

COMMENTS AND CRITICISM
MEASURING CONFIRMATION AND EVIDENCE*

Ellery Eells and Branden Fitelson

University of Wisconsin-Madison

October 19, 2000

Bayesian epistemology suggests various ways of measuring the support that a piece of evidence provides a hypothesis. Such measures are defined in terms of a subjective probability assignment, pr , over propositions entertained by an agent. The most standard measure (where “ H ” stands for “hypothesis” and “ E ” stands for “evidence”) is:

the difference measure: $d(H,E) = pr(H/E) - pr(H)$.⁰

This may be called a “positive (probabilistic) relevance measure” of confirmation, since, according to it, a piece of evidence E *qualitatively* confirms a hypothesis H if and only if $pr(H/E) > pr(H)$, where qualitative disconfirmation is characterized by replacing “ $>$ ” with “ $<$ ” and confirmational irrelevance is characterized by replacing the “ $>$ ” with “ $=$ ”. Other more or less standard positive relevance measures that have been proposed are:

the log-ratio measure: $r(H,E) = \log[pr(H/E)/pr(H)]$

and

the log-likelihood-ratio measure: $l(H,E) = \log[pr(E/H)/pr(E/\sim H)]$.¹

*We thank Marty Barrett, Mike Byrd, Malcolm Forster, Dan Hausman, Ilkka Kieseppa, and Elliott Sober for useful comments.

⁰While a one-place subjective probability assignment, $pr(-)$, is a measure of an agent’s degrees of belief in propositions, $pr(-/-)$ is a two-place *conditional* probability function. $pr(H/E)$, the probability of H conditional on E , is (usually) defined as $pr(H\&E)/pr(E)$. Roughly, according to Bayesian epistemology, the probability of H conditional on E is supposed to correspond the degree of belief the agent would have in H were the agent to learn E .

¹The latter two measures take the *log* of the ratios simply to normalize, so that, like the difference measure, positive and negative confirmation correspond to measures > 0 and < 0 , respectively, and confirmational irrelevance corresponds to measure $= 0$. Advocates of r include Peter Milne, “ $\log[p(h/eb)/p(h/b)]$ is the One True Measure of Confirmation,” *Philosophy of Science*, LXIII, 1 (1996): 21-26; and George N. Schlesinger, “Measuring Degrees of Confirmation,” *Analysis*, LV, 3 (1995): 208-12. Advocates of l include I. J. Good, “Explicativity, Corroboration, and the Relative Odds of Hypotheses,” in his *Good Thinking: The Foundations of Probability and Its Applications* (Minneapolis: Minneapolis UP, 1983); and I. J. Good, “The Best Explicatum for Weight of Evidence,” *Journal of Statistical Computation and Simulation*, XIX (1984): 294-9.

In a recent paper, David Christensen² criticizes these measures, and as a partial solution to the criticisms he raises, he suggests what he regards as an improved measure, which we call:

$$\begin{aligned} \text{the normalized difference measure: } S(H,E) &= [1/\text{pr}(\sim E)][d(H,E)] \\ &= \text{pr}(H/E) - \text{pr}(H/\sim E).^3 \end{aligned}$$

Of course, this measure also is a positive relevance measure.

Christensen's criticisms of the standard measures focus mainly on the difference measure, and he claims that the same alleged difficulties apply also to the other measures, including, ultimately, his own suggested measure, S ; his ultimate conclusion is skepticism about the possibility of defining an adequate probabilistic measure of confirmation. The objections he raises to these measures involve mainly the so-called "problem of old evidence";⁴ the claim is that the standard measures do not fare well in certain situations he describes involving "old evidence," while S does better in these situations but is itself not ultimately satisfactory. In section I of this paper, we argue that the standard measures (in particular the difference measure) can accommodate the examples Christensen describes just fine – that he has not fully appreciated the versatility of the relevant part of a kind of Bayesian resolution (that he discusses) of the problem. In section II, we argue that Christensen's allegedly improved measure S suffers from special difficulties of its own, besides also being vulnerable to the alleged "old evidence" difficulties he describes for the standard measures, and we conclude with reflections on Christensen's skeptical conclusion.

I. Old Evidence and Support of Hypotheses

The problem of old evidence, as usually formulated, arises in situations in which $\text{pr}(E) = 1$. In this case, $\text{pr}(H/E) = \text{pr}(H)$, and E cannot confirm H – on *any* of the *first three* measures of confirmation described above. However, this can happen even when, intuitively, evidence E would seem to provide significant support for hypothesis H : when E was learned should not affect

²"Measuring Confirmation," This JOURNAL, XCVI, 9 (September 1999): 437-61.

³This gives two formulations of measure S , which Christensen proves are equivalent to each other (provided that $\text{pr}(E) \neq 1$).

⁴This problem for Bayesian confirmation theory, to be discussed below, was, famously, raised by Clark Glymour in his *Theory and Evidence* (Princeton: Princeton UP, 1980) and was noticed as an issue for probabilistic theories of confirmation as early as 1968: see I. J. Good, "Corroboration, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor," *The British Journal for the Philosophy of Science*, XIX (1968): 123-43; also his "A Historical Comment Concerning Novel Confirmation," *The British Journal for the Philosophy of Science*, XXXVI, 2 (June 1985): 184-6.

its evidential relevance for H .

Christensen says that the problem of old evidence, or at least a version of it, can arise even when $\text{pr}(E)$ is not equal to 1, but just very close to 1; this is significant since many Bayesians feel uncomfortable about the idea that a rational person should ever assign probability 1 to a nontautology anyway. Christensen claims that “it remains true that as $\text{pr}(E)$ approaches 1, the degree to which E can confirm anything becomes [on measures d and r] vanishingly small” (p. 439).⁵ Christensen emphasizes the version of the problem of old evidence in which $\text{pr}(E)$ is not equal to 1. And he points out that in this case $S(H,E)$ needn’t vanish as $\text{pr}(E)$ approaches 1.⁶ However, since the “old evidence” criticisms he advances against the standard measures of confirmation do not depend on whether $\text{pr}(E)$ equals or is only very close to 1, we will in this section, for simplicity, assume the usual version in which $\text{pr}(E) = 1$. Further, for the rest of this section, we limit ourselves to discussing the criticisms of the difference measure d . As Christensen says about his criticisms, our points also can easily be applied to the log-ratio measure r ; the log-likelihood-ratio measure l will be a topic in section II.

Even given this, there are still two problems of old evidence that Christensen distinguishes, which he calls the *diachronic* and *synchronic* problems.⁷ The diachronic problem involves an actual event of confirmation, at an earlier time at which evidence is learned and confirmation of a hypothesis actually takes place.⁸ Suppose that the diachronic problem can be solved by showing

⁵All page-only references will be to Christensen. Note that $\text{pr}(H) = \text{pr}(H/E)\text{pr}(E) + \text{pr}(H/\sim E)\text{pr}(\sim E)$, so that $\text{pr}(H)$ must approach $\text{pr}(H/E)$ as $\text{pr}(E)$ approaches 1. However, this is consistent with the possibility that, for example, each of these two values fluctuate together from very close to 0 to very close to 1 as $\text{pr}(E)$ approaches 1. But a natural and standard assumption (e.g., as in Richard Jeffrey’s probability kinematics, *The Logic of Decision* (Chicago: Chicago UP, 1983)) is that $\text{pr}(H/E)$ and $\text{pr}(H/\sim E)$ remain constant as the probabilities of the evidence propositions, E and $\sim E$, may vary. In this case, as $\text{pr}(E)$ approaches 1, $\text{pr}(H)$ will approach the fixed value $\text{pr}(H/E)$; and, of course, the degree to which E supports H vanishes to 0, on each of the first three measures of support described above.

⁶In section II, we discuss this and other comparisons between S on the one hand and d , r , and l on the other.

⁷For other ways of classifying problems of old evidence, see the references cited in Christensen.

⁸There are actually several versions of the diachronic problem; Christensen’s concern is the simple kind of problem where E can be thought of as an observation report. For discussion of two other kinds of diachronic problems, not relevant to his specific criticisms, see Christensen and the references cited therein.

that the relevant confirmation event can be modeled as follows: at the time of the confirmation event, the probability of the hypothesis H increases appropriately from its old unconditional probability to a new value equal to the old probability of H conditional on E . Given this, Christensen focuses on the *synchronic* problem of explaining, in Bayesian terms, how E can *still* be regarded as providing appropriately significant support for H , even though $\text{pr}(E)$ is now equal to 1. As he puts it,

it seems clear that two things can be true simultaneously: (1) some beliefs will provide significant evidential support for others, and (2) some of the evidential beliefs will be held with high levels of confidence. These two features are...flatly incompatible on the standard Bayesian analysis of (synchronic) evidential support. (p 443)

Christensen considers the following “historical approach” to the synchronic problem: “it seems appropriate to say that E is (*actual*) evidence for H , for a given individual, if, at some time in the past, the event of its confirming H , for that individual, took place.”⁹ Crucial here is the idea of distinguishing between the *past event of E 's confirming H* from *E 's now being part of one's current body of evidence for H* , where the choice of the terms “confirming” and “evidence” is somewhat arbitrary and meant just to mark the distinction. We have used the term “support” above in some places to be neutral between these two distinct but related ideas.

These two ideas (especially the “evidence” idea) will be elaborated further and improved upon below to deal with the following objection raised by Christensen (pp. 444-5). He describes an example in which the hypothesis (call it H) is that deer live in a nearby wood. The agent comes across (call this evidence D) a pile of deer droppings there. The probability of D becomes 1 and the probability of H increases to near 1. Subsequently he comes upon (call this evidence A) a shed deer antler. However, at this time the probability of H is already very high, so the discovery of A could not significantly increase the probability of H . Thus, on the “historical approach” described above, D provides significant support for H while A provides very little support. Intuitively, however, it would seem that D and A are equally strong evidence for H .¹⁰ And as Christensen puts it,

The historical approach thus makes contemporary evidential support depend in an

⁹From Ellery Eells, “Problems of Old Evidence,” *Pacific Philosophical Quarterly*, LXVI (1985), p. 287; we have changed here the “ T ”s in the original to “ H ”s.

¹⁰Approach of $\text{pr}(H)$ to the particular value 1 is not crucial to the problem Christensen sees here. All that is important is that two pieces of evidence seem symmetrical in their support of H while in the presence of either the other doesn't affect the probability of H much.

unintuitive way on the order in which evidence was discovered. Intuitively, synchronic support should depend on the agent's present epistemic state, not on such historical accidents. Clearly, the synchronic problem cannot be reduced to the diachronic one in this way. (p. 445)

To introduce some notation that will be useful in describing the Bayesian solution to the synchronic problem that answers Christensen's criticisms, we first describe in a more general way the solution to the diachronic problem. Let E_1, E_2, \dots, E_N be all the evidence propositions that have been learned to date that are relevant to the hypothesis H ; say these pieces of evidence came in at times t_1, t_2, \dots, t_N , respectively (so that it is now after t_N); let pr_0 be the subjective probability assignment the agent begins with, just before t_1 ; and, for $i = 1, 2, \dots, N$, let $pri(-) = pr_0(-/E_1 \& E_2 \& \dots \& E_i)$.¹¹ Now we can define "confirmation" as follows:

Definition: E_i confirms H at time t_i if and only if $pri(H) (= pri_1(H/E_i)) > pri_1(H)$; and the *degree to which* E_i confirms H at t_i is the difference between these two probabilities.

Then, in Christensen's deer example, D significantly confirmed H at the time it was learned, but A did *not* significantly confirm H at the time it was learned, when D had become part of "background knowledge." Admittedly, this registers what may be a "historical accident" of the order in which the evidence came in. But this will seem unintuitive only if we leave out the rest of the Bayesian story, which uses the concept we call "being evidence for", to which we now turn.

On the Bayesian account, hypothesis support is a three-place relation among evidence, hypothesis, and background knowledge (where the last is "summarized" in a subjective probability function which changes as evidence comes in). This shows up in the definition of "confirmation" above; and relativity to (different parts of) background knowledge is central to the definition of "being evidence for", given below. As a simple application of the general definition to follow, we turn again to Christensen's deer example. Here we may say that relative to *no relevant background knowledge*, each of D and A is significant evidence for H , in that: $pr_0(H/D) > pr_0(H)$ and $pr_0(H/A) > pr_0(H)$, where in each case the difference is "significant." The relativity to the state of no relevant background knowledge is represented in the probability comparisons here by using, on the right hand sides, the pr_0 -unconditional probability of H , and on the left hand sides that probability conditional on the evidence. Also, relative to *background knowledge* D (A), A (D) is

¹¹ pr_0 should *not* be understood as representing an agent's degrees of beliefs before she has encountered *any* evidence about *anything at all*, or a "pure" initial probability function with which an agent enters the world. Rather, we are simply backing up to the point at which the agent lacks the evidence relevant to the analysis of a particular case of hypothesis support. (Compare Christensen, p. 460, n. 34.)

something less than “significant” evidence for H , in that $\text{pr}_0(H/D \& A)$ ($\text{pr}_0(H/A \& D)$) is only slightly larger than $\text{pr}_0(H/D)$ ($\text{pr}_0(H/A)$). The relativity to background knowledge D (A) is represented here in the probability comparisons by taking both probabilities compared to be conditional on D (A). Note that here the question of the significance of evidence for hypothesis does not depend on the order in which evidence comes in.

For the sake of completeness, we now give a more general characterization of “evidence.”¹² Where again E_1, \dots, E_N includes all the evidence that is relevant to a hypothesis H , and pr_0 is as before, we can ask what the evidential relevance of some E_i is for H relative to some subset of the propositions $E_1, \dots, E_{i-1}, E_{i+1}, \dots, E_N$. Let B range over the 2^{N-1} conjunctions of the members of these 2^{N-1} subsets. Then we offer the following definition:

Definition: E_i is evidence for H , relative to B , if and only if $\text{pr}_0(H/B \& E_i) > \text{pr}_0(H/B)$; and the *degree of E_i 's evidential support for H , relative to B* , is the difference between these two probabilities.

We emphasize that B can be the conjunction of the members of *any* subset of $\{E_1, \dots, E_{i-1}, E_{i+1}, \dots, E_N\}$ and not just of the set $\{E_1, \dots, E_{i-1}, E_{i+1}, \dots, E_N\}$ itself.

Making use of the idea that hypothesis support is a three-place relation, we have characterized two distinct questions that can be asked about support. One can ask whether E actually confirmed H , and this question is implicitly relative to the actual history of the relevant subjective probability assignment, in particular to the actual order in which relevant pieces of evidence came in. On the other hand, one can ask whether evidence E is (still and permanently), part of the evidence for H relative to some background knowledge B , and of course the answer will vary as we vary B . The first question is historical and depends on historical accident pertaining to the order in which evidence actually came in. But we agree with Christensen that there is an aspect of support that is more like a logical relation, being insensitive to historical accident, and permanent. We think these distinct ideas are captured by the two ideas we have defined and have called “confirmation” and “evidence”.

It is worth briefly mentioning another “problem” Christensen sees for the measure d (and all this applies equally to r), which he calls the *probable-hypothesis problem* (pp. 448-9). As he correctly points out, as $\text{pr}(H)$ approaches 1, the difference between $\text{pr}(H/E)$ and $\text{pr}(H)$ must vanish to 0. His complaint is that even though E may, intuitively, be highly significant for H , the measure d can in this case measure only the (vanishingly small) historical impact of E on H . We agree that

¹²This is an improvement over Eells (1984, *op. cit.*), pp. 286-7; see also Eells, “Bayesian Problems of Old Evidence,” in *Scientific Theories*, edited by C. Wade Savage (Minneapolis: University of Minnesota Press: 1990), pp. 208-210.

this is true, but appropriately so for our historical concept of “*confirmation*,” while it is also easy to see that, again, the approach to “*evidence*” detailed above does not put this kind of constraint on the “*evidential*” significance of E for H , as measured by d relative to various bodies of background knowledge B .

II. An Assessment of the Measure S and of Christensen’s Skeptical Conclusion

In the end, Christensen is skeptical about the possibility of *any* adequate probabilistic measure of support, and offers his measure S just as an *improvement* over the standard measures he considers. In the previous section, we argued that Christensen’s “old evidence” criticisms of measure d (and of r) do not point to an advantage of S , when d (or r) is properly applied to the old evidence issue. In this section we consider 1) his assessment of the measure l , 2) his “normalization” idea (dividing a measure by $\text{pr}(\sim E)$) in general, and 3) his reasons for his skeptical conclusion, and we conclude a) that the normalized difference measure S (and normalization in general) does not represent an improvement over any of the standard measures and b) that Christensen’s reasons for his skeptical conclusion are misplaced, in a hope for an “all purpose,” “purely synchronic” measure of support (which seems to have been a motivation for S).

Christensen (p. 440) is careful to point out that l , like S , does not violate what he takes to be two desiderata of measures c of support, discussed above, namely,

D1: $\text{pr}(E)$ ’s approaching 1 *does not imply* that $c(H,E)$ approaches 0

and

D2: $\text{pr}(H)$ ’s approaching 1 *does not imply* that $c(H,E)$ approaches 0.

These are violated by d and r . However, he seems to suggest (p. 440, and see note 8) the following as a desideratum (a sort of variation on **D1**), which is violated by l but not by S :

D3: Provided that $\text{pr}(H)$ does not (at the same time) approach 1, $\text{pr}(E)$ ’s approaching 1 *does not imply* that $c(H,E)$ approaches 0.

Apparently, it is on this basis that Christensen thinks we should favor S over l . We think this criticism of l is too hasty – at least as a defense of S as compared to the log-likelihood-ratio idea. For one thing, it seems natural for defenders of l to ask why *they shouldn’t also* be allowed to use the trick of multiplying their measure by the “normalization factor” $1/\text{pr}(\sim E)$. This would yield the following “normalized” version of the log-likelihood-ratio measure l :

$$lN(H,E) = [1/\text{pr}(\sim E)] \log[\text{pr}(E/H)/\text{pr}(E/\sim H)] = [1/\text{pr}(\sim E)][l(H,E)].$$

It is interesting to note that lN satisfies all three of Christensen’s desiderata **D1–D3**. In addition, it turns out that lN is *even more* insensitive to the kinds of changes in $\text{pr}(H)$ and $\text{pr}(E)$ that seem to worry Christensen. In particular, lN satisfies the following desideratum, but S does not:

D4: $\text{pr}(H)$'s approaching 0 while (at the same time) $\text{pr}(E)$ approaches 1 *does not imply* that $c(H,E)$ approaches 0.

For instance, if $\text{pr}(H) = 0.01$ and $\text{pr}(E) = 0.99$, then $S(H, E) \leq 0.0101$. By contrast, in such a case, there is nothing preventing $\text{LN}(H,E)$ from being quite high. It seems intuitive that in cases where $\text{pr}(H)$ is low it should be *easy* for E to provide significant support to H (even if $\text{pr}(E)$ also happens to be high). According to Christensen's measure S , however, this cannot happen. It seems to us that Christensen should really be comparing his normalized measure S with other normalized measures like LN , not with *non-normalized* measures like l . After all, Christensen doesn't give up on the difference measure entirely – he just normalizes it so that it will satisfy certain desiderata.

On the other hand, one must consider the *costs* of normalization. For example, many contemporary Bayesian resolutions of both the ravens paradox and the problem of evidential variety (or diversity) depend on the following assumption about Bayesian measures of confirmation c :¹³

D5: If $\text{pr}(H/E1) > \text{pr}(H/E2)$, then $c(H, E1) > c(H, E2)$.

It is not an exaggeration to say that most Bayesian confirmation theorists would accept **D5** as a *desideratum* for Bayesian measures of confirmation. Indeed, a wide variety of Bayesian relevance measures satisfy **D5**, including d , r , and l . Unfortunately, if we “normalize” *any* of these three measures, we end up with a measure that violates **D5**. In particular, S and LN *both* violate **D5**.¹⁴

Finally, we turn to why Christensen himself ultimately rejects S and takes his skeptical position about the possibility of an adequate probabilistic measure of support. Turning back to his deer example, Christensen asks us to suppose that (in our notation) $\text{pr}(H) = 0.5$, and $\text{pr}(D)$ and

¹³See, for instance, Paul Horwich's *Probability and Evidence* (New York: Cambridge, 1982) for a Bayesian explication of the ravens paradox which depends on **D5** (pp. 54-63) and an outline of a Bayesian explication of the confirmational value of varied data which also depends on **D5** (pp. 118-122).

¹⁴See Branden Fitelson, “The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity,” *Philosophy of Science*, LXVI, 3 (Proceedings Supplement) (1999): S362-S378, for further discussion of **D5**, including a proof that the relevance measures of Rudolf Carnap (*Logical Foundations of Probability*, 2nd ed., Chicago: University of Chicago Press, 1962), Robert Nozick (*Philosophical Explanations*, Cambridge: Harvard UP, 1981), and Halina Mortimer (*The Logic of Induction*, Englewood Cliffs, New Jersey: Prentice Hall, 1988) also violate **D5**. For further unintuitive features of the measure S , see Ellery Eells' review of James M. Joyce's *The Foundations of Causal Decision Theory*, in *The British Journal for the Philosophy of Science*, LI, 4 (2000, forthcoming), Fitelson, “A Bayesian Account of Independent Evidence with Applications,” *Philosophy of Science*, LXVIII (2001, forthcoming), and Eells and Fitelson, “Symmetries and Asymmetries in Evidential Support” (manuscript).

$pr_0(A)$ are both very low. At t_0 , both D and A would confirm H very strongly (on any measure of confirmation we have considered). Then, at t_1 , the agent discovers what are almost certainly deer droppings. So, both $pr_1(H)$ and $pr_1(D)$ are very high, and as Christensen notes, “ S -support from A becomes very low” just after t_1 . According to Christensen, “this is quite unintuitive; A seems just as good a sign of deer as it was before D became highly probable” (p. 457). For this reason, Christensen concludes (p. 459) that S gives the wrong answer in this example.

Christensen offers an explanation of why S goes wrong in this kind of example: “ S is insensitive to ... the distinction between specific evidence and background assumptions” (p. 459). However, rather than concluding from this interesting observation that choosing a particular mathematical form for a measure of hypothesis support will not by itself be enough to properly handle the problem of old evidence, Christensen ends up with the general, skeptical conclusion that, “It now seems that probabilistic accounts will ... miss our ordinary notion’s dependence on the distinction between background beliefs and specific evidence” (p. 461).

We suggest it is a mistake to think that the synchronic old evidence problem must be solved (if solvable at all) by “purely synchronic means” – i.e., merely by fiddling with the mathematical forms of measures of confirmation, without recourse to different parts of the agent’s background knowledge. The fact that such “purely synchronic” approaches to the problem of old evidence seem doomed to failure is no reason to be skeptical about *all* probabilistic approaches to the problem of old evidence. As we have suggested, when the theory of probabilistic hypothesis support is elaborated to include both a notion (and measure) of “confirmation” and a notion (and measure) of “evidence” that explicitly incorporate, in appropriately different ways, relativity of hypothesis support to parts of background knowledge, the problem of old evidence can be adequately resolved.