Philosophy of Science (2023), **90**, 1173–1182
doi:10.1017/psa.2023.15

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

CONTRIBUTED PAPER

Absolutely Zero Evidence

Veronica J. Vieland[1,2] and Sang-Cheol Seok[2]

[1]The Ohio State University, Columbus, OH, USA and [2]Mathematical Medicine LLC, Chicago, IL, USA
Corresponding author: Veronica J. Vieland; Email: Veronica.Vieland@MathMed.org

(Received 06 January 2023; accepted 16 January 2023; first published online 17 February 2023)

Abstract

Statistical analysis is often used to evaluate the strength of evidence for or against scientific hypotheses. Here we consider evidence measurement from the point of view of representational measurement theory, focusing in particular on the 0-points of measurement scales. We argue that a properly calibrated evidence measure will need to count up from *absolute* 0, in a sense to be defined, and that this 0-point is likely to be something other than what one might have expected. This suggests the need for a new theory of statistical evidence in the context of which calibrated evidence measurement becomes tractable.

1. Introduction

Statistical analysis is common throughout the biological and social sciences, and certain statistical outputs are routinely understood in evidential terms. The most commonly used evidence statistic (ES) is the empirical p-value (P).[1] Small values of P are routinely said to indicate strong *evidence* against a null hypothesis, with *evidence strength* taken to be stronger the smaller the value of P, and failure to achieve sufficiently small P upon replication is interpreted as indicating *evidence* against the initial finding.

There is a robust literature arguing against interpreting P as a measure of evidence. Thus the very persistence of the practice suggests that strength of statistical evidence is something scientists often *want* to measure. Indeed, it is something that they *do* measure, whenever they treat the numerical value of an ES as representing evidence strength. The question is, does any given ES measure evidence in a meaningful way? Here we invoke representational measurement theory (Hand 2004) in considering this question, focusing on one particular aspect of measurement scale, namely, the 0-point.

Vieland (2017) proposed that the first step toward a properly calibrated ES was development of a well-behaved *empirical* measure, that is, one that patently behaves

[1] Other familiar ESs include the simple likelihood ratio (SLR), maximum LR (MLR), and Bayes factor (BF); see following text for additional details.

the way evidence behaves, and she proposed a novel ES (the RLR, see following text) that seemed to fit the bill, at least in certain simple cases. But establishing empirical measurement devices is merely a first step toward actual calibration. In this article we attempt to take a second step, focusing on one particular aspect of RLR that seems, on the face of it, peculiar.

The remainder of this article is organized as follows. Section 2 provides preliminary details regarding the ESs to be considered in what follows. In section 3 we review distinctions among types of measurement scales, and briefly consider the scale types of familiar ESs. The issue of the 0-point for a measurement scale arises in this context. Section 4 considers the 0-points of BF and RLR, which leads to a counterintuitive conclusion regarding the nature of 0 evidence. In section 5 we consider the concept of absolute 0 as it arises in connection with measurement of temperature to illustrate that any dissatisfaction remaining at the end of section 4 may simply reflect the fact that, as yet we lack a suitable theory of statistical evidence, without which the issue of evidence calibration is moot.

2. Preliminaries

In ordinary parlance, the term "evidence" has (at least) two usages: It can refer to the observational inputs to inference but is also used to refer to a particular relationship between data and hypotheses. This relationship is sometimes referred to as support (Hacking 1965) or relative support (Edwards 1992), but more commonly it is referred to as *evidence* or *weight of evidence*. In what follows, we will use *evidence* in this latter sense, referring to a relationship between hypotheses on given data in the context of a statistical model.

Following Vieland (2017) we will illustrate throughout with coin tossing, with probability θ that the coin lands heads and data D = (*x*,*n*), where *x* = observed number of heads on *n* tosses. There is a simple formula in this case for the probability of D as a function of θ: $P_\theta(D) = \binom{n}{x}\theta^x(1-\theta)^{(n-x)}$. While for actual coin-tossing, *n* can take only integer values, in what follows we treat *n* as continuous.

Suppose we are interested in comparing the two hypotheses H_1:θ < ½ (coin is biased toward tails) and H_2:θ = ½ (coin is fair). Following Hacking (1965), we will assume that a fundamental quantity in any treatment of statistical evidence is the simple LR (SLR), where in our example $SLR(\theta|D) = \frac{(\theta_1)^x(1-\theta_1)^{(n-x)}}{\left(\frac{1}{2}\right)^n}$ for any given value of θ_1. One practical question is how to deal with the composite H_1, which allows for multiple values of θ_1. Perhaps the most widely used approach is the maximum LR (MLR), $MLR(\theta|D) = \frac{(\hat{\theta})^x(1-\hat{\theta})^{(n-x)}}{\left(\frac{1}{2}\right)^n}$, where $\hat{\theta} = \frac{x}{n}$, the maximum likelihood estimate of θ. But MLR has the (arguably fatal) flaw of not permitting evidence in favor of H_2. Another commonly used approach is the BF (Kass and Raftery 1995), which in this case reduces to $BF = \int LR(\theta|D)f(\theta)d\theta$, where $f(\theta)$ is the prior probability distribution for θ. Applying a uniform prior, BF is proportional to the simple average LR (ALR), where the average is taken across all possible values of θ in the numerator.

Vieland (2017) proposed another evidence measure, RLR = MLR/ALR. In what follows, we contrast RLR with BF, to make some philosophical points about the

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

development and validation of measurement scales. But before this can be done, some measurement theoretic distinctions are in order.

3. Measurement scale types

Measurements can be classified into different scale types, which can be characterized in (at least) two useful ways. The first is by asking what we can do *with* them. Suppose test subjects are asked to make a judgment of "more beautiful" or "less beautiful" in a series of pairwise comparisons among n pictures of different faces. This allows a rank-ordering of the pictures from least beautiful to most, and we can assign numbers (say, 1 to n) to represent this ordering. We can now compare faces with respect to rank-ordering, for example, we can say "face #100 is judged to be more beautiful than #99." However, it is *not* meaningful to ask whether the amount by which #100 is more beautiful than #99 is greater than the amount by which #50 is more beautiful than #49, or than #2 for that matter. This is because nothing in the way we have constructed the scale allows us to interpret distances from one number to another in terms of underlying units of beauty. This illustrates *ordinal* measurement. What we can do *with* ordinal scales is to make comparisons regarding order, and nothing more.

A second way to characterize scale types is by asking what we can do *to* them. For ordinal variables, we can transform the original scale into any other set of symbols that preserves rank-ordering, for example, we can replace our scale values 1,.,n with their respective logarithms. As long as the transformation preserves rank-order, it preserves the meaning of the original scale.

The logarithmic transformation would no longer be meaning-preserving, however, if we interpreted the difference between numbers on the original scale as having meaningful units. For instance, the Fahrenheit temperature scale provides not only a rank-ordering of temperatures but also assigns meaning to the unit: The difference in temperature between 100°F and 99°F is the same as the difference in temperature between 50°F and 49°F. We can express this by saying that 1°F always "means the same"[2] with respect to temperature. This illustrates an interval scale. What we can do *with* interval scales is to meaningfully make comparisons of order and also of differences. As for what we can do *to* them, interval scales are amenable to any linear transformation (e.g., the formula converting °F to °C), because such transformations preserve both rank-order and a constant meaning for the unit across the scale range.

A logarithmic transformation would disrupt this thermal meaning. $\text{Log}_e(100°\text{F}) - \log_e(99°\text{F}) = 0.01$ while $\log_e(50°\text{F}) - \log_e(49°\text{F}) = 0.02$. Thus the same change in temperature (1°F) becomes represented by different numbers on the logarithmic scale. Application of a nonlinear transformation to measurements made on an interval scale results in a "rubber scale," for which the meaning of the unit changes across the range of the scale (Houle et al. 2011). Clearly, comparing differences on a rubber scale is problematic, in much the same way that comparing differences on an ordinal scale is problematic.

Ratio scales are interval scales with one additional feature: they count up from 0. Virtually all fundamental measurements in the physical sciences are on ratio scales,

[2] This is Hacking's (1972) phrase, from a passage critiquing the LR as an ES because no argument exists to show that a unit change in the LR always "means the same."

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

Table 1. Overview of measurement scale types

Scale Type	Range	Examples	Meaningful Comparisons	Permissible Transformations
Ordinal	ordered symbols	personal preference	order	rank-order preserving
Interval	real numbers	dates; temperature in °F or °C	order, differences	linear
Ratio	positive real numbers	length; mass; temperature in °K	order, differences, ratios	multiplication by a positive constant

including length, weight, mass, and so forth. Measurements made on ratio scales can be compared with respect to order, differences, and ratios. The °Kelvin has ratio scale type, which means that 20°K is twice as hot as 10°K. By contrast, 20°F cannot be meaningfully said to be twice as hot as 10°F. The only arbitrary feature of a ratio scale is the size of the degree, that is, the amount of change in the object of measurement that we choose to assign to a one unit change on the measurement scale. Thus for ratio scales, the only meaning-preserving transformation is multiplication by a positive constant. Note that ratio scales do not require that the 0-point be attainable; 0 may be merely a limiting value of the scale, never achieved in practice[3] (see also following text).

Table 1 (modified from Houle et al. 2011) summarizes the major scale types.

We also note one relevant variation, the *signed* ratio scale, where sign is used by convention to indicate direction. Houle et al. (2011) gives the example of a ratio-scaled measure of left-right symmetry, with sign used to indicate whether symmetry is measured from left to right or right to left; physical work is also measured on a ratio scale under the convention that positive/negative values indicate work done by or to a system, respectively.

What can we say about the measurement scale of BF? When we interpret larger values of BF as stronger evidence in favor of H_1, we are treating BF as providing a rank-ordering of evidence, that is, as being at least ordinally scaled. In fact, we seem to consider BF to be merely ordinal because a common substitution for BF in reporting results is logBF.[4] (The same is true for P and MLR.) But logarithmic transformations of interval or ratio scaled variables create rubber scales, on which differences between measurement values are no longer meaningful. We would wager that virtually everyone interprets a change in BF from, say, 2 to 4 as representing less of a change in the evidence than a change in BF from 4 to 20. This implies that we believe we can meaningfully compare the difference 4–2 with the difference 20–4, or perhaps the ratios 4/2 with 20/4. If we are on an ordinal or rubber scale, however, neither type of comparison is supportable.

Ideally, we would like to be able to say when one study's evidence is twice as strong as another's. Thus we assume we would prefer, if possible, to measure evidence on a

[3] The technical term for 0 as a limiting value is "infimum." We will also use the expression "lower bound" when referring to the bottom of a scale without presupposing whether its minimum value can be achieved or not.

[4] Here "log" may refer to the base2, base10, or natural scale.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

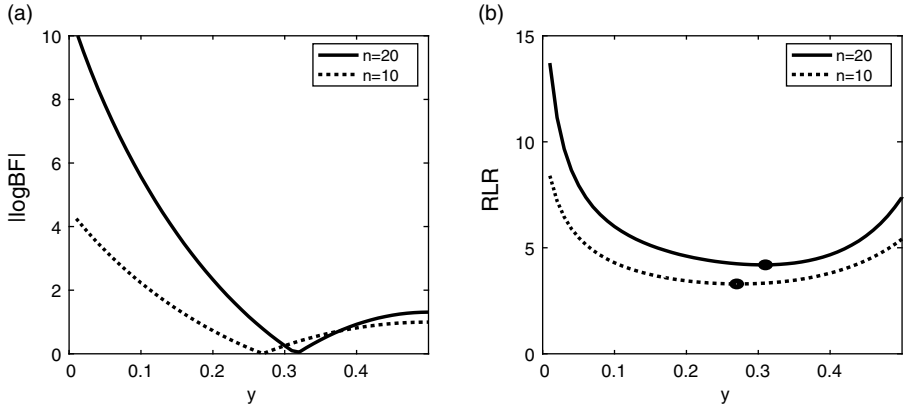

Figure 1. Transition point (TrP) of |log BF| and RLR.
Illustration for the coin-tossing example from the text: (a) |logBF| (uniform prior), (b) RLR. TrP is the point at which |logBF| = 0 or RLR is at its minimum. Values to the left of TrP support H_1, while values to the right support H_2. In both plots, TrP moves to the right as *n* increases. In addition, RLR increases at the TrP as *n* increases.

ratio scale. This presupposes not only a meaningful definition of the unit, but also, establishment of a proper 0-point. The latter might seem the simpler task, but as we argue in the following text, even this proves to be more complicated, and more interesting, than one might have anticipated.

4. 0-points for logBF and RLR

To establish a proper 0-point for the BF, we need to be willing to adjust the scale from the outset. The range of BF is (0, ∞), but minimal evidence is represented by 1 on this scale, with values <1 representing evidence in favor of H_2 and values >1 indicating evidence for H_1. A ratio scale demands, however, that the minimal amount of the object of measurement be assigned a value of 0. We could therefore opt for logBF because log(1) = 0. The range of logBF is (−∞, ∞), which in turn suggests a signed ratio scale, with sign indicating which hypothesis is supported (assuming that is, that we could establish a meaningful unit of measurement for logBF in the first place, which is by no means guaranteed). Equivalently, we can work with |logBF|.

|LogBF| = 0 demarcates the boundary, or what Vieland (2017) called the *transition point* (TrP), between putative evidence for one hypothesis versus the other, with larger departures from logBF = 0 taken to indicate increasing evidence (Figure 1a). The existence of a TrP is a feature of any ES that permits evidence to accumulate in favor of either hypothesis. Note that as *n* increases, the value of $y = x/n$ (the proportion of tosses landing heads) at which the TrP occurs shifts toward $y = ½$.[5] Thus for |logBF|, $y_{(TrP)}$ changes as a function of *n*. But the value of |logBF| at $y_{(TrP)}$ is always = 0. This suggests that the range of |logBF| is [0, ∞). We return to this in moment.

[5] This corresponds to the familiar phenomenon of smaller effect sizes becoming "significant" as *n* increases.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

The range of RLR is (0, ∞). However, RLR = 0 does not correspond to the TrP. Rather, the TrP corresponds to the minimum RLR as a function of y for given n, which is strictly > 0 (Figure 1b). This feature was highlighted in (Vieland 2017, figure 2b). As with |logBF|, the RLR TrP shifts with increasing n. But what was not discussed in Vieland (2017) is that additionally, *the value of RLR at* $y_{(TrP)}$ *increases*. This is in stark contrast to the behavior of |logBF|.

Which if either type of behavior is correct? On the face of it, the |logBF| behavior is more readily understood. The TrP is the value of y at which the data are equally compatible with both hypotheses. Surely then no matter how much data we have, at the TrP we still have 0 evidence favoring one hypothesis or the other. How can we make sense of the idea that the evidence strength increases while the data remain entirely impartial?

On the other hand, 0 appears to play two different roles on this scale. As $n \to 0$, $|\text{logBF}| \to 0$ regardless of y. But we also have $|\text{logBF}|_{(y=\text{TrP})} = 0$, so another way to get to 0 is by letting $y \to$ TrP. 0 is both the infimum as $n \to 0$; and also, 0 is the actual value of |logBF| when $y =$ TrP. This in turn means that the range of |logBF| is {{0},(0, ∞)}, which is odd if not necessarily illegitimate.

Two ways to get to the same 0-point is not in itself a problem. For instance, there are multiple recipes for bringing a physical system toward 0°K, for example, by letting entropy S→0 or pressure P→0. But they all lead to the same state (viz., S ≈ 0 and P ≈ 0). On the face of it, the situation with |logBF| is not like this. We can approach $|\text{logBF}| = 0$ by letting $n \to 0$ or by letting $y \to$ TrP, but in the latter case n can be as large as we like while $|\text{logBF}| = 0$, and in the former case |logBF| can only approach but never achieve 0. These appear to be two *different* states.

And there seems to be a conceptual difference too. $n \to 0$ can be described as allowing the quantity of relevant information to →0. But $y =$ TrP does seem to convey information regarding the hypotheses. In our coin-tossing example, where we presume a single, underlying true value of θ, there is no physical explanation for a coin that always lands such that $y =$ TrP (even setting aside the matter of stochastic variability), precisely because the TrP changes with n. For this very reason, however, we can see that $y =$ TrP conveys some kind of information about the coin and/or the model, information that becomes more troublesome as n increases, in the sense of perhaps increasingly calling the whole model into question. Whether this type of information should play a role in evidence regarding the hypotheses is debatable.[6] But again, it would appear that $|logBF| = 0$ is being used to mean two different things: 0 relevant information, and, nonzero relevant information that is equivocal between the hypotheses.[7] In the absence of a general theory of evidence, it is not possible to say definitively that the underlying states corresponding to $|logBF| = 0$ are in fact not equivalent; but neither is it possible to assert that they are.

[6] There is a rich philosophical literature on related topics, see, e.g., Fitelson (2002), Crupi et al. (2007), and McCain and Poston (2014).

[7] It is also the case that logSLR($y =$ TrP) = 0 for all n. We have suspected for some time that the SLR fails to provide a calibrated measure of evidence strength, even while it permits us to determine which is the better supported hypothesis (Hacking 1972). What is new here is the idea that part of the problem might involve something fundamentally amiss with the 0-point.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

Playing devil's advocate for the moment, suppose we wanted to eradicate this duality of the 0-point from our evidence measurement scale. How might we go about this? The only solution we can see would be to allow the evidence at the TrP to increase with increasing n. This would achieve a scale with 0 as its infimum, representing the situation in which evidence $\to 0$ when and only when the amount of relevant information $\to 0$. Interestingly, *this is precisely the behavior exhibited by RLR*.[8] It is at the very least interesting to note that RLR, which is arguably the better empirical measure in other regards (Vieland 2017), does not force us into the conundrum posed by the 0-point of |logBF|.

The question of the 0-point for either ES appears to be nontrivial. It seems that a simple appeal to intuition, or an *a priori* decision regarding treatment of the 0-point, is insufficient. Apparently to resolve the matter we need a *theory* of statistical evidence, in the context of which the precise relationships among data, hypotheses, information, and evidence can be articulated. Insofar as there exists a lower bound on strength of evidence, and surely there must be one, it begins to look like that lower bound may turn out to be considerably more interesting than we had anticipated. There is an instructive precedent for this, with which we will close our argument.

5. Absolute scales

The designation *absolute* arises almost exclusively in connection with Kelvin's temperature scale, which we may therefore take as paradigmatic of this scale type.[9] Often the Kelvin scale is said to be absolute simply because it counts up from 0. But counting up from 0 is a feature of all ratio scales. Is there a sense in which 0°K is "absolute," while, for example, 0 length (say, in centimeters) is not?

One way in which 0°K seems different from 0 length is that the lower bound for temperature is subject to empirical determination.[10] For instance, Amontons inferred a lower bound by extrapolating from experimental data to find the temperature in the limit as pressure went to 0, and experiments aimed at achieving ever lower temperatures are ongoing to this day. Indeed the mere existence of a lower bound on temperature was far from obvious a priori: "*Hot* and *cold*, like *fast* and *slow*, are mere relative terms; and, as there is no relation or proportion between motion and a state of rest, so there can be no relation between any degree of heat and absolute cold, or a total privation of heat; hence it is evident that all attempts to determine the place of *absolute cold*, on the scale of a thermometer, must be nugatory" (Rumford 1804, quoted by Chang 2004, 172). By contrast, neither the existence of a lower bound for length, nor the question of which length ought to be assigned a value of 0, requires any investigation, or even any real thought.

[8] This behavior is also reminiscent of Keynes's (1921) discussion of the "weight of argument," as something that *always* increases with additional relevant data, independently of whether the augmented data increase or decrease support for the argument.

[9] In measurement theory texts the probability scale is sometimes said to be absolute in the sense that no transformations are allowable (Houle et al. 2011), but this is a different use of the term; multiplication by a positive constant is an allowable transformation for the Kelvin scale. Also, P is on the probability scale, but this by no means guarantees that it is even on an ordinal scale when used as an ES.

[10] It is important to remember that the °K was articulated in the course of development of thermodynamics; the statistical mechanics view of the 0-point in terms of the complete absence of particle motion came later, and presupposed thermodynamic theory articulated in terms of the °K.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

Under Kelvin's definition, 0° corresponds to a fully efficient Carnot engine, and there is nothing in the theory that would prevent full efficiency from occurring, however, the laws of thermodynamics break down at extremely cold temperatures.[11] The most we can say is apparently Nernst's law, which tells us that for any reversible process, the $T = 0$ isotherm cannot be intersected by any adiabat other than the $S = 0$ isentrope (Callen 1985, 281). Never mind what that means. The point here is not to understand the physics, but only to note that if we want to resolve questions involving 0°K, then an understanding of physics is required.

In short, we could say that 0°K *is* special because it is *interesting*. The 0-point of temperature and its properties are neither obvious nor trivial to establish, but rather derive from careful study of the intended object of measurement. 0°K is *absolute* insofar as it is an infimum established by physical laws, in a way that 0 length is not. Admittedly the distinction between 0-points for mundane ratio scales like length, and absolute minima like 0°K, is a difference of degree (no pun intended) rather than kind. Issues of measurement scale always depend on the theoretical contexts in which they arise (Houle et al. 2011). But some theoretical contexts are more complex than others. The paradigmatic absolute scale is simply a run-of-the-mill interval scale for which the 0-point is *absolute* in the sense of being part and parcel of the *theory of temperature* (*thermo*dynamics) in the context of which the scale is defined. And this is ultimately the basis for our interest in the 0-points of BF and RLR. Here too, as we have suggested, it seems that we will need a *theory* of evidence before we can resolve something so seemingly simple as what constitutes a minimal amount.

It is interesting to note that Kelvin's own conception of an absolute scale did not involve a 0-point at all. In fact, the first of his two temperature scales had no lower bound. What Kelvin meant by an absolute scale was one that maintained constant meaning for the *degree* (unit) of temperature regardless of the substance being measured (Chang 2004). But for temperature, resolving the 0-point and establishing the unit went hand in hand, informed by experimental results but above all requiring the development of a novel methodological framework, which in turn changed understanding of what temperature *is*.[12] Just so, our understanding of what statistical evidence *is* may need to adapt as we work out the particulars of how we are to go about measuring it.

6. Discussion

A great deal of science relies on an activity that looks like measurement of statistical evidence. But useful measurement requires a cogent measurement theoretic foundation. Because we would like to be able to make meaningful evidential comparisons of order, difference, and ratio, what we need is a ratio scale. We have

[11] Quantum theory precludes 0°K but relating extremely cold quantum systems to thermodynamic theory is nontrivial in its own right. Note also that so-called negative temperature on the Kelvin scale is a red herring: Negative temperatures are warmer than their positive counterparts, so 0 remains the lower bound.

[12] Interestingly, Kelvin's second scale—the one now accepted as correct—turned out to be a linear transformation of the Celsius scale, implying that the °C "means the same" amount of temperature across the range of the scale. Therefore, his first proposal, which was a nonlinear transformation of Celsius, has to have been a rubber scale.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

argued here that a proper ratio scale for statistical evidence measurement will need to be *absolute*, in the sense that determination of its 0-point apparently requires a better theory of evidence than what statisticians have relied on to date.

But of course a meaningful 0-point alone is not sufficient for proper measurement. An absolute scale is, at the end of the day, simply an interval scale with a lower bound of 0 that is interesting in some way, and the hallmark of an interval scale is that the unit "means the same" across the range of the scale and across contexts of application. This returns us to the concept of *absolute* in a sense closer to Kelvin's original intent.

How *does* one confirm constancy of the meaning of a unit for a theoretically constructed object of measurement? Kelvin's theory of temperature was entirely mathematical: The degree was defined in terms of ratios of heat for an ideal gas undergoing a Carnot cycle, a wholly fictional setup that could not be implemented in the laboratory. The constancy of the meaning of the unit was embedded in the mathematics, but for that very reason, unavailable to direct empirical verification. By Kelvin's day there existed good *empirical* measurement devices such as Amontons' air thermometer, which seemed likely, based on experimentation, to be measuring temperature on interval scales. Thus Kelvin was able to validate his measurement scale empirically, *to some extent*, by aligning his calculations with the readings of (apparently) interval-scaled measurement devices under carefully controlled experimental conditions approximating, though never achieving, the conditions of Carnot's cycle.

An entire book could be written, however, in explication of that casual clause "to some extent" in the previous sentence (indeed, vide Chang 2004!). Constancy of the meaning of the °K as measured by actual thermometers was confirmed using a process, in Chang's phrase, of *epistemic iteration*, which to this day leaves us short of certainty, but nevertheless with a rich and productive theoretical framework. The laws of thermodynamics take on their familiar, elegant form only when expressed as a function of temperature measured on the Kelvin scale, and this is the ultimate validation of the °K. But it remains an unassailable fact that there is no such thing as direct verification that any given measurement device is consistently measuring on the Kelvin scale, let alone doing so under all conditions of application.

Apart from access to reasonably good thermometers, Kelvin had something else working in his favor: He was among a community of scientists with a shared desire for a better understanding of temperature. By contrast, it is difficult to convince statisticians of the need for a better understanding of evidence. Perhaps this is because they view the very idea of measurement of evidence on an absolute scale to be, in Rumford's evocative word, *nugatory*. After all, how would we verify that one degree of evidence on any given measurement scale always "means the same" with respect to the evidence, without some independent way of knowing what the evidence is?

The point is well taken, but moot. Vindication of a theoretical measurement construct is not a matter of axiomatics. It happens by epistemic iteration, not in one fell swoop and never to the point of mathematical certainty. Perhaps the first step to solving the evidence measurement problem—and surely this is a problem worth solving—is understanding the limits on what demonstration of a solution would look like.

https://doi.org/10.1017/psa.2023.15 Published online by Cambridge University Press

References

Callen, Herbert B. 1985. *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. New York: John Wiley & Sons.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez. 2007. "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues." *Philosophy of Science* 74 (2):229–52.

Edwards, Anthony W. F. 1992. *Likelihood*, 2nd ed. Baltimore: Johns Hopkins University Press.

Fitelson, Branden. 2002. "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity." *Philosophy of Science* 66:S362–S78.

Hacking, Ian. 1965. *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hacking, Ian. 1972. "Review of Edwards' Likelihood." *British Journal for the Philosophy of Science* 23:132–37.

Hand, David J. 2004. *Measurement: Theory and Practice*. Oxford: Oxford University Press.

Houle, David, Cristophe Pelabon, Günter P. Wagner, and Thomas F. Hansen. 2011. "Measurement and Meaning in Biology." *The Quarterly Review of Biology* 86 (1):3–34.

Kass Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430):773–95.

Keynes, John Maynard. 1921. *A Treatise on Probability*. London: Macmillan and Co.

McCain, Kevin, and Ted Poston. 2014. "Why Explanatoriness Is Evidentially Relevant." *Thought: A Journal of Philosophy* 3 (2):145–53.

Vieland, Veronica J. 2017. "Measurement of Statistical Evidence: Picking Up Where Hacking and Others Left Off." *Philosophy of Science* 84 (5):853–65.

Cite this article: Vieland, Veronica J. and Sang-Cheol Seok. 2023. "Absolutely Zero Evidence." *Philosophy of Science* 90 (5):1173–1182. https://doi.org/10.1017/psa.2023.15