Confirmation and the ordinal equivalence thesis

Olav B. Vassend

August 4, 2017

Abstract

According to a widespread but implicit thesis in Bayesian confirmation theory, two confirmation measures are considered equivalent if they are ordinally equivalent — call this the "ordinal equivalence thesis" (OET). I argue that adopting OET has significant costs. First, adopting OET renders one incapable of determining whether a piece of evidence substantially favors one hypothesis over another. Second, OET must be rejected if merely ordinal conclusions are to be drawn from the expected value of a confirmation measure. Furthermore, several arguments and applications of confirmation measures given in the literature already rely on a rejection of OET. I also contrast OET with stronger equivalence theses and show that they do not have the same costs as OET. On the other hand, adopting a thesis stronger than OET has costs of its own, since a rejection of OET ostensibly implies that people's epistemic states have a very fine-grained quantitative structure. However, I suggest that the normative upshot of the paper in fact has a conditional form, and that other Bayesian norms can also fruitfully be construed as having a similar conditional form.

1 Introduction

According to Bayesian confirmation theory, evidence E confirms hypothesis H relative to background theory K if and only if $Pr_K(H|E) > Pr_K(H)$.[1,2,3] This criterion suffices to answer the qualitative question of whether or not E confirms H, but it does not answer the quantitative question of how much E confirms H, nor does it answer the comparative question of which of two confirmed hypotheses is confirmed more by E. To answer the quantitative and comparative questions, one must adopt a confirmation measure that quantifies the degree to which E confirms H. The following are just a few of the confirmation measures that have been suggested in the literature:

The plain ratio measure, $r(H, E) = \frac{Pr(H|E)}{Pr(H)}$

The log-ratio measure, $lr(H, E) = \log r(H, E)$[4]

The difference measure, $d(H, E) = Pr(H|E) - Pr(H)$

The log-likelihood measure, $l(H, E) = \log \frac{Pr(E|H)}{Pr(E|\neg H)}$

The alternative difference measure, $s(H, E) = Pr(H|E) - Pr(H|\neg E)$[5]

The Kemeny-Oppenheim measure, $k(H, E) = \frac{Pr(E|H) - Pr(E|\neg H)}{Pr(E|H) + Pr(E|\neg H)}$

Note that the domain of a confirmation measure strictly speaking consists of triples, (H, E, Pr); however, for simplicity I will for the most part suppress mention of Pr. It is well known that confirmation measures do not always order hypothesis-evidence pairs in the same way: the measures are sometimes ordinally non-equivalent.

[1] From now on I will suppress mention of the background theory.

[2] Equivalently, if and only if $Pr(H|E) > Pr(H|\neg E)$ or if and only if $Pr(E|H) > Pr(E|\neg H)$. Disconfirmation and absence of confirmation (neutrality) can be defined analogously.

[3] Of course, Pr_K is assumed to be a probability distribution defined on a Boolean algebra of propositions that includes both H and E.

[4] It is not customary to specify the base of the logarithm.

[5] This measure is also sometimes called the "Joyce-Christensen measure," after Joyce (1999) and Christensen (1999)

For instance, confirmation measures r and d are ordinally non-equivalent since they differ in how they rank certain hypothesis-evidence pairs.[6,7] It is obvious that two confirmation measures that are ordinally non-equivalent ought not be considered the same confirmation measure. In other words, the following is uncontroversial:

> *The Ordinal Non-Equivalence Thesis*: If two confirmation measures are ordinally non-equivalent, then the two confirmation measures are not the same confirmation measure.

I have called the ordinal non-equivalence thesis a "thesis," but perhaps it is more appropriate to call it a truism. The task of this paper will be to investigate the converse of the non-equivalence thesis. Namely,

> *The Ordinal Equivalence Thesis*: If two confirmation measures are ordinally equivalent, then they are the same confirmation measure.

According to the ordinal equivalence thesis (OET), r and lr are the same confirmation measure since, even though they have differing functional forms, they rank all hypothesis-evidence pairs in the same order. The ordinal equivalence thesis has arguably become a widespread tacit – and sometimes explicit – commitment among philosophers who work on Bayesian confirmation theory. For example, Branden Fitelson writes:

> "If two relevance measures are ordinally equivalent, then, as far as we are concerned, they are identical. So, when we say 'according to c', we really mean 'according to any measure ordinally equivalent to c"' (Fitelson, 2007, p. 7n7).

Other philosophers reveal their commitment to OET by treating ordinally equivalent measures as interchangeable, which is only legitimate given OET.[8] For example, David Glass and Mark McCartney write:

[6] Example: $Pr(H) = 0.1$, $Pr(H|E) = 0.9$, $Pr(H') = 0.01$, $Pr(H'|E) = 0.5$. Here H is better confirmed than H' according to d, but H' is better confirmed than H according to r.

[7] Interestingly, the standard measures do correlate fairly well (Tentori et al., 2007).

[8] Numerous conversations I have had with philosophers who work on Bayesian confirmation theory have convinced me that it is standard for philosophers to regard ordinally equivalent measures as interchangeable in general.

> "l satisfies (C4) provided division by zero is equated with infinity. To avoid this, the ordinally equivalent measure proposed by Kemeny and Oppenheim (1952) can be used instead." (Glass and McCartney 2015, p62n4.)

Still other philosophers do not unconditionally commit to the ordinal equivalence thesis, but hold that ordinally equivalent measures *often* are interchangeable. For example, Tomoji Shogenji writes that, "For many purposes, ordinally equivalent measures are essentially the same measure" (Shogenji, 2012, p. 5n4). Shogenji may be right that ordinal equivalence suffices for many purposes, but a major goal of this paper is to show that there are also several purposes for which the ordinal equivalence thesis is too weak.

Here is the plan of the paper. In Section 2, I describe various competing theses that we may choose to adopt; each of these theses corresponds to an alternative level of analysis that we may choose to prioritize. In Section 3, I show the shortcomings of the ordinal equivalence thesis by contrasting it with alternative theses, and in particular I show that adopting the ordinal equivalence thesis renders one unable to set various thresholds that can be used to interpret a set of confirmation scores, and that a thesis stronger than OET must be adopted if merely ordinal conclusions are to be drawn from the expected values of a confirmation measure. In Section 4, I show that several arguments given by philosophers already rely on a rejection of OET. In Section 5, I discuss possible reasons why OET has been accepted. A major reason why philosophers have focused on the ordinal level is probably because a rejection of OET seems to imply that human beings have epistemic states that have a very fine-grained quantitative structure. However, I suggest that the normative upshot of the paper has a conditional form, so that it is only applicable when the antecedent of the conditional applies. I furthermore suggest that other Bayesian norms can fruitfully be understood as having a similar conditional form.

2 Formal characterizations of various equivalence conventions

As is well known in the literature, ordinal equivalence can be formally characterized. More precisely, two confirmation measures are ordinally equivalent if and only if there is a strictly increasing function from each measure to the other. We can state the preceding characterization of ordinal equivalence more formally as follows:

> *Ordinal equivalence characterization*: Confirmation measures c and c' are ordinally equivalent if and only if there is a strictly increasing function, f, such that, for all H and E and all probability distributions over H and E, $c(H, E) = f(c'(H, E))$.

To better understand the ordinal equivalence thesis, it is useful to contrast it with alternative theses that we may instead choose to adopt. Inspired by the above characterization of ordinal equivalence, we can use the following abstract schema to derive alternative equivalence theses:

> *Confirmation Equivalence Schema*: Confirmation measures c and c' are equivalent if and only if there is an invertible function, f, such that $c = f(c')$.

Different confirmation equivalence theses can then be characterized by what requirements they put on f. In theory, we could produce infinitely many theses from the above schema since there are potentially infinitely many requirements we could choose to put on f. Certain theses are of more theoretical interest than others, however. Following Stevens (1946), I will call the theses I consider “ordinal,” “interval,” “ratio,” and “absolute,” where these theses are distinguished by the increasingly strong demands they place on f.

> *Ordinal Equivalence Thesis* (OET): The requirement on f is that it be strictly increasing.

> *Interval Equivalence Thesis* (IET): The requirement on f is that it be strictly increasing and linear.
>
> *Ratio Equivalence Thesis* (RET): The requirement on f is that it be strictly increasing and linear with constant term 0.
>
> *Absolute Equivalence Thesis* (AET): The requirement on f is that it be the identity function.

Adopting OET amounts to carving the set of all possible confirmation measures into classes of ordinally equivalent measures and treating the measures in each class as interchangeable. Similarly, IET and RET carve the space of confirmation measures into classes of measures that are what we might respectively call “interval” and “ratio” equivalent. The fourth thesis, AET, is the strongest possible thesis: its equivalence classes contain only a single confirmation measure each.

My choice of singling out the above four theses is not arbitrary. The first three theses correspond to three of the four “levels of measurement” outlined by Stevens (1946) in the context of scientific measurement. As Stevens points out, the strength of the conclusions one is licensed to draw from data depends on the strength of the measurement scale used. What is true in the case of measurement scales is also true in the case of confirmation measures, as I show in the next section when I discuss the consequences of adopting OET by contrasting it with the consequences of instead adopting IET.

3 Consequences of adopting the ordinal equivalence thesis

In the following two subsections, I discuss general consequences associated with adopting OET and treating confirmation measures as mere ordinal measures. In Section 4, I show why these consequences matter for several of the arguments and applications of confirmation measures that have been discussed in the literature.

3.1 Interpreting a set of confirmation scores

Suppose I give you the results of a 100m race with three runners by listing the order in which the runners finished. Then you are not entitled to say that the difference in performance between the winner and the runner-up is roughly the *same* as the difference in performance between the runner-up and the third-place finisher; nor may you conclude that the winner performed substantially better than the other two runners. The ordinal data with which you have been provided simply does not contain this information. Suppose you learn, however, that the winner is Usain Bolt and that the other two runners are recreational runners. Then you have reason to believe that the winner's performance was in fact much better than the performance of the other two runners. If, on the other hand, you learn that all three runners are recreational runners, you no longer have any reason to think that the winner's performance was substantially better than the performance of the other two runners. Thus, if all you learn about the three competitors are their ordinal ranks, you cannot draw conclusions about their performance relative to each other. You can only infer such conclusions on the basis of further information.

But now suppose that you instead learn the *times* of the three runners. Suppose, for instance, that you learn that the winner's time was 10 seconds while the other two runners finished in 14 and 14.5 seconds. Then you really can say that the winner was substantially better than the other two runners, and moreover you can say that the two losers performed about equally well. To be sure, your conclusions still depend on background knowledge about running and about the time scale used, but the conclusions you are entitled to draw are much more robust than in the case where you are just given ordinal ranks in the sense that the conclusions do not depend sensitively on knowledge about the particular runners.

The above example illustrates the differences between ordinal and interval/ratio scales. If we adopt OET, then the proper way to interpret the numerical outputs of confirmation measures is as ranks. Although the outputs of e.g. the log-ratio measure lr can be any real number, only the ordinal properties of the real numbers are being used. Suppose, for instance, that our favored confirmation measure — call it "m" — outputs the three numbers 0.91, 0.9, and 0.1 for evidence-hypothesis pairs (H, E), (H', E), and (H'', E), respectively. In that case we are entitled to say that E confirms H more than H', and that E confirms H' more than H'', but we cannot say that H' and H'' are confirmed to approximately the same degree by E, or that each is much more highly confirmed than H''. To make any of these claims is to go beyond the merely ordinal properties of 0.91, 0.9, and 0.1.

Indeed, if we adopt OET, then any conclusion we draw from m's output is valid only if it still holds when we choose to use a different ordinally equivalent measure. This is because by adopting OET we agree to treat ordinally equivalent measures as interchangeable. But it is easy to transform our m into an ordinally equivalent measure that instead outputs, say, the numbers 3, 2, and 1 for the above three evidence-hypothesis pairs. All one needs to do is device a suitable strictly increasing function. For example, $g(x) = 1.25x + 0.875$ for $x \leq 0.9$, and $g(x) = 100x - 88$ for $x > 0.9$. Of course, g is not a very "natural" function, but that is beside the point. The point is that g is a strictly increasing (even continuous) function that transforms m into m'; therefore, by OET, m and m' are equivalent confirmation measures. Performing the preceding transformation makes it clear that the only conclusion we are justified in drawing from the data is the ordinal ranking itself, $m(H, E) > m(H', E) > m(H'', E)$.

The situation is different if we adopt one of the other equivalence theses. Suppose we adopt IET instead. Then any other confirmation measure in the same equivalence class as m must be of the form $m' = am + b$, with positive a. Thus, it must be the case that:

$$\frac{m(H,E) - m(H',E)}{m(H',E) - m(H'',E)} = \frac{m'(H,E) - m'(H',E)}{m'(H',E) - m'(H'',E)} \tag{1}$$

In other words, any functional transformation of m allowed by IET preserves *relative interval sizes.* The consequence of this is that while a measure outputting 0.91, 0.90, and 0.1 can be transformed into an interval equivalent measure that instead outputs the values 91, 90, and 10, respectively, it is not possible to transform it into a measure that outputs 3, 2, and 1. Thus, if we have narrowed down the range of confirmation measures to a class of interval equivalent measures and we adopt IET, then we are entitled to draw robust conclusions from the distances between the numbers outputted by our measure. If we adopt IET, then we are no longer merely using the ordinal properties of the real numbers — the difference between 0.91 and 0.9 really *is* smaller than the difference between 0.9 and 0.1.

3.1.1 Setting thresholds with IET

But the fact that interval equivalent confirmation measures preserve relative interval sizes does not yet mean that we are able to conclude that, e.g., H and H' are confirmed to roughly the same degree by E. In order to draw a conclusion of this kind, we need specific knowledge about m's behavior that allows us to determine that H and H' are confirmed to roughly the same degree (by E and E' respectively; of course E and E' may be identical) if and only if $|m(H,E) - m(H',E')| < \delta$, for some (small) δ. In the same way, we can establish a threshold that says that H is substantially better confirmed by E than is H' by E' if and only if $m(H,E) - m(H',E') > \epsilon$ for some suitably chosen ϵ.

Royall (1997, p. 11) does the preceding for the likelihood ratio, $Pr(E|H_1)/Pr(E|H_2)$.[9] He considers the following "canonical experiment": suppose an urn contains either all white balls or else an equal number of white and black balls. Suppose you then draw three balls with replacement and all the balls turn out to be white. Intuitively, this seems to be "pretty strong" evidence that all the balls are white rather than that half of them are black. The likelihood ratio favoring all white balls is in this case 8. Thus, Royall concludes, 8 is the threshold that signifies "pretty strong" evidence (in an everyday context, let us add) favoring one hypothesis over another. Of course, the choice of this particular canonical experiment is somewhat arbitrary, but note that the particular choice of canonical experiment does not matter much if we

[9] Note: the likelihood ratio is not a Bayesian measure of confirmation. Rather, it is a direct measure of the evidential support that one hypothesis enjoys vis-a-vis another one. As Fitelson (2007) points out, the standard Bayesian confirmation measure that agrees with using the likelihood ratio to compare the relative support of two hypotheses is the ratio measure. Thus, implicitly, Royall is setting thresholds for interpreting quantities of the form $\frac{r(H,E)}{r(H',E)}$.

accept IET. A different canonical experiment may have instead yielded 7 or 9, say, as the threshold that signifies "pretty strong" or maybe just "strong" evidence. But fortunately the real numbers 7 and 9 are relatively *close* to each other, and when we adopt IET we make use of these facts about the real numbers. Therefore, nothing significant hinges on choosing either 7 or 8 or 9 as the threshold.

The precise values of the thresholds are therefore not important — in fact, the thresholds ought not be treated too precisely; what a set of thresholds allows us to do is to better interpret a set of confirmation scores. Importantly, IET allows us to use the same threshold throughout the whole confirmation scale. That is because IET implies that the difference $m(H, E) - m(H', E')$ has the same meaning (i.e. describes the same difference in confirmation) regardless of where on the scale $m(H, E)$ and $m(H', E')$ happen to be. This is exactly what (1) guarantees will be the case. And the fact that $m(H, E) - m(H', E') = a$ describes the same difference in confirmation regardless of the values of $m(H, E)$ and $m(H', E')$ allows us to say, given the confirmation scores of two hypotheses, whether the two hypotheses are confirmed to essentially the same degree, or whether one of the hypotheses is better confirmed, or much better confirmed, than the other one.

It is important to appreciate the importance of being able to make these kinds of comparisons between $m(H, E)$ and $m(H', E)$. Indeed, the question of *whether* a piece of evidence confirms one hypothesis more than it confirms another hypothesis is essentially uninformative unless we can also at the very least determine whether the difference in confirmation is substantial or trivial. Indeed, even if we are ultimately mostly interested in the ordinal ranking provided by the confirmation measure, having confirmation scores that are at least on an interval scale prevents us from *over-interpreting* a difference in confirmation score between two hypotheses. If $m(H, E) > m(H', E)$, then E confirms H more than it confirms H', but if the difference between the confirmation scores is small, the inequality may be practically insignificant, especially when measurement error is taken into account: that is, the inherent accuracy of our measurement procedure may be such that, had we repeated our measurement, the new E' could easily be such that $m(H, E') < m(H', E)$.

3.1.2 Setting thresholds without IET?

IET allows us to set thresholds that determine, e.g. whether H and H' are confirmed to roughly the same degree by some piece of evidence. Are there equivalence theses weaker than IET that allow us to do the same thing?

In general, in order to make an assessment of the "distance" between two confirmation scores, we need a function that takes as its input two confirmation scores and outputs a (non-negative) number that represents the distance between the two scores. Suppose we have available some such function, D. In order for us to be able to set up a threshold δ according to which x and y are "approximately equal" if and only if $D(x, y) < \delta$, it needs to be the case that $D(x, y) = a$ means the same thing regardless of what x and y happen to be. Thus, in particular, if $D(x, y) = a$ and $D(z, w) = a$, then it should be the case that the distance $D(x, y)$ means the same thing as the distance $D(z, w)$, so that we can say that $D(x, y) = D(z, w)$. In order for this to be the case, the class of admissible transformations must obey something very analogous to (1). More precisely, in order for it to be legitimate to conclude that $D(x, y) = D(z, w)$ from the fact that $D(x, y) = a$ and $D(z, w) = a$, it needs to be the case that $D(f(x), f(y)) = D(f(z), f(w))$ whenever x, y, z, and w are transformed using any admissible transformation f. Hence, the class of all admissible transformations must satisfy the following equation:

$$\frac{D(x, y)}{D(z, w)} = \frac{D(f(x), f(y))}{D(f(z), f(w))} \tag{2}$$

Thus, given a distance measure D, we can say that two confirmation scores are approximately equal, or that one confirmation score is substantially greater than another confirmation score *only if* we adopt an equivalence thesis according to which only transformations that obey (2) are admissible. Now, given very weak conditions on D, the class of transformations that obey (2) will be a proper subset of the class of all strictly increasing functions.[10] It follows that OET will in general will be too weak to set thresholds. In order for us to be able say anything more specific about

[10] There are several conditions we could put on D. For example, one reasonable requirement is that confirmation measures scores can be arbitrarily close to each other according to D.

how strong of an equivalence thesis is required, more specific assumptions must be made about the distance measure, D.

The most natural and simplest distance measure on the real numbers is arguably the absolute distance metric, $D(x, y) = |x-y|$. If we plug the absolute distance metric into (2) we recover (1). Furthermore, the linear functions are the only functions that obey (1); therefore, all admissible transformations must be linear.[11] It follows that IET is the *weakest* thesis that allows us to set thresholds of the sort discussed above, *provided* the distance measure is the absolute value metric. If some other distance measure is used, then some other thesis than IET may instead (indeed, probably must) be adopted. But in any case, OET is too weak, because any comparison of confirmation scores requires a distance measure, and the distance measure will impose the requirement that the admissible transformations obey (2).

3.2 Taking expectations of confirmation measures

As we shall see later, several applications of Bayesian confirmation theory involve calculating the mathematical expectation of some confirmation measure. In general, the *expected value* of some quantity (random variable), x, that can take values x_1, x_2, $\ldots$, x_n, relative to a probability distribution p, is defined as follows: $\mathrm{E}[x] = \sum_i x_i p(x_i)$.

Taking the expectation of a confirmation measure presupposes that the confirmation measure is not interpreted as a mere ordinal measure, even if we only care about the ordinal properties of the expectation. This is because the fact that two confirmation measures are ordinally equivalent does not entail that their *expectations* will be ordinally equivalent.

To see why this is the case, suppose more generally that we are interested in the expected value of quantities, x, y, z, etc. What kind of scale must x, y, z, etc. be on in order for us to be able to draw the *ordinal* conclusion that, for example, $\mathrm{E}[x] \geq \mathrm{E}[y]$? Clearly, in order for us to be able to draw the conclusion that the expected value of x really is greater than or equal to the expected value of y, it must

[11] The proof that only linear functions obey (1) is trivial and omitted.

be the case that, for every admissible transformation, f, of x and y, it is also the case that $\mathrm{E}[f(x)] \geq \mathrm{E}[f(y)]$. Hence, in order for us to draw merely ordinal conclusions from the expected values of x and y, the class of admissible transformations must satisfy the following requirement:

$$\mathrm{E}[x] \geq \mathrm{E}[y] \implies \mathrm{E}[f(x)] \geq \mathrm{E}[f(y)] \tag{3}$$

But the class of all strictly increasing functions does not satisfy the above requirement.[12] In general, if f is a strictly increasing function, then the following will of course be true:

$$\mathrm{E}[x] \geq \mathrm{E}[y] \implies f(\mathrm{E}[x]) \geq f(\mathrm{E}[y]) \tag{4}$$

However, (4) does not entail (3) unless the following condition also holds:

$$f(\mathrm{E}[x]) \geq f(\mathrm{E}[y]) \implies \mathrm{E}[f(x)] \geq \mathrm{E}[f(y)] \tag{5}$$

But there are many strictly increasing functions that violate (5). Hence x, y, z cannot be on a mere ordinal scale even if we want to draw merely ordinal conclusions from their expected values. In general, we can guarantee that (5) (and therefore also (3)) holds if the class of admissible transformations satisfies the following requirement:

$$f(\mathrm{E}[x]) = \mathrm{E}[f(x)] \tag{6}$$

As it happens, the class of linear functions satisfies (6). Hence, if x, y, and z are on an interval scale, then that is sufficient for us to be able to draw ordinal conclusions from their expected values.[13]

[12] Here is a simple counter-example. Suppose we have the following probabilities: $p(H_1) = 0.5$, $p(H_1|E) = 0.6$, $p(H_1|\neg E) = 0.2$, $p(E) = 0.625$, $p(H_2) = 0.4$, $p(H_2|E) = 0.2$, $p(H_2|\neg E) = 0.7333$. As can be verified, we have: $\mathrm{E}[d(H_1, E)] = 0 = \mathrm{E}[d(H_2, E)]$. However, $\mathrm{E}[d(H_1, E)^3] < \mathrm{E}[d(H_2, E)^3]$. Note that this example assumes that H_1 and H_2 are not exhaustive hypotheses; i.e., there must be at least one other hypothesis, H_3, etc. in the partition of hypotheses.

[13] Indeed, under several reasonable conditions, the class of linear functions is the *only* class that satisfies (6).

4 Applications of Bayesian confirmation measures that rely on a rejection of OET

As I pointed out in Section 1, many philosophers have adopted OET, either explicitly or implicitly. However, there are also many examples in the literature of applications of Bayesian confirmation theory that implicitly rely on a rejection of OET. To the extent that one wants to make arguments of the sort discussed in this section, one must therefore reject OET.

4.1 Case 1: Schlesinger's argument against the difference measure

In Section 3.1, I explained that adopting OET prevents one from being able to set thresholds that can be used to determine whether a given degree of confirmation is strong, moderate or insignificant. As it happens, there are examples of arguments in the literature that implicitly rely on the assumption that such thresholds can be set. In particular, (Schlesinger, 1995, p. 211) presents an argument (repeated and endorsed in Zalabardo (2009)) against the difference measure and in favor of the ratio measure of confirmation. The argument asks us to compare a change in probability from $1/10^9$ to $1/100$ with a change from 0.26 to 0.27. According to Schlesinger and Zalabardo, the first probability shift is intuitively "much greater" than the second one. The ratio measure gets the "right" verdict here, but the difference measure does not. As the argument in Section 3.1 shows, Schlesinger and Zalabardo cannot say that the ratio measure judges the shift from $1/10^9$ to $1/100$ to be "much greater" than the shift from 0.26 to 0.27 unless the ratio measure is interpreted as something more than just an ordinal measure. Schlesinger and Zalabardo are consequently tacitly rejecting OET.

4.2 Case 2: Myrvold's Bayesian account of the virtue of unification

In Section 3.1, I also explained that OET prevents one from being able to say that two confirmation scores are "approximately the same"; only confirmation theses at least as strong as IET enable one to say this (if the absolute distance metric is used to measure distance). However, there are arguments in the literature that rely on the assumption that it is legitimate to talk about two confirmation scores being approximately the same. In particular, Myrvold (2003) (or, more recently, Myrvold (2016)) gives a Bayesian account that purports to show how a unifying hypothesis can sometimes be confirmed more by evidence than a non-unifying hypothesis, and he applies his account to several examples. Myrvold's explanation of the examples relies on the use of both a confirmation measure, c, and a measure of unification U, and he requires that the measures jointly exhibit the following property: If $c(H_1, E_1) \approx c(H_2, E_1)$, $c(H_1, E_2) \approx c(H_2, E_2)$, and $U(E_1, E_2; H_1) > U(E_1, E_2; H_2)$, then $c(H_1, E_1 \& E_2) > c(H_2, E_1 \& E_2)$. As argued earlier, the use of approximation signs requires that the confirmation measures not be interpreted as mere ordinal measures. Myrvold's account can be salvaged even with OET if the approximation signs are replaced by equality signs. But in that case the unrealistic assumption must be made that H_1 and H_2 are independently confirmed to *exactly* the same degree by the evidence.

The next case I will consider comes from Fitelson (1999). Fitelson shows how several arguments given in the literature are sensitive to the choice of confirmation measure because the arguments depend crucially on properties that some measures have but others lack. According to Fitelson, the problem is that these arguments rely on properties that vary between ordinally non-equivalent measures. In the following, I will show that one of the arguments also implicitly relies on a rejection of OET.

4.3 Case 3: The Gillies-Popper-Miller argument

Gillies's (1986) reconstruction of an argument due to Popper and Miller (1983) depends on the confirmation measure used having the following decomposition property: $c(H, E) = c(H \vee E, E) + c(H \vee \neg E, E)$.[14] According to Gillies, this decomposition allows us to neatly separate H's confirmation score into a deductive part and an inductive part. Redhead (1985) points out that not all measures have the preceding decomposition property, and Fitelson (1999) notes that the Gillies-Popper-Miller argument is consequently sensitive to the choice of confirmation measure. More precisely, Redhead points out that the confirmation measure r does not have the decomposition property. Gillies responds to Redhead's criticism by claiming that r is a flawed confirmation measure, and that d, which does have the decomposition property, is better. In response, Fitelson points out that l also lacks the decomposition property. Presumably, Gillies could respond to Fitelson by claiming that l, too, is a flawed measure of confirmation. However, we can make the further observation that there are measures ordinally equivalent to d that do not have the preceding decomposition property. For example, the measure d^3, which of course is ordinally equivalent to d, does not have the decomposition property. Thus, Gillies's argument does not merely rely on d's being better than r and l; it implicitly relies on d^3's *not* being a good measure of confirmation. Since d and d^3 are ordinally equivalent, Gillies's argument is implicitly rejecting OET. More generally, the preceding discussion shows that the decomposition property is not necessarily shared by confirmation measures that are ordinally equivalent. Hence, anyone who proposes an argument that relies on the decomposition property will quite likely have to reject OET.

The next case I will consider concerns how the Paradox of the Ravens has been handled in the literature.

4.4 Case 4: Solutions to the Paradox of the Ravens

The Paradox of the Ravens is a paradoxical conclusion that arises from the combination of two very reasonable premises: Nicod's Criterion and the Equivalence Condition. Nicod's Criterion says that universal generalizations of the form $\forall x(Ax \rightarrow Bx)$ are confirmed by instances of the form $Ac \& Bc$. The Equivalence Condition says that if e confirms S, then e confirms every sentence logically equivalent to S. To-

[14] I thank a referee for helpful comments on this paragraph.

gether, the Equivalence Condition and Nicod's Criterion entail a conclusion that seems counter-intuitive, namely that a non-black non-raven confirms the proposition that every raven is black. Since Nicod's Criterion and the Equivalence Condition are widely accepted, the standard solution[15] is to embrace the paradoxical conclusion while explaining it away by conceding that a non-black non-raven confirms the proposition that every raven is black, but only to a "minute degree" (Vranas, 2004) in ordinary circumstances.

Standard solutions that have been given to the Paradox of the Ravens clearly violate OET. For example, Fitelson and Hawthorne (2004, pp. 31-7) give a quantitative solution that depends crucially on the non-ordinal properties of the likelihood ratio, l. In particular, their Theorem 4 (p. 34) gives a bound on the ratio of two likelihood ratios that can be violated if we transform the two likelihoods into ordinally equivalent measures by the method I used earlier on p. 8.

In general, quantitative solutions to the Paradox of the Ravens inevitably reject OET. However, there are also non-quantitative solutions to the Paradox of the Ravens. These solutions have the more modest goal of showing that a non-black non-raven confirms the proposition that all ravens are black *less* than a black raven does, without making the quantitative claim that the confirmation is much less. Since these solutions only make ordinal claims, they do not rely on a violation of OET. However, a proper solution to the Paradox of the Ravens arguably *should* be quantitative. As an analogy, suppose I ask you why the sun looks the size of a tennis ball even though it is so far away, while a tennis ball looks tiny from just 100 yards away. If you answer that it is because the sun is bigger than a tennis ball, you have given me relevant information, but you have not really provided an adequate explanation. Similarly, our intuition in the Paradox of the Ravens is that a non-black non-raven should (in most circumstances) barely, if at all, confirm the proposition that all ravens are black, or at least that it should confirm this proposition much less than a black raven does. A proper solution to the Paradox of the ravens should entail this conclusion, and therefore cannot be just ordinal.

[15] Which, of course, is not the only solution. See Rinard (2014) for instance.

4.5 Case 5: The use of mathematical expectations

Several uses to which confirmation measures have been put rely on taking expected values of confirmation measures. Here I will discuss just two such applications.[16]

First, a confirmation score tells you how much a piece of evidence confirms a single hypothesis. However, it's also often interesting to know how much the evidence influences the whole partition of hypotheses; or, in other words, how big the divergence is between the posterior distribution and the prior distribution, given the evidence. The natural way to generalize a confirmation measure to a divergence measure is by taking an expectation. For example, Crupi and Tentori (2014) suggest the following definition:

$$\text{InfDis}(p(\mathbf{H}|E), p(\mathbf{H})) = \sum_j c(H_j, E) * p(H_j|E) \tag{7}$$

Plugging different confirmation measures into (7) then gives rise to different divergence measures. For example, plugging in the log-ratio measure gives rise to the well known KL divergence (Kullback and Leibler, 1951). Conversely, any divergence measure may be regarded as an implicit generalization of a confirmation measure. Divergence measures such as the KL divergence have been applied in many ways in the Bayesian literature. For example, they form the foundations of one of the most prominent versions of objective Bayesianism (Bernardo, 1979).

Crucially, confirmation measures that are ordinally equivalent will in general not give rise to divergence measures that are ordinally equivalent, for the reasons given in Section 3.2. Indeed, ordinally equivalent confirmation measures can give rise to very different divergence measures. Consider, for example, the log-likelihood measure and the Kemeny-Oppenheim measure. These are ordinally equivalent, but the log-likelihood measure judges distances between probabilities that are close to 0 or 1 to be much larger than does the Kemeny-Oppenheim measure, because the log-likelihood measure is unbounded while the Kemeny-Oppenheim measure is bounded between 0 and 1. Thus, the log-likelihood measure and the Kemeny-Oppenheim measure

[16] For examples of other applications, see Good (1985).

give rise to divergence measures that will often ordinally disagree if the probabilities involved are extreme (close to 0 or close to 1).[17] Hence, if we want to be able to draw merely ordinal conclusions from Bayesian divergence measures, we cannot treat the confirmation measures on which they are based as mere ordinal measures.

Second, as has recently been pointed out by Brössel and Huber (2014), confirmation measures also have an application in experimental design. More precisely, from a Bayesian point of view, the best experiment to conduct is the one that can be expected to have the greatest onfirmational impact, where the expectation is calculated over the prior probabilities of the possible evidence, given the candidate experimental design. The confirmation measure that is standardly used (implicitly) for this purpose in the literature on Bayesian experimental design is the log-ratio measure. Brössel and Huber instead use as their illustration the Kemeny-Oppenheim measure of confirmation. I. J. Good, on the other hand, advocated using the log-likelihood measure for the same purpose (Good, 1985). Interestingly, as was just pointed out, the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent, and are for that reason generally regarded as equivalent in the philosophical literature. However, as we have seen, the fact that the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent does not imply that their *expectations* will be ordinally equivalent; hence, the experiment that maximizes expected confirmation with respect to the log-likelihood measure will in general not be equivalent to the experiment that maximizes expected confirmation with respect to the Kemeny-Oppenheim measure—let me hasten to add that neither Brössel and Huber nor Good claim that the expectations of these measures are equivalent.

5 Methodological and Concluding Remarks

Carnap (1962) first drew the distinction between the "comparative" and "quantitative" questions of confirmation. As the previous sections make clear, we can draw finer distinctions than that. In particular, the interval level occupies an intermediate

[17] Numerical examples are easy to come up with, but tedious. Note also that if there are many hypotheses, then at least some of the probabilities *must* be small.

position between the merely comparative (ordinal level) and the fully quantitative (ratio level). For whatever reason, the comparative question analyzed on the ordinal level has become the question that occupies philosophers' attention. Why is that? One possibility is that some philosophers simply believe the ordinal level to be the most interesting level of analysis. I do not think this belief is warranted, but even if it is granted, the arguments in the previous sections show that the quantitative levels of analysis are by no means uninteresting. Several well known arguments that make use of confirmation measures implicitly rely on a rejection of OET. Moreover, many conclusions that we want to draw from the output of a confirmation measure may only be legitimately drawn if the measure is assumed to be more than a mere ordinal measure.

A different reason why philosophers may have focused on the ordinal level of analysis is that they think the question of which confirmation is ordinally correct must be settled before the more fine-grained question of which confirmation measure has the right quantitative structure can be approached. Although this idea seems intuitive, it is mistaken. Indeed, if we instead start with the desideratum that we want a confirmation measure that we can interpret as, say, an interval measure and not just an ordinal measure, then that puts significant restrictions on the functional form that the confirmation measure can take, as argued in Vassend (2015).

Indeed, focusing on the interval level leads to a very different perspective on confirmation measures. By definition, each class of interval equivalent confirmation measures is a proper subset of a class of ordinally equivalent measures; but even so, it is possible for two ordinally non-equivalent confirmation measures to exhibit quantitative behavior that is more similar than the quantitative behavior exhibited by two measures that are ordinally equivalent. For example, from a quantitative point of view, the log-likelihood measure and the log-ratio measure are arguably "more similar" to each other than the log-likelihood measure is to the Kemeny-Oppenheim measure, even though the latter two measures are ordinally equivalent and the first two are not, because the log-likelihood measure and the log-ratio measure will often have numerically similar outputs.[18] One consequence of this is that the log-ratio

[18] In particular, if the hypothesis space is large, it will generally be the case that $p(E|\neg H) \approx p(E)$,

and log-likelihood measures arguably give rise to divergence measures that are more similar to each other than are the divergence measures derived from the log-likelihood measure and Kemeny-Oppenheim measure. Focusing only on the ordinal level of analysis therefore leads us to neglect quantitative similarities and dissimilarities that cut across ordinally equivalent classes.

A final probable reason why philosophers have focused their attention on the ordinal level of analysis and have implicitly accepted OET is that many philosophical Bayesians are subjective Bayesians who hold that an agent's probability function is supposed to represent the degrees of belief of the agent. It is already controversial whether agents' degrees of belief have the kind of quantitative structure that probability functions have. Several philosophers have endorsed an "anti-realism" about probabilistic representations of belief states (e.g. Easwaran (2016)).

To reject OET for confirmation measures is apparently to contend that agents' epistemic states have an even more fine-grained structure than is attributed to agents according to probabilism. If, for example, IET is accepted, then not only do rational agents have degrees of belief that are representable by probabilities; all differences between differences (according to some confirmation measure) of an agent's probability function also represent actual features of the epistemic state of the agent. For Bayesians who already worry about the psychological realism of probabilistic degrees of belief, the complex structure seemingly attributed to agents' epistemic states according to IET may be a bridge too far.

On the other hand, it is undoubtedly the case that we sometimes do have quantitative intuitions about confirmation, so there is a basis in human epistemological

for most H's, and hence the log-likelihood measure and log-ratio measure will have numerically similar outputs. Indeed, if the hypothesis space is parameterized by a continuous parameter, Θ, then, for every $\theta \in \Theta$, we have $l(\theta, E) = \log \frac{Pr(E|\theta)}{Pr(E|\neg\theta)} = \log \frac{Pr(E|\theta)}{\int_{\Theta*} Pr(E|\theta)Pr(\theta)d\theta}$, where $\Theta*$ is Θ with θ taken out. But removing a single point from the parameter space will not have any effect on the integral, so $\int_{\Theta*} Pr(E|\theta)Pr(\theta)d\theta = \int_{\Theta} Pr(E|\theta)Pr(\theta)d\theta = Pr(E)$. Therefore, $l(\theta, E) = \log \frac{Pr(E|\theta)}{Pr(E|\neg\theta)} = \log \frac{Pr(E|\theta)}{Pr(E)} = lr(\theta, E)$, and so $l(\theta, E)$ is actually *identical* to $lr(\theta, E)$ when the hypothesis space is continuous. As far as I know, this fact has not been noted before. On the other hand, the fact that the Kemeny-Oppenheim measure and the log-likelihood measure are ordinally equivalent means that they will always agree on whether $c(H, E) > c(H', E)$, but they will often strongly disagree on whether the *difference* between $c(H, E)$ and $c(H', E)$ is small, large, or trivial; their *interval* judgments are in other words quite different.

experience for looking at the quantitative structure of confirmation measures. For example, in the case of the paradox of the ravens, our intuition is that – in ordinary circumstances – a non-black non-raven confirms the the proposition that all ravens are black *much less* than a black raven does. And we often feel that a piece of evidence fails to really discriminate between two hypotheses, so that the hypotheses are intuitively confirmed to roughly the same extent.

Of course, the fact that we sometimes have strong quantitative intuitions about confirmation does not mean we always do. But nor, should it be added, do we always have strong ordinal intuitions. The Bayesian framework idealizes away these human limitations, but the norms of Bayesianism presumably still hold for more limited agents whenever the norms are applicable. Indeed, Bayesian norms, more generally, can fruitfully be construed as conditional norms. For example, even though human beings lack a completely ordered set of degrees of beliefs, Bayesian Dutch book arguments tell you that, provided you do have degrees of belief and you intend to use them in order to choose which bets to accept or reject, and you want to avoid sure losses, then your degrees of belief need to be probabilistic. Accepting this norm does not entail believing that your degrees of belief are always probabilistic or even that it is always rationally required of you to have credences that are probabilistic.[19] However, if you make it your goal to use a set of credences to choose how to act, then accepting the norm implies you have to accept that the credences that are relevant to your actions, at least, ought to be probabilistic.

Accuracy-based arguments for probabilism may reasonably be construed as establishing a similar conditional norm: if your goal is to have accurate credences in a set of propositions, then your credences in those propositions need to be probabilistic.[20] But if you do not care about the truth value of some proposition, or you are not attending to the proposition, then the conditional norm does not apply to you with respect to that proposition.

In the same way, the arguments in this paper establish the following conditional

[19] Of course, many Bayesians want to argue for this stronger unconditional norm as well.

[20] Of course, philosophers often want to go further; they want to say, for example, that you *ought* to have the goal of avoiding sure losses or having accurate credences.

norm: if you have a set of confirmation scores and you intend to interpret the scores in a certain way (e.g. to say that some of them are approximately the same) or use them in a certain way (e.g. take their expectations), then your confirmation scores cannot be on just an ordinal scale. This conditional norm only "kicks in" if you use your confirmation measure in certain ways, and accepting the norm does not entail believing that your confirmation judgments always will be or always ought to be on an interval scale. The norm therefore does not make unrealistic presuppositions about human psychology, nor does it make unreasonable demands.[21]

References

Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.

Brössel, P. and Huber, F. (2014). Bayesian Confirmation: A Means With No End.

Carnap, R. (1962). *Logical Foundations of Probability.* Chicago: University of Chicago Press, second edition.

Christensen, D. (1999). Measuring Confirmation. *Journal of Philosophy*, XCVI:437–61.

Crupi, V. and Tentori, K. (2014). Measuring Information and Confirmation. *Studies in the History and Philosophy of Science*, 47:81–90.

Easwaran, K. (2016). Dr. Truthlove or: How I learned to Stop Worrying and Love Bayesian Probabilities. *Nous*, 50(4):816–853.

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, 66:S362–78.

[21] Thanks to Branden Fitelson, Malcolm Forster, Elliott Sober, and Mike Titelbaum for reading a draft of this paper. Thanks also to the audience at a presentation of an earlier version at the 2014 pacific APA, in particular Brad Armendt, Kenny Easwaran, Sam Fletcher, and Greg Gandenberger. Finally, thanks to several anonymous reviewers at *Synthese*.

Fitelson, B. (2007). Likelihoodism, bayesianism, and relational confirmation. *Synthese*, 156:473–489.

Gillies, D. (1986). In Defense of the Popper-Miller Argument. *Philosophy of Science*, 53:110–13.

Good, I. J. (1985). Weight of evidence: A brief survey. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 249–270. Elsevier Science Publishers.

Joyce, J. (1999). *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press.

Kemeny, J. G. and Oppenheim, P. (1952). Degree of Factual Support. *Philosophy of Science*, 19(4):307–324.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Myrvold, W. (2003). A Bayesian Account of the Virtue of Unification. *Philosophy of Science*, 70(2):399–423.

Myrvold, W. (2016). On the Evidential Import of Unification. Unpublished manuscript.

Popper, K. and Miller, D. (1983). The impossibility of inductive probability. *Nature*, 302:687–88.

Redhead, M. (1985). On the Impossibility of Inductive Probability. *The British Journal for the Philosophy of Science*, 36(2):185–191.

Rinard, S. (2014). A new bayesian solution to the paradox of the ravens. *Philosophy of Science*, 81(1):81–100.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm.* CRC Press.

Schlesinger, G. (1995). Measuring degrees of confirmation. *Analysis*, 55:208–12.

Shogenji, T. (2012). The Degree of Epistemic Justification and the Conjunction Fallacy. *Synthese*, 184(1):29–48.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684):577–80.

Tentori, K., Crupi, V., and Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, 103:107–119.

Vassend, O. B. (2015). Confirmation Measures and Sensitivity. *Philosophy of Science*, 82(5):892–904.

Vranas, P. (2004). Hempel's Raven Paradox: A Lacuna in the Standard Bayesian Solution. *British Journal for the Philosophy of Science*, 42:393–401.

Zalabardo, J. (2009). An Argument for the Likelihood Ratio Measure of Confirmation. *Analysis*, 69:630–5.