Tracking Confirmation

Igor Douven*†

Confirmation is a graded notion: evidence can confirm a hypothesis to a greater or lesser degree. There has been debate about how to measure degree of confirmation. Starting from the observation that we would like evidence to be a discriminating indicator of truth, we conduct computer simulations to determine how well the various known measures of confirmation predict the extent to which a given piece of evidence fulfills that role, given a hypothesis of interest. The outcomes show that some measures are markedly better indicators of truth than others.

1. Introduction. Bayesian approaches have come to dominate confirmation theory. Most central to these approaches is the concept of probability, typically interpreted subjectively, as indicating a person's credence (or degree of belief) in a given proposition. Another notion almost as central is that of evidence. We not only believe propositions to differing degrees, but credences can also differ under different suppositions that we might make or can change upon receipt of new information. In the Bayesian paradigm, the notion of evidence has a purely formal definition: E is evidence *for* H if $p(H|E) > p(H)$, is evidence *against* H if $p(H|E) < p(H)$, and is *irrelevant* to H if $p(H|E) = p(H)$, all relative to some person's credence function p.

There is a third concept, which again builds on the concept of probability, that has not received quite as much attention as probability or evidence but that may have as much theoretical significance as those two. It is what Good (1950, 1983, 1985) calls the "weight of evidence" and what nowadays is more commonly referred to as the "degree of confirmation." Good (1983, xi) relates that he came to know the notion through his work with Jeffreys

Received September 2019; revised January 2020.

*To contact the author, please write to: IHPST,13 rue du Four, 75006 Paris, France; e-mail: igor.douven@univ-paris1.fr.

†I thank the anonymous reviewers for their incisive comments, which greatly helped to improve this article. The Julia code for the simulations reported in this article can be downloaded from https://github.com/IgorDouven/Tracking.

Philosophy of Science, 88 (July 2021) pp. 398–414. 0031-8248/2021/8803-0002$10.00
Copyright 2021 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

and Turing at Bletchley Park and that in his view "it is almost as much an intelligence amplifier as the concept of probability itself." Psychologists have shown that the notion of degree of confirmation is also deeply rooted in human reasoning, possibly even more deeply than the notion of probability (Tentori, Chater, and Crupi 2016; see also Tentori, Crupi, and Russo 2013).

In the context of developing a new theory of knowledge and building on ideas from Nozick (1981), Roush (2005, chap. 5) argues that evidence should be a *discriminating indicator*—it should indicate that a hypothesis is true if it is and *only* if it is—and that degree of confirmation should measure how well evidence discriminates between a hypothesis being true and its being false. In Roush's terminology, evidence should *track* truth, and degree of confirmation should measure how well evidence succeeds in that respect. For Good, Roush, and various other authors (e.g., Eells and Fitelson 2002; Hahn, Oaksford, and Bayindir 2005; Fitelson 2006), the degree to which evidence E confirms hypothesis H is given by the ratio between the likelihoods $p(E|H)$ and $p(E|\neg H)$, or at least by a measure ordinally equivalent to it (e.g., one that takes the logarithm of the said ratio). By now, however, many measures of confirmation have been proposed in the literature, besides the aforementioned likelihood ratio measure (see, e.g., Fitelson 1999; Eells and Fitelson 2002; Crupi, Tentori, and Gonzalez 2007; Glass 2013).

Authors have been discussing which of those measures is preferable on the basis of vignettes that are supposed to elicit intuitions favoring this measure or militating against that measure. Such vignettes may, for instance, ask readers to imagine a card being drawn from a full deck and then to judge how strongly the finding that the card is a jack confirms the hypothesis that it is a face card (Eells and Fitelson 2002). Another approach has been to invoke general principles that supposedly ought to be regarded as general desiderata for measures of confirmation. For instance, Fitelson (2006) argues that *logicality* is one such desideratum, meaning that a measure should reach its maximum when the evidence entails the hypothesis and its minimum when the evidence entails the negation of the hypothesis. Typically, such general principles are also argued for or against by invoking our pretheoretical responses to them.

While this methodological approach is common practice in philosophy, and has also often proved valuable, there are cases in which the back-and-forth between authors appealing to different sets of intuitions has not led to any real progress, let alone to a resolution of a debate. In the debate about measures of confirmation, this methodology may have helped to narrow down the range of plausible candidates, but it has not been able to identify a clear winner. Even if we had strong intuitions favoring a particular measure, one wonders whether the traditional philosophical approach is most appropriate for settling the issue of which measure of confirmation to adopt. If the question is which measure does best at tracking truth, why should we involve

https://doi.org/10.1086/712792 Published online by Cambridge University Press

our intuitions to begin with? Presumably, the question is empirical: some measures will, as a matter of fact, be better at tracking truth than others. For instance, given two measures of confirmation, a high value of one measure may simply be more frequently associated with the truth of the hypothesis at issue, given the truth of the evidence at issue, than an equally high value of the other measure.

The current article approaches the issue of which measure of confirmation to prefer not quite empirically but quasi-empirically, by turning to computer simulations. The simulations to be presented, while stylized, are realistic enough to offer some insight into the real-world behavior of the various measures of confirmation, specifically, their performance as discriminating indicators. It turns out that some measures do markedly better in this respect than others. If people are sensitive to such differences, then one would expect them to estimate degree of confirmation in a way that is more in accordance with more discriminating measures. This article's findings will be seen to be in line with empirical results already reported in the literature (Crupi et al. 2007). Indeed, those results are most naturally explained by my findings. The same findings also suggest further empirical work, which could illuminate the question of the broader significance of those findings.

2. Measures of Confirmation. From a Bayesian perspective, it is reasonable to think that the degree to which evidence E confirms hypothesis H must be some function of the credences assigned to E, H, or some of their finite truth-functional combinations. One can construct an almost endless variety of such functions. In practice, only a handful of those have been given serious attention. For many years, philosophers have focused on the measures listed in table 1, which is taken from Crupi et al. (2007) but extended with the likelihood ratio measure mentioned in the introduction.[1]

Fitelson (1999) was the first to point out that what may appear an embarrassment of riches is in effect just an embarrassment. As he shows, in touting the strengths of Bayesianism, authors have often used the notion of degree of confirmation informally. When their arguments are properly formalized, however, it emerges that they implicitly rely on different measures of confirmation and that an argument that is valid on the assumption of one measure often is not on the assumption of another. While this does not demonstrate that we should come to agree on a unique measure of confirmation, it does

1. The likelihood ratio measure ℓ is ordinally equivalent to Kemeny and Oppenheim's (1952) measure κ (see, e.g., Crupi 2016), and most authors concerned with confirmation appear to hold that differences among ordinally equivalent measures are of little interest. The results in this article show that they may well be wrong in this respect, most notably, because of two ordinally equivalent measures, one may be better at measuring how well a piece of evidence tracks truth (in Roush's sense).

https://doi.org/10.1086/712792 Published online by Cambridge University Press

TABLE 1. STANDARD BAYESIAN MEASURES OF CONFIRMATION

	Reference
$\ell(H, E) = p(E\|H)/p(E\|\neg H)$	Good (1950)
$\mathscr{k}(H, E) = (p(E\|H) - p(E\|\neg H))/(p(E\|H) + p(E\|\neg H))$	Kemeny and Oppenheim (1952)
$r(H, E) = (p(H\|E)/p(H)) - 1$	Finch (1960)
$c(H, E) = p(H \wedge E) - p(H)p(E)$	Carnap (1962)
$d(H, E) = p(H\|E) - p(H)$	Carnap (1962)
$n(H, E) = p(E\|H) - p(E\|\neg H)$	Nozick (1981)
$m(H, E) = p(E\|H) - p(E)$	Mortimer (1988)
$s(H, E) = p(H\|E) - p(H\|\neg E)$	Christensen (1999)
$g(H, E) = 1 - p(\neg H\|E)/p(\neg H)$	Rips (2001)

show that we should aim to narrow the class of admissible such measures so that we can at least agree on a family of measures that share important characteristics. We can then reevaluate the various arguments for and against Bayesianism in light of the measures that make the cut and see which arguments hold water given any of those measures.

That raises the question of which characteristics measures of confirmation should have. As previously mentioned, Fitelson (2006) advocates logicality as one desideratum. This is motivated by the idea that confirmation formalizes a notion of partial entailment, generalizing that of deductive entailment (Kemeny and Oppenheim 1952; Carnap 1962). Thus conceived, $E \vDash H$ and $E \vDash \neg H$ should emerge as limiting cases on any measure of confirmation, the former constituting the case of maximally strong confirmation of H by E, the latter the case of maximally strong disconfirmation of H by E.

Eells and Fitelson (2002) consider further possible desiderata, focusing especially on symmetries, some of which they regard as desirable properties for a measure of confirmation, others as undesirable properties (see also Fitelson 2001). For instance, in Eells and Fitelson's view *hypothesis symmetry* is a desirable feature, where a measure μ satisfies this feature precisely if $\mu(H, E) = -\mu(\neg H, E)$, for all H, E, and p. By contrast, Eells and Fitelson believe it undesirable for a measure of confirmation to have the property of *commutativity symmetry*, which holds if and only if (iff) $\mu(H, E) = \mu(E, H)$, for all H, E, and p. They demonstrate that, of the popular measures, d, $\mathscr{k}$, and ℓ are the only ones to satisfy their desiderata.

Crupi et al. (2007) propose a more general desideratum, which they state in terms of a function ν that is again fueled by the idea of confirmation as a generalization of entailment. This function is defined as follows: $\nu(H, E) = a$ if $E \vDash H$, $\nu(H, E) = -a$ if $E \vDash \neg H$, and $\nu(H, E) = 0$ otherwise, given any H and E and with a some positive constant. Crupi et al.'s new desideratum then is that a measure μ should satisfy a symmetry principle σ (such as the principles mentioned in the previous paragraph) in case of confirmation/disconfirmation iff ν satisfies σ in the same case. For example, because

https://doi.org/10.1086/712792 Published online by Cambridge University Press

$E \models H$ iff $\neg H \models \neg E$, it holds that $\nu(H, E) = \nu(\neg E, \neg H) = a$, regardless of whether E confirms or disconfirms H. It hence follows from Crupi et al.'s desideratum that $\mu(H, E) = \mu(\neg E, \neg H)$ should hold for a measure μ to be acceptable.

One important result reported by Crupi et al. (2007) is that none of the measures in table 1 satisfies their desideratum. But another is that, from each of those measures, one can easily obtain a measure that does satisfy the desideratum. The trick is to "normalize" the measures via what they reduce to in the limiting cases in which $E \models H$ and $E \models \neg H$. For instance, in the limiting case of confirmation (so if $E \models H$), the measure $\mathscr{d}$ amounts to $p(\neg H)$, and in the limiting case of disconfirmation (if $E \models \neg H$), $\mathscr{d}$ amounts to $-p(H)$. These authors then define their new measure $\mathscr{z}(H, E)$ to be $\mathscr{d}(H, E)/p(\neg H)$ if $p(H|E) \geq p(H)$ and $\mathscr{d}(H, E)/p(H)$ otherwise. There is nothing essential to the reference to $\mathscr{d}$ here; as Crupi et al. show, the parallel construction for any of the other measures yields the same measure $\mathscr{z}$.

Crupi et al. (2007) go beyond making a theoretical case for $\mathscr{z}$. They also present a reanalysis of results of an urn experiment documented in Tentori et al. (2007). In this reanalysis, they compare $\mathscr{z}$ and the standard Bayesian measures of confirmation with regard to how strongly predictions based on them correlate with the judgments concerning the impact of new evidence from Tentori et al.'s participants. Tentori and coauthors had found Kemeny and Oppenheim's (1952) measure $\mathscr{k}$ to outperform the other standard measures, meaning that the degrees of confirmation predicted by that measure correlated more strongly with registered judgments of confirmational impact than did the degrees of confirmation predicted by any of the other standard measures. However, Crupi et al.'s reanalysis shows that when $\mathscr{z}$ is included in the comparison, that measure does better still than $\mathscr{k}$.

Although Crupi and coauthors offer strong theoretical and empirical reasons favoring their $\mathscr{z}$ measure, these reasons do not entirely clinch the case for that measure (nor do the authors claim otherwise). For one, there is the famous difficulty of arguing from *is* to *ought*; for all we know, people's judgments of evidential impact are subject to a systematic bias, in which case the superiority of $\mathscr{z}$ in predicting those judgments has little normative significance. For another, the new desideratum Crupi et al. propose, from which the theoretical advantage of $\mathscr{z}$ is to derive, while not implausible, is also not entirely compelling. Support for it is to come from a vignette involving a rather complicated chance setup about which it is difficult to have strong intuitions. That also explains why later authors have gone on to argue for other measures of confirmation. Most notably, Zalabardo (2009) and Glass and McCartney (2015) make different cases for the old likelihood ratio measure, by appeal to different intuitions.

As mentioned in the introduction, though, we are not forced to approach the issue of which measure of confirmation to go with by assessing the various

https://doi.org/10.1086/712792 Published online by Cambridge University Press

candidates in light of our pretheoretical convictions. For as also mentioned there, Nozick and Roush have proposed to conceive of such measures as instruments for estimating the degree to which evidence succeeds in telling the truth of a hypothesis from its falsity. And conceived that way, these measures are open to empirical scrutiny. Suppose that over a certain period we keep a record of the degrees to which the pieces of information we receive during that period confirm or disconfirm (if either) various hypotheses of interest to us, in which we compute these degrees using all measures of confirmation we deem at least somewhat plausible. At the end of that period, we may not know the truth values of all the hypotheses that were of interest, but we will probably know the truth values of a fair number of them. Insofar as we have registered those truth values, we can check whether, for instance, higher values of a given measure were more strongly associated than lower values with a hypothesis later turning out to be true. More generally, we could ask whether a high degree of confirmation/disconfirmation reliably indicates the truth/falsity of the hypothesis at issue, while less extreme degrees are more weakly associated with truth/falsity. Using standard statistical machinery, we could quantify more finely how accurately degrees of confirmation, given any measure, predict truth values. By doing this for all measures, we could then see which of them is the most discriminating indicator.

While this is straightforward on paper, carrying out this empirical research may be a real challenge in practical terms, if only because our credences are often not accessible to us with a degree of precision that allows us to differentiate among what the confirmational impact of new information is according to different measures of confirmation. To avoid this difficulty, as well as more mundane ones related to consistent bookkeeping, we turn to computer simulations.

3. Tracking Truth. The question before us is which of the measures in the foregoing discussion discriminates most reliably between a hypothesis being true and its being false. All these measures have a "neutral" point, which indicates that the evidence is irrelevant to the hypothesis; a value above that point indicates confirmation, a value below it, disconfirmation.[2] Ideally for these measures, any value above the neutral point indicates truth, and any value below that point indicates falsity, with the indication being stronger, the higher (in the case of truth) or the lower (in the case of falsity) that value is.

2. For all measures with the exception of ℓ, the neutral point is 0; some authors use the term "likelihood ratio measure" for the log of ℓ, which yields a measure that has 0 as its neutral point as well.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

Even in this ideal situation, we would not expect all measures to do equally well, qua discriminators. For even then, a high value of $\mu(H,E)$ may make it a safer bet to classify H as true if E is known than an equally high value of $\mu^*(H, E)$, for different measures μ and μ^*. And things might not be ideal in the first place: perhaps for a measure μ, values above and below the neutral point are not significantly differently associated with the truth of the hypothesis, meaning that μ would be worthless as a discriminating indicator.

To compare the measures with regard to their discriminatory power, we proceed as follows: $\Omega = \{\omega_1, \ldots, \omega_n\}$ is a set of possible worlds with a mass function π defined on it, where $\pi(\omega_i) > 0$ for all i, with $1 \leq i \leq n$ and $\Sigma_i \pi(\omega_i) = 1$. Propositions are subsets of Ω, and where $\mathbf{1}_\varphi$ is the index set of proposition $\varphi \subseteq \Omega$, the probability of φ, $p(\varphi)$, is given by $\Sigma_{i \in \mathbf{1}_\varphi} \pi(\omega_i)$. Now designate some random world in Ω as the actual one and also randomly pick two propositions from $\mathcal{P}(\Omega)$, which we may designate as H and E, making sure only that both are nontrivial (i.e., both are nonempty proper subsets of Ω) and that whichever world we picked as the actual one is a member of E. As the notation suggests, the idea is to interpret H as some hypothesis of interest and E as a potential piece of evidence, where the requirement that the actual world is in E ensures that the evidence is veridical.

Once these choices have been made, we have all the probabilities we need to calculate, for any of our measures, the evidential impact that E has on H, according to the given measure. Furthermore, we can check whether H is true or not: it was picked randomly without any requirement to include the actual world, but of course it may well do so. We can repeat the following steps as often as we like: (i) randomly pick a mass function π on Ω; (ii) pick an actual world, and pick two propositions, subject to the aforementioned requirements; (iii) calculate for each measure the degree to which the proposition that plays the evidence role confirms the proposition that plays the hypothesis role, relative to π; and (iv) register whether the latter is true. The repetitions will output a vector of true/false labels and, for each measure, an associated vector of degrees of support.

What we would like to know, then, is how well the degrees help us to discriminate the labels, that is, how accurately those degrees predict whether the situation was one in which the hypothesis was true or one in which it was false. Given a binary dependent variable (in our case, true/false) and a continuous independent variable (in our case, degree of confirmation), the usual way of determining how helpful the independent variable is for predicting the dependent variable uses the technique of logistic regression. If τ designates the probability that a hypothesis is true, running a logistic regression on data of the kind at issue yields an equation of the form

$$\log \frac{\tau}{1 - \tau} = \beta_0 + \beta_1 x,$$

https://doi.org/10.1086/712792 Published online by Cambridge University Press

where x is degree of confirmation (according to a specific measure), and coefficients β_0 and β_1 have been estimated from the data through a maximum likelihood procedure.[3]

Our current interest is not so much in these coefficients but rather in which measure gives rise to the model that best classifies the labels, for that ipso facto answers our question of which measure is the most discriminating indicator. A standard way of evaluating the quality of a logistic regression model, qua binary classifier, is first to convert the output (log odds) to probabilities,[4] next to assume some threshold such that a probability above or equal to that threshold is interpreted as indicating truth and a probability below that threshold is interpreted as indicating falsity, and finally to compare the true positive rate with the false positive rate (i.e., the proportion of true hypotheses correctly identified as such vs. the proportion of false hypotheses incorrectly classified as true). This evaluation makes use of the so-called area under the curve (AUC), which looks at all possible classification thresholds. The AUC value is generally interpreted as a measure of how well a logistic regression model discriminates between the two levels of the dependent variable (see, e.g., Fawcett 2006).

When we run simulations along these lines, we should reckon with the possibility that the size of Ω matters to how labels relate to degrees. So we should let that size vary within some reasonable range. To model the kind of situation in which we pick two propositions at random, given that these should be nontrivial, we need Ω to have at least four members. In our simulations, we look at numbers of worlds from 5 to 100, in steps of 5. Specifically, for each of those numbers of worlds, we repeat 1,000 times the process of randomly picking an actual world, hypotheses H and E, and weight function π, with each π being drawn anew from the flat Dirichlet distribution for the given number of worlds.[5] We then calculate, for each of those situations and each measure of confirmation, the degree to which E confirms H, and

3. However, I verified that it would make no essential difference to the results to be reported if the logistic regressions in the simulations were replaced by Bayesian logistic regressions, which do not find the coefficients by maximum likelihood estimation but rather by sampling or similar methods.

4. Where o is the log odds we get from the model for some input degree of confirmation d, the probability (according to the model) that the hypothesis at issue is true given that it was confirmed by the evidence to a degree of d equals $\exp(o)/(1 + \exp(o))$.

5. The Dirichlet distribution is a multivariate probability distribution, which can be thought of as a probability distribution over probability distributions. A flat Dirichlet distribution is one whose vector of parameters consists strictly of 1's. Intuitively, such a distribution makes any distribution as likely to be selected as any other. By choosing different values for the parameters, one can, e.g., make distributions that give about equal weight to all worlds more likely to be chosen or distributions that give most weight to one of the worlds.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

we also register the truth value of H. Next, for each measure of confirmation, we run a logistic regression with registered truth values as the dependent variable and degrees of confirmation (according to the given measure) as the independent variable. And finally we calculate the AUC value of the resulting model. For each number of possible worlds we consider, we repeat these simulations 250 times, to get a sense of the reliability of our results.

Figure 1*a* shows—for each measure and each number of possible worlds we considered—the AUC score, averaged over the 250 simulations for the given combination. It is immediately evident from that figure that Crupi and colleagues' z measure by far outperforms the competition, at least generally speaking.[6] Even where its performance appears close to Kemeny and Oppenheim's k—namely, for 5 worlds—a two-sided t-test conducted on the results of the 250 simulations for that measure and z shows that the latter does reliably better, at a level of $\alpha < .0001$; t-tests for the other numbers of worlds came out significant at the same α level.[7]

In the above simulations, each $\omega \in \Omega$ had the same probability $1/|\Omega|$ of being actual. A variant of these simulations well worth considering lets the probability with which a world ω is picked to be actual in a given simulation depend on the probability mass that is assigned to ω at the start of that simulation. The variant makes particularly good sense if we think of π as determining the objective probability with which the various worlds in Ω are actual and think of our credences as tracking those objective probabilities.[8]

Keeping everything else the same, we obtain the results shown in figure 1*b*. While the picture is a bit more complicated now, it is still the case, from 25 worlds onward, that z outperforms the other measures (at $\alpha < .0001$); for 5, 10, and 15 worlds, k does slightly although significantly ($\alpha < .0001$) better; and for 5 worlds, r does a little better as well, and also significantly so ($\alpha < .0001$).[9]

Depending on which of the two sets of simulations we look at, z is either always or for the most part a significantly more discriminating indicator than

6. Crupi et al. (2007) in effect allow for monotone transformations of z, but those leave AUC values unaffected.

7. In parallel, we conducted Bayesian t-tests (assuming the Jeffreys-Zellner-Siow prior, as specified in Rouder et al. [2009]), which yield a so-called Bayes factor indicating the extent to which the evidence afforded by the data favors or disfavors the null hypothesis. In these tests, all Bayes factors came out smaller than .0001, meaning that in all cases the data were far less likely to occur under the null hypothesis than under the alternative hypothesis.

8. Thanks to an anonymous reviewer for suggesting this variant.

9. Bayes factors obtained from parallel Bayesian t-tests also indicate that, in each case, the data were much more likely under the alternative hypothesis than under the null hypothesis.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

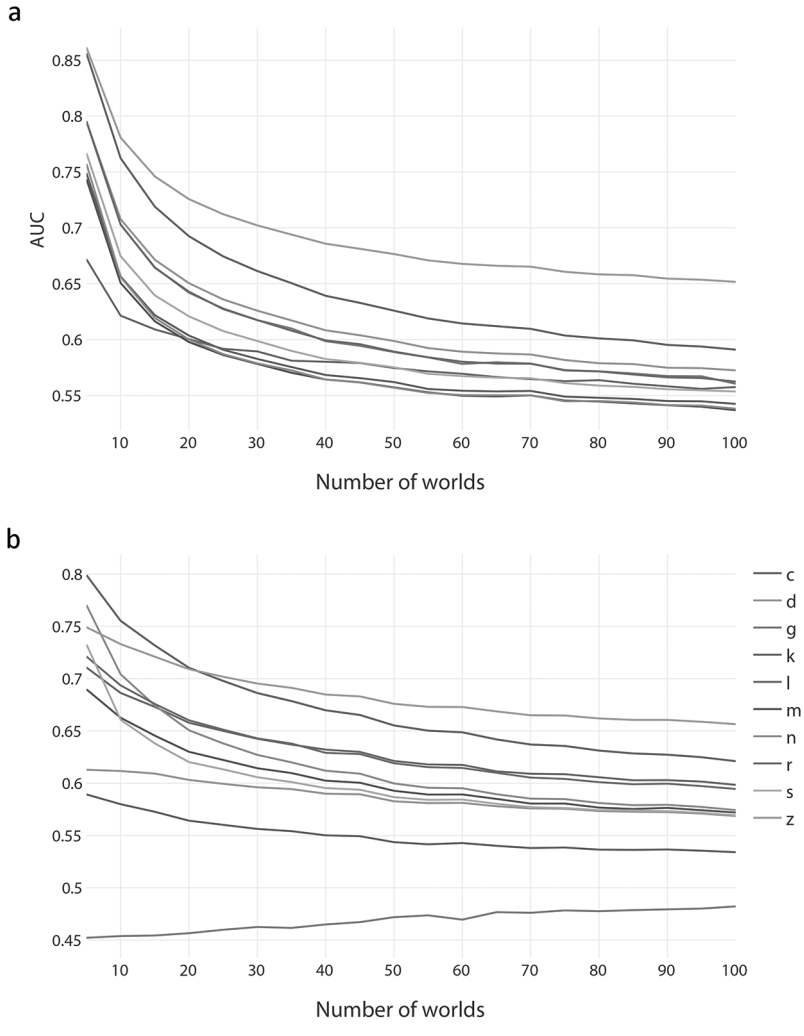


Figure 1. Mean AUC values of the various measures obtained in simulations for possible worlds from 5 to 100, increasing in steps of 5. *a*, Simulations in which each world had the same chance of being selected as the actual world; *b*, simulations in which each world's chance of being selected as the actual world was equal to the probability mass attributed to it in the simulation. Color version available as an online enhancement.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

the other measures. While that is an important finding in itself, it also helps to explain what had remained something of a mystery in Crupi et al. (2007). For, why does $\mathfrak{z}$ predict the judgments gathered for the experiment from Tentori et al. (2007) more accurately than the other measures? It is difficult to imagine that ordinary people are sensitive to the feature that $\mathfrak{z}$ respects certain supposedly desirable symmetries but violates certain supposedly undesirable ones. None of those symmetries strikes one intuitively as being either so compelling or so misguided that that would be a plausible assumption. Yet, if the notion of evidential impact is psychologically salient in the way some research suggests, then it is imaginable how, through use and trial and error, people have come to implicitly prefer to measure that impact in a way that makes the result as discriminating as possible or at least as discriminating as they could figure out how to make it. In this connection, it is further noteworthy that $\mathfrak{k}$, which came out second best in the empirical part of Crupi et al. (2007), also emerges as a clear second in our simulations.

4. Can We Do Better? The methodology applied in the previous section offers a different approach to the question of which measure of confirmation is preferable, an approach that does away with previous reliance on case-based intuitions and that gives full weight to the practical significance of measuring evidential impact. As mentioned, only a handful of all possible functions that we can build from probabilities assigned to a hypothesis H, evidence E, and their finite truth-functional combinations have been probed as measures of confirmation. Among those never seriously considered, there might well be hiding some with greater discriminatory power than $\mathfrak{z}$. Given the sheer number of candidates, however, systematically searching the solution space seems out of reach here. In this kind of case, looking at variants of solutions already known to do well is often a good idea.

Fitelson and Hitchcock (2011) point at the closeness between the discussion concerning measures of confirmation and that concerning measures of causal strength. For instance, Lewis (1986) proposes what is in effect $\mathfrak{l}$ as a measure of causal strength. And the measure $\mathfrak{d}$, which Christensen (1999) proposes as a measure of confirmation, has also been doing duty as a measure of causal strength in the literature on causality, where it mostly goes by the name of "Δp rule" (Shanks 1995). In her influential paper, Cheng (1997) argues that this rule is inferior to what she dubs the "power PC [probabilistic contrast] account."

To my knowledge, the power PC account has not yet been considered as possibly contributing to confirmation theory, but it might be worth doing so, given that, formally, this measure is a relatively minor variant of Crupi et al.'s $\mathfrak{z}$ measure. Using the measure $\mathfrak{s}$ where Cheng uses Δp, her measure of the *generative power* of E for H is defined to be

https://doi.org/10.1086/712792 Published online by Cambridge University Press

$$\mathcal{k}(H, E) \coloneqq \begin{cases} \dfrac{\mathcal{s}(H, E)}{p(\neg H|\neg E)} & \text{if } \mathcal{s}(H, E) \geq 0 \\ \dfrac{\mathcal{s}(H, E)}{p(H|\neg E)} & \text{otherwise.} \end{cases}$$

When $\mathcal{s}(H, E) < 0$, Cheng prefers to speak of the *preventive power*, or *inhibitory power*, of E for H and, therefore, takes the negative of the second clause as stated here, but that is just a semantic issue.

In section 2, we defined $\mathcal{z}$ in terms of $\mathcal{d}$ but also noted that the former can be defined just as well in terms of any of the other measures in table 1. To facilitate comparison with Cheng's measure, we restate it here in terms of Christensen's measure:[10]

$$\mathcal{z}(H, E) \coloneqq \begin{cases} \dfrac{\mathcal{s}(H, E)}{p(\neg H)/p(\neg E)} & \text{if } \mathcal{s}(H, E) \geq 0 \\ \dfrac{\mathcal{s}(H, E)}{p(H)/p(\neg E)} & \text{otherwise.} \end{cases}$$

This makes it easy to see that the only differences between $\mathcal{z}$ and Cheng's measure reside in the denominators in both clauses.

Superficially at least, the difference between the measures appears to be small. Is it of any theoretical significance? Crupi and colleagues' decision to divide by $p(\neg H)/p(\neg E)$ and $p(H)/p(\neg E)$ in the first and second clauses, respectively, is motivated by their wish to obtain a measure with range $[-1, 1]$ that reaches its maximum if $E \vDash H$ and its minimum if $E \vDash \neg H$ (see also Crupi and Tentori 2013). It takes little effort to show that $\mathcal{k}$ satisfies not only this condition but also the various constraints that follow from Crupi et al.'s (2007) main desideratum (see sec. 2 above), an overview of which is provided by table 3 in their paper.

To see that $\mathcal{k}$ has the same range as $\mathcal{z}$, it suffices to note that $\mathcal{k}$ can be rewritten as

$$\frac{p(H|E) - p(H|\neg E)}{1 - p(H|\neg E)},$$

for the case in which $\mathcal{s}(H, E) \geq 0$, and as

10. Crupi and colleagues formulate the condition for the first clause to apply as $p(H|E) \geq p(H)$, but that is equivalent to the condition given here. To see this, note that $p(H|E) \geq p(H)$ iff $p(H|E) - p(H) \geq 0$. Moreover, by the law of total probability, $p(H)$ is a convex combination of $p(H|E)$ and $p(H|\neg E)$, and hence $p(H|E) \geq p(H)$ iff $p(H|E) \geq p(H|\neg E)$.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

$$-\frac{p(H|\neg E) - p(H|E)}{p(H|\neg E)},$$

for the alternative case. (Note also that $p(H|\neg E) > p(H|E)$ in the latter case.)

As for the other conditions, first consider the case in which $\mathcal{s}(H, E) \geq 0$. Then we have that

$$\mathcal{z}(H, E) = \frac{p(\neg E)}{p(\neg H)} \times \mathcal{s}(H, E)$$

and also that

$$\mathcal{ch}(H, E) = \frac{p(\neg E)}{p(\neg H \wedge \neg E)} \times \mathcal{s}(H, E),$$

or equivalently

$$\mathcal{ch}(H, E) = \frac{p(\neg E)}{p(\neg H)p(\neg E|\neg H)} \times \mathcal{s}(H, E),$$

or again

$$\mathcal{ch}(H, E) = \frac{1}{p(\neg E|\neg H)} \times \frac{p(\neg E)}{p(\neg H)} \times \mathcal{s}(H, E),$$

or finally

$$\mathcal{ch}(H, E) = \frac{1}{p(\neg E|\neg H)} \times \mathcal{z}(H, E)\,.$$

By parallel reasoning, we find that if $\mathcal{s}(H, E) < 0$, then $\mathcal{ch}(H, E)$ can be rewritten as $1/p(\neg E|H) \times \mathcal{z}(H, E)$.

From this, it follows that if $E \vDash H$, then $\mathcal{ch}(H, E) = \mathcal{z}(H, E)$, given that then $\neg H \vDash \neg E$ and so $1/p(\neg E|\neg H) = 1$; it similarly follows that $\mathcal{ch}(H, E) = \mathcal{z}(H, E)$ if $E \vDash \neg H$. It further follows that there is a piecewise-defined function f such that $\mathcal{ch}(H, E) = f(\mathcal{z}(H, E))$, where f is injective and $f(x) = -f(-x)$. Because it also holds that $\mathcal{ch}(H, E) > / = / < 0$ if $p(H|E) > / = / < p(H)$, all conditions of lemma 3 in Crupi et al. (2007, 249) are fulfilled for $\mathcal{ch}$ to enjoy all consequences of the new desideratum proposed by these authors.

The discussion above also makes it easy to see that, in general, $\mathcal{ch}(H, E) > \mathcal{z}(H, E)$ if $\mathcal{s}(H, E) \geq 0$, and $\mathcal{ch}(H, E) < \mathcal{z}(H, E)$ if $\mathcal{s}(H, E) < 0$. Might this at least be of practical significance, even if theoretically the two measures are close cousins? To find out, we apply the methodology from the previous section and rerun the same simulations, now just focusing on $\mathcal{z}$ and the new measure $\mathcal{ch}$ in the simulations in which all worlds have the same chance of

https://doi.org/10.1086/712792 Published online by Cambridge University Press

being actual and on $\mathcal{Z}$, $\mathcal{CH}$, and $\mathcal{K}$ in the simulations in which a world's probability mass determined its chance of being actual. (Here $\mathcal{K}$ is taken into account as well because it sometimes performed best in the corresponding simulations above.) The results of these simulations are shown in figure 2. We see that, in the first set of simulations, the new measure improves at least a little over $\mathcal{Z}$. The *t*-tests on the AUC values for the measures show that while the differences between the means are not significant for 5 and 10 worlds, they are from there on: from 15 to 65 worlds at an α level of .05, and above 65 worlds at an α level of .005.[11] In the second set of simulations, the new measure improves a lot over $\mathcal{Z}$, with all *t*-tests being significant at $\alpha < .0001$. The new measure also does consistently better than $\mathcal{K}$, and even for 5 and 10 worlds, where the two are relatively close, the difference is significant at $\alpha = .0001$.[12]

Admittedly, these are just the outcomes of simulations, and it is another question whether the difference between the measures is big enough to matter in people's reasoning. One way to check would be by reusing the data set that Crupi and colleagues used to compare the measures discussed in their paper. We might find that $\mathcal{CH}$ predicts human judgments of evidential impact more accurately than $\mathcal{Z}$. Better still, one could run new experiments more specifically designed to differentiate between $\mathcal{Z}$ and $\mathcal{CH}$. I am, in fact, hoping that the results reported here will inspire psychologists to undertake this work. For now, I hope to have shown that the new methodology may be fruitfully deployed in a quest for measures of confirmation that are possibly more discriminating than the ones we already have.

5. Conclusion. We started from Roush's observations that we want our evidence to reliably discriminate between truth and falsity and that a measure of confirmation should quantify how successful a piece of evidence is in this regard. These observations inspired an approach to assessing measures of confirmation that circumvents questions of how intuitively compelling this or that condition is as a desideratum for such measures. Applying the approach to the most prominent measures of confirmation led to the finding that Crupi et al.'s (2007) measure $\mathcal{Z}$, which already enjoyed considerable backing—theoretical as well as empirical—also outstrips the competition in power to discriminate between truth and falsity. As we saw, however, there is no reason why we should leave matters there. We already discovered

11. Here, Bayes factors do not quite tell the same story: from 65 worlds onward, they indicate moderate to strong support for the alternative hypothesis, but for lower numbers of worlds, they indicate only weak support for the alternative or even slight support for the null hypothesis.

12. These results are in line with the outcomes from parallel Bayesian *t*-tests, which all yielded a Bayes factor smaller than .0001.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

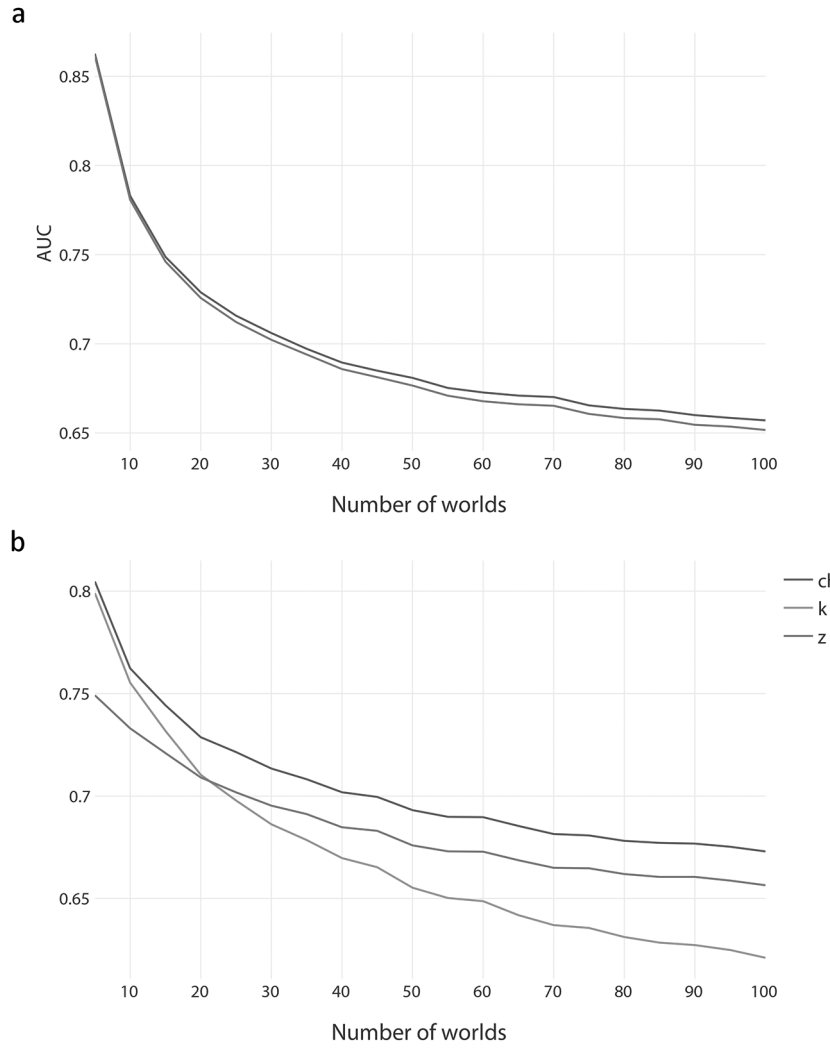


Figure 2. Mean AUC values for *ch* and *z* (*a*) and *ch*, *k*, and *z* (*b*) obtained in simulations for possible worlds from 5 to 100, increasing in steps of 5. *a*, Simulations in which all worlds had an equal chance of being selected as the actual world; *b*, simulations in which each world's chance of being selected as the actual world equaled its probability mass in the simulation. Color version available as an online enhancement.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

one measure that does at least somewhat better than the one proposed by Crupi and colleagues. Others may be out there, possibly some that do quite a bit better. The methodology put forward in this article is easy to implement and thereby facilitates the identification of such measures.

REFERENCES

Carnap, Rudolf. 1962. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Cheng, Patricia. 1997. "From Covariation to Causation: A Causal Power Theory." *Psychological Review* 104:367–405.

Christensen, David. 1999. "Measuring Confirmation." *Journal of Philosophy* 96:437–61.

Crupi, Vincenzo. 2016. "Confirmation." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. https://plato.stanford.edu/archives/win2016/entries/confirmation/.

Crupi, Vincenzo, and Katya Tentori. 2013. "Confirmation as Partial Entailment: A Representation Theorem." *Journal of Applied Logic* 11:364–72.

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

Eells, Ellery, and Branden Fitelson. 2002. "Symmetries and Asymmetries in Evidential Support." *Philosophical Studies* 107:129–42.

Fawcett, Tom. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27:861–74.

Finch, Henry A. 1960. "Confirming Power of Observations Metricized for Decisions among Hypotheses." *Philosophy of Science* 27:293–307, 391–404.

Fitelson, Branden. 1999. "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity." *Philosophy of Science* 66 (Proceedings): S362–S378.

———. 2001. "Studies in Bayesian Confirmation Theory." PhD diss., University of Wisconsin–Madison.

———. 2006. "Logical Foundations of Evidential Support." *Philosophy of Science* 73:500–512.

Fitelson, Branden, and Christopher Hitchcock. 2011. "Probabilistic Measures of Causal Strength." In *Causality in the Sciences*, ed. P. McKay Illari and F. Russo, 600–627. Oxford: Oxford University Press.

Glass, David H. 2013. "Confirmation Measures of Association Rule Interestingness." *Knowledge-Based Systems* 44:65–77.

Glass, David H., and Mark McCartney. 2015. "A New Argument for the Likelihood Ratio Measure of Confirmation." *Acta Analytica* 30:59–65.

Good, Irving J. 1950. *Probability and the Weighing of Evidence*. London: Griffin.

———. 1983. *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press.

———. 1985. "Weight of Evidence: A Brief Survey." In *Bayesian Statistics*, vol. 2, ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, 249–70. Amsterdam: North-Holland.

Hahn, Ulrike, Mike Oaksford, and Hatice Bayindir. 2005. "How Convinced Should We Be by Negative Evidence?" In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, ed. B. Bara, L. Barsalou, and M. Bucciarelli, 887–92. Mahwah, NJ: Erlbaum.

Kemeny, John, and Paul Oppenheim. 1952. "Degrees of Factual Support." *Philosophy of Science* 19:307–24.

Lewis, David K. 1986. "Postscripts to 'Causation.'" In *Philosophical Papers*, vol. 2, 173–213. Oxford: Oxford University Press.

Mortimer, Halina. 1988. *The Logic of Induction*. Paramus, NJ: Prentice Hall.

Nozick, Robert. 1981. *Philosophical Explanations*. Oxford: Oxford University Press.

Rips, Lance J. 2001. "Two Kinds of Reasoning." *Psychological Science* 12:129–34.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson. 2009. "Bayesian *t* Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin and Review* 16:225–37.

Roush, Sherrilyn. 2005. *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

https://doi.org/10.1086/712792 Published online by Cambridge University Press

Shanks, David R. 1995. "Is Human Learning Rational?" *Quarterly Journal of Experimental Psychology* 48:257–79.
Tentori, Katya, Nick Chater, and Vincenzo Crupi. 2016. "Judging the Probability of Hypotheses versus the Impact of Evidence: Which Form of Inductive Inference Is More Accurate and Time-Consistent?" *Cognitive Science* 40:758–78.
Tentori, Katya, Vincenzo Crupi, Nicolao Bonini, and Daniel Osherson. 2007. "Comparison of Confirmation Measures." *Cognition* 103:107–19.
Tentori, Katya, Vincenzo Crupi, and Selena Russo. 2013. "On the Determinants of the Conjunction Fallacy: Confirmation versus Probability." *Journal of Experimental Psychology: General* 142:235–55.
Zalabardo, José. 2009. "An Argument for the Likelihood-Ratio Measure of Confirmation." *Analysis* 69:630–35.

https://doi.org/10.1086/712792 Published online by Cambridge University Press