to very bad credal states—for example, if they included a tautology and a contradiction in a single cell, then they have thereby ensured they can never be perfectly accurate, or even probabilistic. Generalized Calibration still holds: it promises only to tell us about the epistemically best credal state meeting the relevant constraints, and doesn't imply that credence is particularly good, or rationally permissible. Resolutions that guarantee that any credal state that satisfies them is epistemic-utility-dominated are naturally taken to be rationally defective resolutions to make. But that doesn't mean that we should ignore them!

accuracy-first version of epistemic utility theory.

4 Conclusion

So Additivity, Continuity, and the Calibration Test entail Strict Propriety. This furnishes us with a strong argument that the legitimate epistemic utility functions are the additive and continuous strictly proper ones. That argument is not vulnerable to the objections against calibration raised by Joyce (1998) and Levinstein (2017). With such an argument in hand, the arguments for Probabilism and Conditionalization and other credal norms that appeal to these functions are strengthened.

Acknowledgements

The authors thank four referees for the journal for their very useful feedback and suggestions. Williams acknowledges that the project leading to this paper has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 818633). Pettigrew acknowledges support from a British Academy Mid-Career Fellowship (MF21-210022).

J Robert G Williams
School of Philosophy, Religion and History of Science
University of Leeds
Leeds, UK
j.r.g.williams@leeds.ac.uk

Richard Pettigrew
Department of Philosophy
University of Bristol
Bristol, UK
Richard.Pettigrew@bristol.ac.uk

References

Briggs, R. A. and Pettigrew, R. [2020]: 'An accuracy-dominance argument for conditionalization', *Noûs*, **54**(1), pp. 162–181.

DeGroot, M. H. and Fienberg, S. E. [1983]: 'The Comparison and Evaluation of Forecasters', *The Statistician*, **32**(1/2), pp. 12–22.

Greaves, H. and Wallace, D. [2006]: 'Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility', *Mind*, **115**(459), pp. 607–632.

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

Joyce, J. M. [2009]: 'Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief', in F. Huber and C. Schmidt-Petri (*eds*), *Degrees of Belief*, Springer.

Levinstein, B. A. [2017]: 'Accuracy Uncomposed: Against Calibrationism', *Episteme*, **14**(1), pp. 59–69.

Nielsen, M. [2021]: 'Accuracy-Dominance and Conditionalization', *Philosophical Studies*, **178**(10): pp. 3217-3236

Oddie, G. [1997]: 'Conditionalization, Cogency, and Cognitive Value', *British Journal for the Philosophy of Science*, **48**, pp. 533–41.

Pettigrew, R. [2016]: *Accuracy and the Laws of Credence*, Oxford: Oxford University Press.

Predd, J., Seiringer, R., Lieb, E. H., Osherson, D., Poor, V. and Kulkarni, S. [2009]: 'Probabilistic Coherence and Proper Scoring Rules', *IEEE Transactions of Information Theory*, **55**(10), pp. 4786–4792.

Ramsey, F. P. [1926 [1931]]: 'Truth and Probability', in R. B. Braithwaite (*ed.*), *The Foundations of Mathematics and Other Logical Essays*, London: Kegan, Paul, Trench, Trubner & Co., chap. VII, pp. 156–198.

Savage, L. J. [1971]: 'Elicitation of Personal Probabilities and Expectations', *Journal of the American Statistical Association*, **66**(336), pp. 783–801.

Schervish, M. J. [1989]: 'A general method for comparing probability assessors', *The Annals of Statistics*, **17**, pp. 1856–1879.

Seidenfeld, T. [1985]: 'Calibration, Coherence, and Scoring Rules', *Philosophy of Science*, *52*(2), 274–294.