From Evidential Support to a Measure of Corroboration

Jan Sprenger*

June 14, 2014

Contents

*Contact information: Tilburg Center for Logic, General Ethics and Philosophy of Science (TiLPS), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: j.sprenger@uvt.nl. Webpage: www.laeuferpaar.de

Abstract

According to influential accounts of scientific method, e.g., critical rationalism, scientific knowledge grows by repeatedly testing our best hypotheses. In comparison to rivaling accounts of scientific reasoning such as Bayesianism, these accounts are closer to crucial aspects of scientific practice. But despite the preeminence of hypothesis tests in statistical inference, their philosophical foundations are shaky. In particular, the interpretation of "insignificant results"—outcomes where the tested hypothesis has survived the test—poses a major epistemic challenge that is not sufficiently addressed by the standard methodology for conducting such tests.

In this paper, I argue that a quantitative explication of *degree of corroboration* can fill this important methodological and epistemological gap. First, I argue that this concept is distinct from the Bayesian notion of evidential support and that it plays an independent role in scientific reasoning. Second, I demonstrate that degree of corroboration cannot be suitably explicated in a probabilistic relevance framework, as proposed by Popper (1954, 1934/2002). Third, I derive two measures of corroboration that possess a large number of attractive properties, establish an insightful relation between corroboration and evidential support and are not committed to a Bayesian or a frequentist framework. In sum, the paper rethinks the foundations of inductive inference by providing a novel logic of hypothesis testing.

1 Introduction. Motivating the concept of corroboration

The idea of acquiring scientific knowledge by *testing* hypotheses and appraising how well they have stood up to the test is as old as the scientific revolution. For critical rationalists such as Karl R. Popper (1934/2002), the critical attitude that we express by repeatedly testing our best scientific theories even constitutes the basis of rational inquiry about the world. However, only in the middle of the 20th century, the design and interpretation of statistical hypothesis tests has been formalized. In effect, they have acquired a predominant role in scientific reasoning and are a crucial part of publication standards. The most frequent form of scientific inference is the *null hypothesis significance test (NHST)*: it tests a precise hypothesis h_0—the "null" or default hypothesis—against an unspecific alternative h_1. In its most simple form, the null hypothesis posits a precise value for a real-valued parameter θ ($h_0 : \theta = \theta_0$), while the alternative ($h_1 : \theta \neq \theta_0$) is a disjunction of uncountably many precise hypotheses. Such tests are useful for finding out whether there is a non-negligible difference between two different experimental conditions, e.g., a medical drug and a placebo treatment.

The outcomes of NHST are traditionally described either as the "acceptance" or the "rejection" of the null hypothesis. If the results are very unlikely under the null, it is rejected in favor of the alternative (e.g., Neyman and Pearson 1933; Fisher 1956; Gillies 1971). While a rejection is usually taken as evidence against the null hypothesis and quantified by means of a p-value or significance level, there is little methodological guidance on what the *acceptance of the null hypothesis* could mean, in a positive sense. Statistics textbooks (e.g., Chase and Brown 2000; Wasserman 2004) restrict themselves to the claim that an acceptance of the null hypothesis does not mean more than failure to reject the null, or failure to demonstrate a statistically significant phenomenon. This is remarkable for at least two reasons: statistically insignificant results can hide substantial effects (Ziliak and McCloskey 2008), and also the absence of significant results can be a scientifically interesting conclusion. As an example, consider the monitoring of a freshly admitted medical drug for harmful side effects. The producer of the drug, clinicians and the general public all have an interest in knowing to which degree the null hypothesis—that the drug has no unexpected side effects—is backed by the evidence, but the standard methodology for hypothesis testing does not specify how we should quantify such a judgment, let alone how we should do it in an objective way.

A concept that could fill this lacuna in the NHST methodology is *degree of corroboration*, famously developed by Karl R. Popper in his "Logic of Scientific Discovery" (1934/2002: ch. 10):

> By the degree of corroboration of a theory I mean a concise report evaluating the state (at a certain time *t*) of the critical discussion of a theory, with respect to the way it solves its problems; its degree of testability; the severity of tests it has undergone; and the way it has stood up to these tests. Corroboration (or degree of corroboration) is thus an evaluating *report of past performance*. Like preference, it is essentially comparative. (Popper 1979: 18, original emphasis. See also Popper 1934/2002: 248.)

Adequately explicated corroboration judgments would solve many problems: they would appraise the performance of the null hypothesis in an experiment, rather than just stating the failure to find significant results. They would indicate when the "acceptance" of the null hypothesis provides a reason to trust it. They would explain why highly corroborated hypotheses are preferred to weakly corroborated ones. More generally, explicating degree of corroboration might revive a critical rationalist epistemology of science, by showing how hypothesis tests increase scientific knowledge (e.g., Rowbottom 2011). In the light of these promises, it is notable that neither philosophers nor statisticians have found an adequate explication of degree of corroboration, and that efforts to do so have faded since the 1960s (Popper 1954; Good 1960, 1968).

The paper is structured as follows. Section 2 conceptually demarcates degree of corroboration from Bayesian explications of evidential support. Section 3 discusses, and ultimately rejects, Popper's own explication of corroboration. Section 4 advances a formal argument to the effect that a probabilistic relevance framework is not suited for explicating corroboration. Section 5 introduces the new framework for measuring corroboration and derives two measures of corroboration from a parsimonious set of plausible axioms. Finally, Section 6 explains the attractive properties of this measure and relates it to the concept of evidential support while Section 7 summarizes and concludes. While my own definition of degree of corroboration in Sections 5-6 is definitely inspired by Popper, the approach of the paper is systematic, not exegetical, and the proposed explication will in some ways deviate from Popper's own take on scientific reasoning.

2 Evidential support versus corroboration

The point of measuring corroboration is to quantify the extent to which a hypothesis has stood up to an attempt to refute it. Thus, degree of corroboration gives an evaluating—and supposedly objective—report of past performance. For the case of a hypothesis that makes deterministic predictions, corroborating evidence is intuitively defined as evidence that conforms to the predictions of the tested hypothesis. The more specific it is, the more it corroborates the hypothesis.

This rationale essentially corresponds to the hypothetico-deductive model of theory confirmation (Gemes 1998): logical consequences of a theory confirm it. While this model may be adequate as a *qualitative* theory of corroboration, it is not applicable to NHSTs which deal with statistical predictions of a hypothesis. Here, a different, quantitative model has to be developed (see also Popper 1934/2002: 265–266).

However, it is not evident that we need corroboration judgments for explicating this aspect of NHSTs. There is already a concept that describes how the epistemic status of a hypothesis is raised by observations: *evidential support*. Standardly, evidential support is explicated in Bayesian terms, that is, in terms of degrees of belief: evidence e supports hypothesis h if and only if $p(h|e) > p(h)$, that is, if e increases the agent's subjective degree of belief in h (e.g., Fitelson 2001). Why do we need another, closely related concept?

This skepticism is expressed in the **Monism Thesis**: the concept of corroboration can be reduced to the (Bayesian) concept of evidential support.

This thesis sounds especially attractive in the light of the shaky epistemic foundations of NHST and their frequent misuse (e.g., Cohen 1994; Fidler 2013): perhaps we should abandon the entire business of (frequentist) hypothesis testing, perform a Bayesian analysis based on the interpretation of probability as subjective degree of belief and replace a judgment of corroboration by a judgment of evidential support. For Bayesians such as Howson and Urbach (2006), this could be the preferred option.

I shall now present four objections to the Monism thesis. This does not rule out that a proper explication of corroboration can also be interpreted as a measure of evidential support, or vice versa: rather, the point is to show that the two *concepts* are not redundant and need different explication strategies.

> Objection 1: Inference to the true hypothesis is the target notion of evidential support, but not necessarily of corroboration.

Scientific hypotheses and models are idealizations of the external world that are judged by their ability to capture relevant causal relations and to predict future events, rather than literally true descriptions of the external world (see the survey of Frigg and Hartmann 2006). In other words, the epistemic function of corroboration consists in determining whether the data are consistent with the tested hypothesis, or whether the results agree "well enough" with the null hypothesis h_0 that we may use it as a proxy for a more general statistical model. In other words, the "acceptance" of h_0 does not imply that it should be regarded as true or empirically adequate, but that it is a useful and tractable idealization of a more general statistical model (Bernardo 2012; Gelman and Shalizi 2013). That is, corroboration is a guide to practical preference over competing hypothesis, but not as a guide to truth (Popper 1934/2002: 281–282). Evidential support, on the other hand, is traditionally defined as the degree to which our confidence in the *truth* of a hypothesis is raised. Convergence theorems show how inference to the best-supported hypothesis guides us to the true hypothesis (e.g., Gaifman and Snir 1982; see Brössel 2014 for a similar result regarding the systematic power of a theory). Unlike corroboration, which is defined as an evaluating report on past data, evidential support is supposed to justify inductive inference.

> Objection 2: (Change in) Degree of belief is a central concept for evidential support, but not for corroboration.

Evidential support is based on comparing past and present degrees of belief. This can be traced easily in the qualitative definition of evidential support (e supports h if and only if $p(h|e) > p(h)$), but also in popular support measures such $d(h,e) = p(h|e) - p(h)$ and $r(h,e) = p(h|e)/p(h)$. More generally, Crupi, Chater and Tentori (2013) have argued that all measures of evidential support $\mathfrak{c}(h,e)$ should possess the "final probability incrementality" property

$$\mathfrak{c}(h,e) \;>/=/<\; \mathfrak{c}(h,e') \qquad \text{if and only if} \qquad p(h|e) \;>/=/<\; p(h|e'). \tag{1}$$

This condition demands that e supports h more than e' if and only if e raises the probability of h to a higher level than e' does. This condition makes sense for a concept of evidential support that is specified as a generalization of strict deductive entailment, or as the degree to which e raises the agent's degree of belief in h (Eells and Fitelson 2002; Crupi, Tentori and González 2007). However, it is much less obvious for degree of corroboration: a corroboration judgment seems, at least in principle, to make sense even if we do not have subjective degrees of belief in the tested hypothesis or refuse to elicit them. It is about past performance, not about epistemic or psychological attitude. In a nutshell, rather than a (subjective) measure of belief change, corroboration ought to be an (objective) measure of past performance.

> Objection 3: On a Bayesian account, hypotheses with prior probability $p(h) = 0$ cannot be confirmed evidentially. Yet, they are perfectly acceptable candidates for being corroborated.

As a consequence of Bayes' Theorem, any hypothesis with prior probability $p(h) = 0$ also has posterior probability $p(h|e) = p(h)\,p(e|h)/p(e) = 0$. By the qualitative definition of evidential support, no such hypothesis can be evidentially supported since $p(h|e) = p(h)$. But certainly, they can be *corroborated*: after all, scientists often deal with an uncountable set of candidate hypotheses where all singleton hypotheses receive zero weight (e.g., different values of a physical parameter). Testing whether such hypotheses are good and useful idealizations of reality, or quantifying the empirical corroboration of any such hypothesis certainly makes sense. This objection is especially troubling since Bayesian models of NHSTs often assign zero weight to the null hypothesis, e.g., by assigning a continuous prior over the parameter space. Whatever the measure of evidence that the Bayesian uses for appraising the null in such tests (e.g., a density-based measure such as the Bayes factor), it cannot be a Bayesian measure of evidential support in the proper sense.

> Objection 4: Corroboration is a way more asymmetric notion than evidential support.

The logic of NHSTs is *asymmetric*: in general, a rejection of the tested hypothesis h gives rise to much stronger conclusions than an acceptance would do. A reason for this is that unlike the null, the alternative $\neg h$ is usually not a precise hypothesis, like in our introductory example of testing $\theta = \theta_0$ against $\theta \neq \theta_0$.

It is not obvious how this asymmetry can be expressed by measures of evidential support. Consider two of the most reputable ones, the log-likelihood-measure l (Kemeny and Oppenheim 1952; Fitelson 2001; Bovens and Hartmann 2003), and the Crupi-Tentori-measure z (Crupi, Tentori and Gonzalez 2007; Crupi and Tentori 2013):

$$l(h,e) = \log \frac{p(e|h)}{p(e|\neg h)} \qquad z(h,e) = \begin{cases} \frac{p(h|e)-p(h)}{1-p(h)} & \text{if } p(h|e) \geq p(h) \\ \frac{p(h|e)-p(h)}{p(h)} & \text{if } p(h|e) < p(h) \end{cases}$$

According to both measures, $\neg h$ is supported by e to the same degree that h is *undermined* by e:

$$-l(h,e) = l(\neg h,e) \qquad\qquad -z(h,e) = z(\neg h,e)$$

Such symmetry properties are sensible adequacy conditions for measures of evidential support (Eells and Fitelson 2002; Crupi, Tentori and González 2007), but they are at odds with the asymmetric roles of hypotheses in NHST and unattractive for degree of corroboration. There, it is not even clear what it could mean that $\neg h$ is corroborated.

These objections undermine the Monism Thesis sufficiently to motivate an explication of degree of corroboration on independent grounds. That is, we will set up adequacy conditions on a measure of corroboration that differ from the standard adequacy conditions on evidential support (see Crupi 2014; Crupi and Tentori 2014). This does not rule out that a support measure may perform a double duty as an adequate measure of corroboration: it just means that both concepts are explicated independently. On the basis of our proposed explication, we re-investigate the relationship between corroboration and evidential support (Section 6). We begin by discussing Popper's classical proposal for a measure of corroboration.

3 Popper's measure of degree of corroboration

Popper's first writings on degree of corroboration, that is, chapter 10 of the "Logic of Scientific Discovery", do not engage in a quantitative explication. Apparently, this task is deferred to a scientist's common sense. However, this move makes the entire concept of corroboration vulnerable to the charge of subjectivism: without a quantitative criterion, it is not clear which corroboration judgments are sound and which aren't (Good 1968: 136). Especially if we aim at gaining *objective knowledge* from hypothesis tests, we need a precise explication of degree of corroboration.

Popper faces this challenge in a couple of *BJPS* articles (Popper 1954, 1957, 1958) that form, together with a short introduction, appendix ix) of his "Logic of Scientific Discovery". In these articles, Popper develops and defends a measure of degree of corroboration. Popper argues that this measure cannot be a probability in the sense of Carnap (1950), that is, it is no measure of the plausibility of the tested hypothesis conditional on the observed evidence. In Popper's view, even an unlikely hypothesis can be highly corroborated if it is sufficiently informative and well-supported by the evidence.

To characterize appropriate corroboration measures Popper comes up with a list of desiderata reproduced below. Their rationale is twofold: first, corroboration increases with the mutual relevance of e and h, second, informative hypotheses are preferred over uninformative ones.

Regarding the formal nature of the desiderata, we assume that e and h are among the closed sentences $\mathfrak{L}$ of a language L. A corroboration measure is described by a function $\mathfrak{L}^2 \times \mathfrak{P} \rightarrow \mathbb{R}$, where $\mathfrak{P}$ is the set of probability measures on the σ-algebra generated by $\mathfrak{L}$. This function assigns a real-valued degree of corroboration to any pair of sentences together with a probability (degree of belief) function. For the sake of simplicity, we will omit explicit reference to background knowledge and assume that it is implicit in the probability function $p(\cdot)$.

I $c(h,e) >/=/< 0$ if and only if $p(e|h) >/=/< p(e)$.

This is a classical *positive probabilistic relevance condition*: e corroborates h just in case h makes e more expected. Vice versa, if h makes e less expected, the degree of corroboration is negative. This condition is also in line with Popper's remark (1979: 18) that corroboration is, like preference, essentially contrastive.

II $-1 = c(h, \neg h) \leq c(h,e) \leq c(h,h) \leq 1$.

III $c(h,h) = 1 - p(h)$.

IV If $e \models h$ then $c(h,e) = 1 - p(h)$.

V If $e \models \neg h$ then $c(h,e) = -1$.

These conditions determine under which conditions the measure of corroboration takes its extremal values. Minimal degree of corroboration is obtained if the evidence refutes the hypothesis (V). Conversely, the most corroborating piece of evidence e is the one that verifies h. In this case, degree of corroboration is equal to the *improbability* of h (II, III, IV), which is supposed to express the informativity, testability and empirical content of h (Popper 1934/2002: 268–269; see also Popper 1963: 385–387; Rowbottom 2013: 742–744). This is motivated as follows:

> Science does not aim, primarily, at high probabilities. It aims at a *high informative content*, well backed by experience. But a hypothesis may be very probable simply because it tells us nothing, or little. (Popper 1934/2002: 416)

Assigning a corroboration bonus to highly informative and testable hypotheses fits, of course, into a critical rationalist picture about aims and method of science. The probability $p(h)$ is interpreted in Carnap's (1950) logical sense—a point that need not worry us now, but to which we return later.

VI $c(h,e) \geq 0$ increases with the power of h to explain e.

VII If $p(h) = p(h')$, then $c(h,e) > c(h',e')$ if and only if $p(h|e) > p(h'|e')$.

These conditions reiterate the positive relevance rationale from condition I, and make it more precise. Regarding condition VI, Popper (2002: 416) defines explanatory power according to the formula $\mathcal{E}(e,h) = (p(e|h) - p(e))/(p(e|h) + p(e))$, another measure of the positive relevance between e and h. Condition VII states that corroboration essentially co-varies with posterior probability whenever the prior probabilities are equal.

VIII If $h \models e$, then

a) $c(h,e) \geq 0$;

b) $c(h,e)$ is an increasing function of $1 - p(e)$;

c) $c(h,e)$ is an increasing function of $p(h)$.

IX If $\neg h$ is consistent and $\neg h \models e$, then

a) $c(h,e) \leq 0$;

b) $c(h,e)$ is an increasing function of $p(e)$;

c) $c(h,e)$ is an increasing function of $p(h)$.

Condition VIII demands that corroboration gained from a successful deductive prediction co-vary with the informativity of the evidence and the prior probability of the hypothesis. The latter requirement stands in a certain tension with conditions III and IV, which emphasize the inverse relationship between prior probability and degree of corroboration. Condition IX mirrors condition VIII for the negative case.

These desiderata pull into different directions. Some of them are motivated by considerations of positive relevance and evidential support (I, II, VI, VII, VIIIb), others assign a bonus to the informativity, content or improbability of h (III, IV). In particular, degree of corroboration is maximal if and only if (!) a hypothesis with probability zero is entailed by the evidence. That is, Popper's desiderata reconcile two essential criteria for theory acceptance (Hempel 1960; Huber 2008; Brössel 2013): the support in favor of h, and the logical strength, informativity and empirical content of h.

Popper then develops a corroboration measure that satisfies all these desiderata, namely:

$$c_P(h,e) = \frac{p(e|h) - p(e)}{p(e|h) - p(eh) + p(e)} \tag{2}$$

Before I explain my own take on Popper's proposal, I would like to examine several objections made in the literature.

Rowbottom (2013) objects to Popper that if he were consistent with his claim made elsewhere that universal generalizations always have probability zero, he should restrict his measure to that case, because these hypotheses are also the most important ones in science. Then, c_P can be written as

$$c'_P(h,e) = \frac{p(e|h) - p(e|\neg h)}{p(e|h) + p(e|\neg h)} \tag{3}$$

which is ordinally equivalent to the log-likelihood measure l of evidential support. Rowbottom continues as follows:

> Compare two scenarios in which e is found to be true, the first in which $p(e|h) = 1$ and $p(e) = 0.1$, and the second in which $p(e'|h) = 0.1$ and $p(e') = 0.01$. According to (3), h is equally corroborated, i.e. has a corroboration value of 9/11, in each scenario. This is patently absurd, however, since in the former scenario e is entailed by h [...] (and discovery of $\neg e$ would have falsified the conjunct), whereas in the latter scenario h makes no notable contribution to predicting e [...] (and discovery of $\neg e$ would hardly have been a blow for h [...]). (Rowbottom 2013: 740)

Rowbottom then concludes that $c'_P(h,e)$ is not suitable as a measure of corroboration. Corroboration should be sensitive to the fact that if $p(e|h)$ were very high, an observation of $\neg e$ would virtually falsify h. When e is observed, h has survived a severe refutation attempt and should count as better corroborated than if e' had been observed. $c'_P(h,e)$ fails to rescue this intuition.

A natural reply is that the severity of a test is a *methodological* virtue, but irrelevant for the *evidential* interpretation of the results (Sprenger 2009).

Hence, it should not affect degree of corroboration. Moreover, in continuous or large discrete sample spaces, every piece of evidence e typically has a very low probability of being observed; often it is zero. What makes e corroborating evidence for h is not so much the high value of $p(e|h)$, but the fact that the probability of e (respectively the value of the density function) is much higher than for competing hypotheses. This is, after all, the rationale behind statistical hypotheses tests that typically deal with continuous sample spaces. Hacking (1965), Spielman (1974) and Royall (1997) have, among others, advanced forceful arguments that any statistical hypothesis test must make reference to explicit or implicit alternatives (see also Sprenger 2014). Popper could refer to these arguments in order to deflect Rowbottom's criticism.

A second criticism, observed by Díez (2011: 196), is based on the observation that by VII, e corroborates h more than e' if and only if it raises the probability of h to a higher value than e' does ($c(h,e) > c(h,e')$ iff $p(h|e) > p(h|e')$). According to Díez, the co-variation of posterior probability and corroboration clashes with Popper's dismissal of posterior probability as a criterion for theory choice: "this rule is equivalent to the following rule: choose always the hypothesis which has the highest degree of *ad hoc* character" (Popper 1963: 385). However, condition VII only states which of two pieces of evidence confirms a peculiar theory to a higher degree. In other words, it is restricted to theories with the same informative content. Therefore, the "ad hoc" criticism does not apply in this case.

Third, Díez objects that neither $h \models e$, nor $e \models h$, nor $h \equiv e$ is a sufficient condition for maximal corroboration. Some of these conditions (e.g., $h \equiv e$) are indeed compelling sufficient conditions for maximal *evidential support* (see Crupi 2014). However, the relevance of e for h is not the only factor that affects degree of corroboration: also the informativity of h determines its corroborability (see condition IV). From Popper's point of view, it does make sense that $c(h,h) > c(h',h')$ if and only if h has more empirical content than h'.

In my view, a fourth criticism poses bigger problems for Popper By VIIIc, degree of corroboration co-varies with the prior probability of h whenever h entails e. That is, if h and h' successfully predict e ($h \models e$ and $h' \models e$), then the corroboration ranking tracks the prior probability of h

and h'. This runs contrary to Popper's intentions about the significance of empirical content/testability as a contributing factor to degree of corroboration. Since deductive entailment between theory and evidence is a classical case of prediction in science and a showcase for critical rationalist reasoning, this result is especially worrisome.

Fifth and last, there is an inconsistency in Popper's suggestions for interpreting the probabilities in c_P. Since he is opposed to any subjective interpretation, he proposes a frequentist interpretation for the likelihood $p(e|h)$ and the marginal likelihood $p(e)$, and a logical interpretation for the probability of the hypothesis $p(h)$, which is required to calculate $p(eh) = p(e|h)p(h)$. These moves are quite *ad hoc*, and Popper does not specify a bridge principle for combining these different types of probabilities. Moreover, determining the relative frequency of e or the logical probability of h is a hard problem for which Popper provides little guidance. Of course, we could just interpret all probabilities in a subjective way, but this move would not suit Popper's general philosophical framework (Popper 1934/2002, ch. 8). It would also require an additional and far from obvious argument that a subjective interpretation does not compromise the alleged objectivity of a measure of corroboration.

Summing up, Popper's measure c_P suffers from severe formal and conceptual shortcomings. The crucial question is now: which conclusions do we draw from Popper's failure to adequately explicate degree of corroboration? Should we just come up with a different probabilistic relevance measure? Or change the framework altogether?

4 Corroboration and positive relevance

This section shows two impossibility results for corroboration measures that (i) are built on the notion of positive probabilistic relevance between e and h, that is, e corroborates h whenever $p(e|h) > p(e)$; (ii) preserve Popper's intuition that corroboration should in general not co-vary with prior probability; (iii) satisfy some weak and plausible constraints.

The first condition is mainly formal in nature (cf. Schupbach and Sprenger 2011; Crupi 2014):

Formality There exists a function $f : [0,1]^3 \rightarrow \mathbb{R}$ such that for all $e, h \in \mathfrak{L}$

and $p(\cdot) \in \mathfrak{P}$,

$$c(h, e) = f(p(e|h), p(e), p(h)).$$

This condition states that degree of corroboration depends on the joint probability distribution of e and h, since the three arguments of f are sufficient to determine the entire distribution, degenerate cases left aside. In order to keep the playing field level, we have focused on the same quantities that figure in Popper's measure of corroboration.

Now we state the first substantial condition:

Weak Law of Likelihood (WLL) For mutually exclusive hypotheses $h_1, h_2 \in \mathfrak{L}$, $e \in \mathfrak{L}$ and $p(\cdot) \in \mathfrak{P}$, if

$$p(e|h_1) \geq p(e|h_2) \qquad\qquad p(\neg e|\neg h_1) \geq p(\neg e|\neg h_2) \tag{4}$$

with one inequality being strict, then $c(h_1, e) > c(h_2, e)$.

The WLL has been defended as capturing a "core message of Bayes' Theorem" (Joyce 2008) and as a non-negotiable adequacy condition on measures of evidential support (e.g., Brössel 2013). If h_1 predicts e better than h_2, and $\neg h_1$ predicts $\neg e$ better than $\neg h_2$ does, then h_1 performs better than h_2. Since this reasoning only applies to the predictive performance of the competing hypotheses, it is even more compelling for corroboration than for evidential support. The version given here is in one sense weaker and in one sense stronger than Joyce's original formulation: it is stronger because only one inequality has to be strict (see also Brössel 2013: 395–396); it is weaker because the WLL has been restricted to mutually exclusive hypotheses, where our intuitions are more reliable.

Another condition deals with irrelevant evidence:

Screened-Off Evidence Let $e_1, e_2, h \in \mathfrak{L}$ and $p \in \mathfrak{P}$. If e_2 is probabilistically independent of e_1, h, and $e_1 \wedge h$, then $c(h, e_1) = c(h, e_1 \wedge e_2)$.

This condition prominently figures in several explications of evidential support and explanatory power (e.g., Kemeny and Oppenheim 1952; Schupbach and Sprenger 2011). But it is also very sensible with respect to degree of corroboration. Extra evidence which is irrelevant in any respect ($e_2 \perp\!\!\!\perp e_1, h, e_1 \wedge h$) should not change the evaluation of an experiment where h has been tested and evidence e_1 has been observed. Imagine, for example, that a scientist tests the hypothesis that a high pitch facilitates voice recognition. As the scientists's university is interested in improving the planning of lab experiments, she also collects data on the times when participants drop in, which slots are busy, which ones are quiet, etc. Plausibly, these data satisfy the independence conditions of `Screened-Off Evidence`, and equally plausibly, they do not influence its degree of corroboration.

The next adequacy condition is motivated by the problem of irrelevant conjunctions, a well-known challenge for Bayesian measures of evidential support (e.g., Fitelson 2002; Hawthorne and Fitelson 2004). Assume that a hypothesis h, such as General Theory of Relativity (GTR), logically implies a phenomenon e, such as the perihelion shift of Mercury. This observation corroborates GTR: logical implication is a special case of probabilistic relevance.

However, once we add an utterly irrelevant proposition h' = "the chicken came before the egg" to the hypothesis, it seems that e corroborates $h \wedge h'$—the *conjunction* of GTR and the chicken-egg hypothesis—not more than h (if at all). After all, h' was not tested by the observations we made. It has no record of past performance to which it could appeal. This motivates the following constraint:

Irrelevant Conjunctions Assume the following conditions on $h, h', e \in \mathfrak{L}$ and $p \in \mathfrak{P}$ are satisfied:

(1) h and h' are consistent and $p(h \wedge h') < p(h)$;

(2) $p(e) \in (0,1)$;

(3) $h \models e$;

(4) $p(e|h') = p(e)$.

Then it is always the case that $c(h \wedge h', e) \leq c(h, e)$.

This requirement states that for any non-trivial hypothesis h' that is consistent with h and irrelevant for e ((1), (4)), $h \wedge h'$ is no corroborated more than h whenever h non-trivially entails e ((2), (3)). Indeed, it would be strange if corroboration could be increased "for free" by attaching irrelevant propositions. Plausibly, this requirement could be strengthened to a strict inequality, but for our purposes, the weaker formulation is sufficient, and in this version, it is also weaker than Popper's VIIIc.

Finally, we want to account for the intuition that highly corroborated hypotheses are *informative* propositions backed by the evidence (see Popper's conditions II-IV). Unlike evidential support, corroboration contains an element of severe testing: the hypothesis should run a risk of being falsified, and high informativity and empirical content contribute to this goal. This motivates the following desiderata, one of them being slightly weaker than the other:

Strong Informativity The informativity/empirical content of a proposition can increase degree of corroboration, ceteris paribus. That is, there are $h, h', e, e' \in \mathfrak{L}$ and $p \in \mathfrak{P}$ with $p(e|h) > p(e)$, $p(e'|h') > p(e')$ such that

(1) $p(e|h) = p(e'|h')$, $p(e) = p(e')$;

(2) $1/2 \geq p(h) > p(h')$;

(3) $c(h, e) > c(h', e')$.

Weak Informativity Degree of corroboration $c(h, e)$ does not generally co-vary with the prior probability of h. That is, there are $h, h', e, e' \in \mathfrak{L}$ and $p \in \mathfrak{P}$ with $p(e|h) > p(e)$, $p(e'|h') > p(e')$ such that

(1) $p(e|h) = p(e'|h')$, $p(e) = p(e')$;

(2) $1/2 \geq p(h) > p(h')$;

(3) $c(h, e) \geq c(h', e')$.

The intuition behind `Weak Informativity` can also be expressed as follows: corroboration does not, in the first place, assess the prior plausibility of a hypothesis; therefore $c(h, e)$ should not in general co-vary with the prior plausibility of h. To this, `Strong Informativity` adds that low prior probability/high empirical content can even be corroboration-conducive. Note that the requirement $1/2 \geq p(h), p(h')$ is purely technical and philosophically innocuous.

At this point, it is possible to demonstrate that the listed conditions are incompatible with each other. First, a consequence of `Weak Law of Likelihood` is that corroboration is an increasing function of the prior probability of a hypothesis, which clashes directly with Strong/Weak Informativity:

Theorem 1 No measure of corroboration $c(h,e)$ constructed according to Formality can satisfy Weak Law of Likelihood and Weak/Strong Informativity at the same time.

Second, and perhaps more surprisingly, Strong Informativity clashes with Irrelevant Conjunctions and Screened-Off Evidence:

Theorem 2 No measure of corroboration $c(h,e)$ constructed according to Formality can satisfy Screened-Off Evidence, Irrelevant Conjunctions and Strong Informativity at the same time.

Thus, the intuition behind Strong/Weak Informativity cannot be satisfied if other plausible adequacy constraints on degree of corroboration are accepted. All proofs are given in the appendix. Notably, the result of Theorem 2 can be extended to Weak Informativity if we make the assumption that irrelevant conjunctions *dilute* the degree of corroboration, rather than not increasing it. Of course, all this does not show that explicating degree of corroboration is a futile project. Rather, it reveals a fundamental and probably insoluble tension between the two main contributing factors of corroboration that Popper identifies (see the quote on p. 10): probabilistic relevance and empirical content.

The two theorems suggest two interpretations: (i) either we cannot adequately explicate corroboration in a probabilistic relevance framework, or (ii) the entire concept of corroboration is overloaded with intuitions pointing into different directions. However, the problem does not seem to lie with the adequacy conditions. Screened-Off Evidence is highly plausible for both corroboration and evidential support. Law of Likelihood and Irrelevant Conjunctions are complementary; yet both of them lead to impossibility results. Finally, if we give up Strong/Weak Informativity, we lose a crucial characteristic of corroboration in hypothesis testing, namely that it applies in particular to precise and informative hypotheses.

This diagnosis points us to re-thinking the entire conceptual framework, as expressed in Formality. Perhaps it is neither necessary nor sufficient to base a corroboration judgment on the joint probability distribution of h and e? If we explicate corroboration in terms of probabilistic relevance, judgments of corroboration compare the merits of h with the merits of $\neg h$, defined as the aggregate of alternatives to h. However, a comparison to such an aggregate does not make much sense in many contexts of statistical hypothesis testing where we deal with a multitude of distinct alternatives h_i, $i \in \mathbb{N}$. To calculate $p(e|\neg h) = \sum_i p(e|h_i)$, we would have to know the prior probabilities $p(h_i)$, an assignment that many scientists refuse to make in practice. This framework also fails at describing how hypotheses with probability zero can be corroborated, one of the central distinctions between evidential support and degree of corroboration.

The formal results of this section can then be regarded as formal vindications of the arguments advanced against the Monism Thesis in Section 2. They show that we cannot jointly satisfy a set of reasonable desiderata about degree of corroboration in a probabilistic relevance framework. All this suggests that we should develop explications of degree of corroboration in a different conceptual framework.

5 A new framework for measuring corroboration

One of the main objections to probabilistic relevance explications of corroboration consists in the way the alternative hypothesis is treated. In NHST, it is common that the null is a precise hypothesis $h_0 : \theta = \theta_0$ which is tested against a composite hypothesis $h_1 : \theta \neq \theta_0$. Such composite hypotheses are rather an umbrella for distinct alternatives than a probabilistic aggregate of alternatives, but in evaluations in terms of evidential support, they are treated as a single hypothesis, namely the negation of h_0. In practice, however, we want to *simulteanously test the null against a set of distinct alternatives*, not to test it against a single, aggregate hypothesis. That is, degree of corroboration should be sensitive to the fine-structure of the alternatives.

A simple example may illustrate this thesis. Suppose one wants to infer the mean value θ of a Normal distribution, where the null hypothesis $h_0 : X \sim N(0, 1)$ is tested against the alternatives $h_1 : X \sim N(2, 1)$ and $h_2 : X \sim N(-2, 1)$. Then, some observations (e.g., $x \approx 2$) will be well explained by h_1 and be poorly explained by h_2, while other observations (e.g., $x \approx -2$)

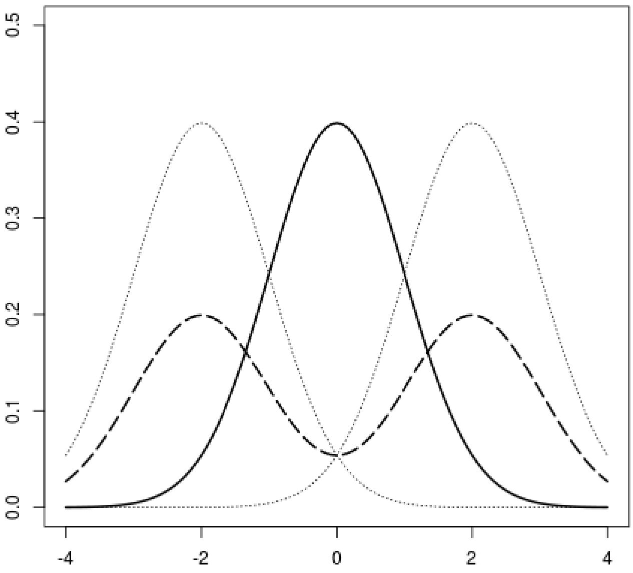

Figure 1: Testing the null hypothesis $h_0 : X \sim N(0,1)$ (full line) against the aggregate of $h_1 : X \sim N(2,1)$ and $h_2 : X \sim N(-2,1)$ (dashed line). The dotted lines represent h_1 and h_2 themselves.

will be well explained by h_2 and be poorly explained by h_1. This tension gets lost when we only consider the *aggregate* of h_1 and h_2, which is a bimodal distribution with peaks at -2 and 2: both observations ($x \approx -2$ and $x \approx 2$) are well explained by the aggregate alternative hypothesis and favor it over the null. Thus, this testing problem is conceptually and mathematically quite different from the original one against two distinct hypotheses. See Figure 1 for a graphical illustration.

The rest of the section develops two measures of degree of corroboration that are sensitive to the partition of alternative hypotheses $\mathcal{H} = \{h_0, h_1, h_2, ...\}$. They summarize judgments of evidential favoring into a single number that expresses the performance of the null hypothesis h_0 in a test with evidence e. The explications focus on the evidential aspect of corroboration and leave out some methodological virtues, such as the severity of the test or issues pertaining to experimental design. This is in line with Popper's own remarks that such virtues cannot be fully formalized (Popper 1956/83: 154). From now on it will be assumed that the data have been collected in genuine tests of h_0.

A first measure of corroboration is derived from three adequacy criteria CA1-CA3 motivated below.

The first requirement is based on a thought from Section 2: degree of corroboration indicates whether h_0 is a suitable proxy for a more general model. In other words, if a hypothesis h_0 is highly corroborated, the loss in accuracy that we suffer by replacing the general model $\mathcal{H}$ by h_0 is reasonably small. For example, assume the null states that manipulating some independent variables has no effect on the data. In practice, there will always be some small effect, but we want to know whether it is *negligibly* small. This would be the practical significance of an "inference to the null hypothesis". This question is highly relevant to scientific practice, and it is the one that we may reasonably regard as the motivation behind the entire null hypothesis testing business. Therefore we demand

CA1 Corroboration should quantify the *average score gain* of replacing the general model $\mathcal{H}$ with the null hypothesis h_0. That is, for a suitable scoring rule $S(h_0, e)$, the degree of corroboration that e provides for

h_0 relative to $\mathcal{H}$ can be defined as

$$c_{\mathcal{H}}(h_0, e) = \sum_{h_i \in \mathcal{H}} \omega_i \cdot (S(h_0, e) - S(h_i, e)) \tag{5}$$

where ω_i denotes the relative weight that each element of $\mathcal{H}$ receives.

That is, degree of corroboration sums up the score differences between h_0 and the alternatives h_i, weighted by ω_i. Note that the ω_i need not be interpreted as degrees of belief that a particular hypothesis is true or empirically adequate. Scientists do not always entertain such degrees of belief in the hypotheses they investigate. They rather regard them as useful idealizations (see also the discussion in Section 2). For example, in the assessment of global climate models, most physical scientists are convinced that none of the considered models is true or empirically adequate, and they use a broad set scientific values for weighting these models (e.g., Frame et al. 2007). Hence, the ω_i are supposed to reflect the relative standing of the alternatives in the scientific community, including cognitive values such as fruitfulness, scope, etc. In other words, the above definition is neutral with respect to the Bayesian/frequentist divide in statistical inference. To repeat, the main move of CA1 is to replace a vague explicandum—degree of corroboration—by a precise and fruitful explicatum, namely average gain in predictive power by accepting the null hypothesis.

The next step is to find a suitable scoring rule $S(h_0, e)$. For this, we impose two more adequacy criteria:

CA2 There exists a real-valued, continuous function $f : [0, 1] \to \mathbb{R}$ such that $S(h_0, e) := f(p(e|h_0))$. In other words, the score of h_0 on evidence e only depends on the probability of e under h_0.

CA3 The scoring rule $S(h_0, e)$ is additive with regard to evidence that is independent under h_0. In other words, if $e \perp\!\!\!\perp e' | h_0$, then

$$S(h_0, e \wedge e') = S(h_0, e) + S(h_0, e')$$

CA2 expresses the natural idea that score depends on and increases with predictive performance. If a likely event occurs, then the score is high; if an unlikely event occurs, the score is low. CA3 demands that scores on independent pieces of evidence add up. Similar requirements and derivations can, for different contexts, also be found in Good (1952), Bernardo (1999) and Williamson (2010). I leave it to future research to find explications of degree of corroboration based on different scoring rules.

It can be demonstrated easily that CA2-CA3 leads to a logarithmic scoring rule $S(h_0, e) = \log p(e|h_0)$, and that they uniquely determine, together with CA1, the following measure of degree of corroboration:

Theorem 3 The only measure of corroboration that satisfies CA1-CA3 has the form

$$C_{\mathcal{H}}(h_0, e) = \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e|h_0)}{p(e|h_i)}. \tag{6}$$

where the logarithm has an arbitrary positive basis.

Note that even if the weights ω_i sum up to infinity (e.g., in the case of improper Bayesian priors), the degree of corroboration can be finite. When many alternatives are hard to distinguish empirically from h_0, the log-likelihood ratio will be close to zero, and this may suffice for assigning a finite value to $C_{\mathcal{H}}$. Standardly, however, we will assume that $\sum_i \omega_i = 1$.

Our explication satisfies all four conceptual requirements that we have advanced for degree of corroboration in Section 2: First, there is no implicit presumption that one of the hypotheses is true, or that high degree of corroboration is truth-conducive. Second, $C_{\mathcal{H}}$ is independent of (an increase in) subjective degree of belief. Third, hypotheses with zero probability can be corroborated straightforwardly. Fourth, by means of splitting the alternative into several individual hypotheses, $C_{\mathcal{H}}$ preserves the essential asymmetry of corroboration judgments.

Notably, the independence property CA3 of the scoring rule $S(h_0, e)$ is preserved by $C_{\mathcal{H}}$. If two pieces of evidence e and e' are independent under the competing hypotheses, their degree of corroboration adds up:

$$\begin{aligned}
C_{\mathcal{H}}(h_0, e \wedge e') &= \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e \wedge e'|h_0)}{p(e \wedge e'|h_i)} \\
&= \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e|h_0) \cdot p(e'|h_0)}{p(e|h_1) \cdot p(e'|h_i)} \\
&= \sum_{h_i \in \mathcal{H}} \omega_i \left(\log \frac{p(e|h_0)}{p(e|h_i)} + \log \frac{p(e'|h_0)}{p(e'|h_1)} \right) \\
&= C_{\mathcal{H}}(h_0, e) + C_{\mathcal{H}}(h_0, e')
\end{aligned}$$

For example, the corroboration gained in two sequential, independent experiments is the sum of the individual degrees of corroboration. This property makes $C_{\mathcal{H}}$ very useful for the meta-analysis of several experiments.

Of course, $C_{\mathcal{H}}$ is not the only sensible measure of corroboration that one can develop from the qualitative constraints discussed at the beginning of this section. There is also an obvious objection to this measure, namely that it is far too easy to obtain maximal, that is, infinite corroboation. Whenever one of the alternatives h_i is incompatible with e, $C_{\mathcal{H}}(h_0, e) = \infty$. But clearly, a hypothesis that performs poorly with respect to most relevant alternatives should not count as maximally corroborated just because another hypothesis happens to assign probability zero to the observed evidence.

In response, three arguments can be advanced. First, the chosen explication of corroboration also has a definite advantage: it is easy to add up degree of corroboration from different experiments. Testing a hypothesis in many experiments naturally emerges as better than testing it in just one experiment. Second, for the purpose of statistical testing, the above worry is quite theoretical since the relevant probability densities are usually strictly positive in the relevant probability space. Third, we can modify $C_{\mathcal{H}}$ in a way that resolves this problem while preserving its most important qualitative properties.

This last suggestion will now be elaborated in detail. First, we replace CA1 by a slightly modified condition:

CA1′ Corroboration should quantify the *average score gain* of replacing the general model $\mathcal{H}$ with the null hypothesis h_0. That is, for a suitable scoring rule $S(h_0, e)$, the degree of corroboration that e provides for h_0 relative to $\mathcal{H}$ can be defined as

$$c_{\mathcal{H}}(h_0, e) = \sum_{h_i \in \mathcal{H}} \omega_i \cdot f_S(h_0, h_i, e) \qquad (7)$$

where ω_i denotes the relative weight that each element of $\mathcal{H}$ receives ($\sum_i \omega_i = 1$), and $f_S(h_0, h_i, e)$ is a monotonous transformation of $S(h_0, e) - S(h_i, e)$.

That is, the score difference $S(h_0, e) - S(h_i, e)$ may now be replaced by a monotonous transformation of this quantity. This keeps the basic qualitative structure intact, but allows for a more intuitive scaling of degrees of corroboration. CA2′=CA2 and CA3′=CA3 remain unchanged.

CA4′ $f_S(h_0, h_1, e)$ is the simplest function of the form

$$f_S(h_0, h_1, e) = \frac{\sum_{j=1}^{m} \sum_{k=1}^{m} c_{jk}\, p(e|h_0)^j\, p(e|h_1)^k}{\sum_{j=1}^{n} \sum_{k=1}^{n} d_{jk}\, p(e|h_0)^j\, p(e|h_1)^k} \tag{8}$$

with the properties

- $f_S(h_0, h_1, e) = 0$ if $p(e|h_0) = p(e|h_1)$.
- $f_S(h_0, h_1, e) = 1$ if $p(e|h_0) = 1$ and $p(e|h_1) = 0$.
- $f_S(h_0, h_1, e) = -1$ if $p(e|h_0) = 0$ and $p(e|h_1) = 1$.

This requirement is in parts motivated by CA2 which demands that $S(h_0, e)$ be a function of $p(e|h_0)$ only. Hence, $f_S(h_0, h_1, e)$ only depends on $p(e|h_0)$ and $p(e|h_1)$. The form of the function specified in (8) is very flexible since any function in the interval $[0, 1]^2$ can be approximated arbitrarily well by a rational function. It is therefore no substantial philosophical constraint on the measure of corroboration that we choose. The three conditions at the end of CA4 fix the neutral value of f_S at zero and the maximal/minimal values at 1 and -1, in order to obtain a balanced aggregate score.

Theorem 4 CA1′–CA4′ jointly determine the unique function

$$f_S(h_0, h_1, e) = \frac{p(e|h_0) - p(e|h_1)}{p(e|h_0) + p(e|h_1)}$$

and the corroboration measure

$$C'_{\mathcal{H}}(h_0, e) = \sum_{h_i \in \mathcal{H}} \omega_i \cdot \frac{p(e|h_0) - p(e|h_i)}{p(e|h_0) + p(e|h_i)} \tag{9}$$

$C'_{\mathcal{H}}$ does not have the property that a logical implication $e \models \neg h_i$ leads to an infinite corroboration value since the scores are bounded by ±1. Since the structure of f_S equals the well-studied Kemeny-Oppenheim measure of evidential support (Kemeny and Oppenheim 1952), we can also deliver an intuitive interpretation for observed degrees of corroboration:

$$C'_{\mathcal{H}}(h_0, e) \in \begin{cases} [0;1/4] & \text{weak corroboration} \\ [1/4;1/2] & \text{moderate corroboration} \\ [1/2;3/4] & \text{substantial corroboration} \\ [3/4;1] & \text{strong corroboration} \end{cases}$$

Negative corroboration could then be read as evidence that undermines the null hypothesis: there is no predictive gain in adopting h_0 as a simplification or a proxy for the more general parametric model. Hence, unless there is strong theoretical reason to stick to h_0, we should replace it by a different hypothesis.

I leave it to the reader to choose between $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$. What counts for the purpose of this paper is that both are sound explications of degree of corroboration that share a lot of desirable properties. This claim will be elaborated in the following section.

6 From corroboration back to evidential support

This section investigates the properties of our corroboration measures $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ and relates them to measures of evidential support. Crucially, it will be shown that they satisfy the desiderata on measures of corroboration that we imposed in Section 4, at least in a modified version that is applicable to the novel framework.

First, a general observation. Most (normalized) measures of evidential support satisfy the constraint $\mathfrak{c}(h_0, e) = 0$ if and only if $p(e|h_0) = p(e|\neg h_0)$. This corresponds to the idea that probabilistically independent evidence neither raises or lowers the probability of a hypothesis. However, it is *not* the case that $C_{\mathcal{H}}(h_0, e) = 0$ if and only if $p(e|h_0) = p(e|\neg h_0)$, and analogously for $C'_{\mathcal{H}}$. Should this violation of a standard neutrality constraint give us reason to worry?

I do not think so. One of the rationales behind the construction of $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ was to eliminate the idea that the alternative to h_0 should be constructed as an aggregate hypothesis $\neg h_0$: this view is at odds with asymmetric nature of hypothesis tests and the questions they ask. Instead, corroboration should describe how well a hypothesis fares with respect to a *set* of alternatives. The neutrality point is then not defined as the point where e leaves the probabilities of h_0 and $\neg h_0$ unchanged, but as *the point where evidence for and against h_0 cancels out*. This redefinition of evidential neutrality is one of the main conceptual innovations with respect to the evidential support paradigm.

All this implies that `Weak Law of Likelihood` cannot be formulated consistently for measures of corroboration, since it depends on $p(e|\neg h_0)$. However, $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ satisfy the stronger

Law of Likelihood (LL) For mutually exclusive hypotheses $\mathcal{H} = \{h_0, h_1, \ldots\}$, $\mathcal{H} \subset \mathfrak{L}$, $e \in \mathfrak{L}$ and $p(\cdot) \in \mathfrak{P}$ and a measure of corroboration $c_{\mathcal{H}}(h, e)$:

$$c_{\mathcal{H}}(h_i, e) >/=/< c_{\mathcal{H}}(h_j, e) \qquad p(e|h_i) >/=/< p(e|h_j)$$

That both measures satisfy LL can be seen by the following result:

Theorem 5 For the difference in degree of corroboration between two hypotheses $h_0, h_1 \in \mathfrak{L}$, the following equalities hold:

$$\begin{aligned} \Delta C_{\mathcal{H}}(h_0, h_1, e) &:= C_{\mathcal{H}}(h_0, e) - C_{\mathcal{H}}(h_1, e) = \log \frac{p(e|h_0)}{p(e|h_1)} \\ \Delta C'_{\mathcal{H}}(h_0, h_1, e) &:= C'_{\mathcal{H}}(h_0, e) - C'_{\mathcal{H}}(h_1, e) \\ &= (p(e|h_0) - p(e|h_1)) \sum_{h_i \in \mathcal{H}} 2\omega_i \frac{p(e|h_i)}{(p(e|h_0) + p(e|h_i))\,(p(e|h_1) + p(e|h_i))} \end{aligned}$$

These equations show that the ordinal relations between $C_{\mathcal{H}}(h_0, e)$ and $C_{\mathcal{H}}(h_1, e)$ only depend on whether $p(e|h_0)$ is greater than $p(e|h_1)$. Analogously for $C'_{\mathcal{H}}$. Thus, `Law of Likelihood` is satisfied, in agreement with the idea that degree of corroboration is an indicator of past performance. We also observe that adding irrelevant conjunctions h' to h_0, that is, hypotheses with the property $p(e|h_0) = p(e|h_0 \wedge h')$, will not affect the degree of corroboration. A fortiori, both measures satisfy the `Irrelevant Conjunctions` property. Actually, $c(h_0, e) = c(h_0 \wedge h', e)$ if h_0 entails e may be the only option to sail between Skylla (Popper's VIIIc: corroboration co-varies with prior probability) and Charybdis (irrelevant conjunctions increase degree of corroboration).

To evaluate the measures with regard to `Screened-Off Evidence`, we have to modify the definition of that property. I suggest to rewrite `Screened-Off Evidence` as follows: if $e' \perp\!\!\!\perp e$, $e' \perp\!\!\!\perp h_i$, and $e' \perp\!\!\!\perp (e \wedge h_i)$ for all $h_i \in \mathcal{H}$, then $C_{\mathcal{H}}(h_0, e \wedge e') = C_{\mathcal{H}}(h_0, e)$. This is a natural generalization from two competing hypotheses (h_0 and $\neg h_0$) to a larger set $\mathcal{H}$ whose members h_i may denote different values of a parameter of interest. It is then easy to observe that both $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ satisfy `Screened-Off Evidence` since they only depend on the likelihoods of the hypothesis on the evidence.

We also gain a nuanced and interesting picture of the sensitivity of corroboration to the prior standing of the corroborated hypothesis. For our measure of corroboration, we can isolate the prior weight of h_0, ω_0, from the relations of the other weights to each other:

$$C'_{\mathcal{H}}(h_0, e) = (1 - \omega_0) \cdot \sum_{h_i \in \mathcal{H} \setminus \{h_0\}} \frac{\omega_i}{1 - \omega_0} f_S(h_0, h_1, e) \tag{10}$$

since the summand containing h_0 vanishes anyway. (We obtain the same calculations for $C_{\mathcal{H}}$ by letting $f_S(h_0, h_1, e) = \log(p(e|h_0)/p(e|h_1))$.) Since $\omega_0 = 1 - \sum_{i \neq 0} \omega_i$, the factors $\omega_i/(1 - \omega_0)$ only depend on the ratios of the ω_i to each other. Hence, equation (10) can be read as "degree of corroboration of h_0 = improbability of $h_0 \times$ average predictive gain by adopting h_0". Together with the rescaling $\omega'_i := \omega_i/(1 - \omega_0)$, this simple operation allows us to partially derive $C'_{\mathcal{H}}$ with respect to ω_0:

$$\frac{\partial C'_{\mathcal{H}}(h_0, e)}{d\omega_0} = - \sum_{h_i \in \mathcal{H} \setminus \{h_0\}} \omega'_i f_S(h_0, h_1, e),$$

since $f(S(h_0, e) - S(h_1, e))$ is by construction independent of the ω_i. It then transpires that $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ decrease in ω_0, that is, they increase with the *improbability* of h_0. This is because the term on the right side of (6) has the same sign as $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$.

Popper proposed that informativity, testability and empirical content, as measured by the improbability of a hypothesis, are always corroboration-conducive factors. For our measures, this depends on whether or not $c_{\mathcal{H}}(h_0, e) > 0$. For a positively corroborated hypothesis, a low weight is indeed beneficial because the average gain in predictive score vis-à-vis the alternatives is bigger than for a hypothesis that already had a high weight beforehand. This shows that $C_{\mathcal{H}}$ and $C'_{\mathcal{H}}$ satisfy `(Strong/Weak) Informativity`, the desiderata that were distinctive for corroboration as opposed to evidential support. For a negative degree of corroboration, the same reasoning amplifies the "degree of refutation" of h_0. Thus, we can recover and refine the Popperian picture at the same time.

Finally, some remarks on the relationship between corroboration and evidential support. One of the most popular explications of weight of evidence, the degree to which e favors h_0 over h_1, is the aforementioned (log-)likelihood ratio $\log(p(e|h_0)/p(e|h_1))$ which is supported by a wide range of theoretical and empirical arguments (Good 1983/2009; Royall 1997; Lele 2004; Sober 2008). Theorem 5 has shown that for $C_{\mathcal{H}}$, the weight of evidence in favor of h_0 is its *excess degree of corroboration* over h_1. That is, $\Delta C_{\mathcal{H}}(h_0, h_1, e) = \log(p(e|h_0)/p(e|h_1))$. A similar relation holds for $C'_{\mathcal{H}}$. This suggests that (contrastive) evidential support can also be derived from corroboration judgments and performance differences. Symmetries in evidential support such as $\mathfrak{c}(h_0, e) = -\mathfrak{c}(\neg h_0, e)$ then naturally emerge as a consequence of symmetries in corroboration differences, such as $\Delta C_{\mathcal{H}}(h_0, h_1, e) = -\Delta C_{\mathcal{H}}(h_1, h_0, e)$.

Actually, $l(h, e)$ is not the only measure of evidential support for which such a relation can be established. For the log-ratio measure $r(h, e) = p(e|h)/p(e)$ defended by Milne (1996), we obtain

$$\begin{aligned}
C_{\mathcal{H}}(h_0, e) &= \sum_{h_i \in \mathcal{H}} \omega_i \log \left(\frac{p(e|h_0)}{p(e)} \cdot \frac{p(e)}{p(e|h_i)} \right) \\
&= \log \frac{p(e|h_0)}{p(e)} - \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e|h_i)}{p(e)} \\
&= r(h_0, e) - \sum_{h_i \in \mathcal{H}} \omega_i \, r(h_i, e)
\end{aligned}$$

which we can interpret as the difference between the support for h_0 and the average support for all hypotheses in $\mathcal{H}$—an observation that I owe to Wayne Myrvold. Dependent on the preferred explication of evidential support, we can either express corroboration in terms of support differences or derive contrastive evidential support from differences in degree of corroboration. All this suggests that corroboration and evidential support are tightly related concepts in inductive inference. How both concepts interact precisely is an exciting issue for further research.

7 Summary and discussion

The concept of degree of corroboration defines how the failure to reject a hypothesis affects its epistemic status. In other words, explicating corroboration helps to positively appraise a hypothesis that has survived a severe test. The failure to adequately formalize this concept has been a long-standing lacuna in statistics, science and philosophy: standard statistical procedures such as null hypothesis significance tests (NHSTs) are silent on non-significant results, and the critical rationalist research program in philosophy of science lacks a quantitative dimension—especially if compared to the rich Bayesian theory of evidential support. This contribution shows what a formalization of corroboration could look like, and how it fruitfully complements the Bayesian perspective on inductive inference.

In the first place, this contribution motivates why corroboration judgments cannot be replaced by judgments of evidential support, and why corroboration and support play complementary roles. Based on this characterization in Sections 1 and 2, I investigate Popper's attempt to explicate this concept in terms of probabilistic relevance. After debunking Popper's own explication, I show that positive probabilistic relevance is a problematic framework for explicating degree of corroboration. This argument culminates in two impossibility results which show that no measure of corroboration can jointly satisfy several plausible adequacy criteria.

Motivated by these findings, I develop a constructive account of degree of corroboration, which is neutral with respect to the methodological divide between various schools of inductive inference (e.g., Bayesians and frequentists). The model is based on the idea that corroboration is, unlike evidential support, assessed with respect to a *partition of alternatives* to the tested hypothesis h_0, rather than by comparing h_0 to its negation $\neg h_0$. In my explication, degree of corroboration compares the average predictive score difference between h_0 and the alternatives to h_0, with respect to evidence e. The idea is that a high degree of corroboration entitles us to replace a general model $\mathcal{H}$ with a precise hypothesis h_0 without incurring too many losses. This fits actually well with Popper's idea that corroboration serves for determining pragmatic preferences over different scientific hypotheses, but does not ground any confidence in their truth.

The chosen explication is shown to have several desirable properties: for instance, it allows for the corroboration of hypotheses with zero probability (a standard problem for the Bayesian), it shows how the informativity of a hypothesis contributes to its corroborability, how irrelevant evidence leaves degree of corroboration unchanged, etc. Moreover, corroboration differences between two hypotheses turn out to be closely related to contrastive evidential support.

All in all, this paper does not only explicate a concept that has unjustifiably fallen into oblivion: it also improves the assessment of the results of statistical hypothesis tests, and it shows how evidential support and degree of corroboration can be complementary notions in the assessment of scientific theories. Future work will explore other axiomatic characterizations for measures of corroboration, expand on their application to statistical tesing and explore the quantitative (dis)agreements between corroboration and Bayesian measures of evidence (e.g., the Bayes factor). For now, I conclude that our formalizations of corroboration lay the foundations for a new logic of statistical hypothesis testing (and NHST in particular), beyond the Bayesian/frequentist divide.

A Proofs of the theorems

Proof of Theorem 1: By `Weak Informativity`, there are $x, y, z > z'$ with $z + z' < 1$:

$$f(x, y, z) \leq f(x, y, z').$$

Choose a probability function $p(\cdot)$ such that $p(h_1) = z$, $p(h_2) = z'$, $p(h_1 \wedge h_2) = 0$, $p(e|h_1) = p(e|h_2) = x$, $p(e) = y$. This is always possible because it was assumed that $z + z' < 1$. Then it is straightforward to show that

$$\begin{aligned} p(e|\neg h_1) &= \frac{1}{1 - p(h_1)} \left[p(e|h_1) p(h_2) + p(e|\neg h_1 \neg h_2) p(\neg h_1 \neg h_2) \right] \\ p(e|\neg h_2) &= \frac{1}{1 - p(h_2)} \left[p(e|h_1) p(h_1) + p(e|\neg h_1 \neg h_2) p(\neg h_1 \neg h_2) \right] \end{aligned}$$

From this we can infer

$$\begin{aligned} & p(e|\neg h_1) - p(e|\neg h_2) \\ = \;& p(e|h_1) \left[\frac{p(h_2)}{1 - p(h_1)} - \frac{p(h_1)}{1 - p(h_2)} \right] + p(e|\neg h_1 \neg h_2)(1 - p(h_1) - p(h_2)) \\ & \cdot \left[\frac{1}{1 - p(h_1)} - \frac{1}{1 - p(h_2)} \right] \\ = \;& p(e|h_1) \frac{p(h_2) - p(h_2)^2 - p(h_1) + p(h_1)^2}{(1 - p(h_1))\,(1 - p(h_2))} + (1 - p(h_1) - p(h_2)) \\ & \cdot \frac{p(e|\neg h_1 \neg h_2) \cdot (p(h_1) - p(h_2))}{(1 - p(h_1))\,(1 - p(h_2))} \\ = \;& p(e|h_1) \frac{(p(h_1) - p(h_2)) \cdot (p(h_1) + p(h_2) - 1)}{(1 - p(h_1))\,(1 - p(h_2))} + (1 - p(h_1) - p(h_2)) \\ & \cdot \frac{p(e|\neg h_1 \neg h_2) \cdot (p(h_1) - p(h_2))}{(1 - p(h_1))\,(1 - p(h_2))} \\ = \;& \frac{1}{(1 - p(h_1))\,(1 - p(h_2))} (p(h_1) - p(h_2)) \cdot (p(h_1) + p(h_2) - 1) \cdot (p(e|h_1) - p(e|\neg h_1 \neg h_2)) \\ < \;& 0 \end{aligned}$$

because e was assumed to be positively relevant to h_1 and h_2, and because the prior of h_1 exceeds the prior of h_2. Hence, the conditions for applying

`Weak Law of Likelihood` are satisfied:

$$f(x,y,z) = c(h_1,e) > c(h_2,e) = f(x,y,z')$$

in contradiction with the inequality $f(x,y,z) \leq f(x,y,z')$ that we got from `Weak Informativity`. □

Proof of Theorem 2: Let us assume that the conditions of `Screened-Off Evidence` are satisfied:

$$\begin{aligned} p(e_2 h) &= p(e_2)p(h) \\ p(e_1 e_2) &= p(e_2)p(e_1) \\ p(e_1 e_2|h) &= p(e_2)p(e_1|h) \end{aligned}$$

By setting $a := p(e_2)$, $x := p(e_1|h)$, $y = p(e_1)$ and $z = p(h)$, we can then derive the general equality

$$f(ax,ay,z) = c(h,e_1e_2) = c(h,e_1) = f(x,y,z) \tag{11}$$

where `Screened-Off Evidence` has been used in the middle equality.

Now we observe that by `Strong Informativity`, there are $x > y$ and $z > z'$ such that

$$f(x,y,z) < f(x,y,z').$$

By an application of (11), we then obtain

$$f(1,y/x,z) < f(1,y/x,z'). \tag{12}$$

Now choose a probability function $p(\cdot)$ such that for sentences $h,e,h' \in \mathfrak{L}$ that satisfy the conditions of `Irrelevant Conjunctions`, $p(h) = z$, $p(h \wedge h') = z'$, $p(e) = y/x$. This implies

$$f(1,y/x,z) \geq f(1,y/x,z'),$$

since $c(h,e) \geq c(h \wedge h',e)$, but it contradicts (12). Hence, the theorem is proven. □

Proof of Theorem 3: Let $e \perp\!\!\!\perp e'|h_0$. From CA2 it follows that $S(h_0, e \wedge e') = f(p(e|h_0) \cdot p(e'|h_0))$, and from CA3 it follows that $S(h_0, e \wedge e') = S(h_0, e) + S(h_0, e') = f(p(e|h_0)) + f(p(e'|h_0))$. This leads to the requirement

$$f(p(e|h_0) \cdot p(e'|h_0)) = f(p(e|h_0)) + f(p(e'|h_0))$$

which is only satisfied by the logarithmic scoring rule $S(h_0, e) = \log_a p(e|h_0)$, for all $a \geq 0$. To see that this uniqueness property holds, remember that the exponential functions are the only continuous functions with the property $g(x + y) = g(x) \cdot g(y)$. They define an isomorphism between the additive group of real numbers and the multiplicative group of postive reals. They are the only functions who do so, and the logarithms are their inverse.

If there were another continuous function f with the property $f(xy) = f(x) + f(y)$, it could not be surjective because in that case, it would have to be a logarithm. Hence, f is not surjective and therefore also bounded (because of continuity). Then adding further summands shows that such a construction cannot work: $f(x_0 \cdot x_1 \cdot x_2 \cdot \ldots) = f(x_0) + f(x_1) + f(x_2) + \ldots$ This shows that f can be raised to an arbitrary value, contradicting boundedness. Hence $S(h_0, e) = \log p(e|h_0)$.

The rest of the proof is straightforward. By CA1 and the above, we obtain

$$\begin{aligned} C_{\mathcal{H}}(h_0, e) &= \sum_{h_i \in \mathcal{H}} \omega_i \cdot (S(h_0, e) - S(h_i, e)) \\ &= \sum_{h_i \in \mathcal{H}} \omega_i \cdot (\log p(e|h_0) - \log p(e|h_i)) \\ &= \sum_{h_i \in \mathcal{H}} \omega_i \cdot \log \frac{p(e|h_0)}{p(e|h_i)} \end{aligned}$$

□

Proof of Theorem 4: As before, CA2′-CA3′ determine that $S(h_0, e) = \log p(e|h_0)$. We will now prove the theorem by considering different forms of f_S in increasing complexity and demonstrate that the form stated in Theorem 4 is indeed the simplest one.

Assume first that $m = 0$ and $n = 1$. In that case, the neutrality condition $f_S(h_0, h_1, e) = 0$ if $p(e|h_0) = p(e|h_1)$ cannot be satisfied unless $c_{00} = 0$ because the numerator is a constant. Hence, we can neglect this possibility.

Now assume that $m = 1$ and $n = 0$. Here, the neutrality condition $f_S(h_0, h_1, e) = 0$ if $p(e|h_0) = p(e|h_1)$ leads to the equation

$$c_{00} + (c_{10} + c_{01})p(e|h_0) + c_{11}p(e|h_0)^2 = 0 \tag{13}$$

which is satisfied in general if and only if $c_{00} = c_{11} = 0$ and $c_{10} = -c_{01}$. Clearly, the resulting function $f(h_0, h_1, e) = p(e|h_0) - p(e|h_1)$ is not ordinally equivalent to $S(h_0, e) - S(h_1, e) = \log p(e|h_0) - \log p(e|h_1)$, regardless of the value of c_{10} and the base of the logarithm. Hence, we can neglect this possibility, too.

Now assume that $m = n = 1$. Again, the neutrality condition leads to the conclusion $c_{00} = c_{11} = 0$ and $c_{10} = -c_{01}$. Now, let us set $p(e|h_0) = 1$, $p(e|h_1) = 0$, and vice versa. Then, the maximality constraint implies $d_{10} = d_{01} = 1$ and the simplest function that maintains ordinal equivalence with $S(h_0, e) - S(h_1 e)$, as demanded by CA1′, is obtained by setting $d_{00} = d_{11} = 0$. □

Proof of Theorem 5: For $C_{\mathcal{H}}$, the calculation is straightforward:

$$\begin{aligned}
\Delta C_{\mathcal{H}}(h_0, h_1, e) &= C_{\mathcal{H}}(h_0, e) - C_{\mathcal{H}}(h_1, e) \\
&= \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e|h_0)}{p(e|h_i)} - \sum_{h_i \in \mathcal{H}} \omega_i \log \frac{p(e|h_1)}{p(e|h_i)} \\
&= \sum_{h_i \in \mathcal{H}} \omega_i \log \left(\frac{p(e|h_0)}{p(e|h_i)} \cdot \frac{p(e|h_i)}{p(e|h_1)} \right) \\
&= \left(\sum_{h_i \in \mathcal{H}} \omega_i \right) \log \frac{p(e|h_0)}{p(e|h_1)} \\
&= \log \frac{p(e|h_0)}{p(e|h_1)}
\end{aligned}$$

For $C'_{\mathcal{H}'}$, we have to work a bit harder:

$$\begin{aligned}
& C'_{\mathcal{H}}(h_0, e) - C'_{\mathcal{H}}(h_1, e) \\
= \; & \sum_{h_i \in \mathcal{H}} \omega_i \frac{p(e|h_0) - p(e|h_i)}{p(e|h_0) + p(e|h_i)} - \sum_{h_i \in \mathcal{H}} \omega_i \frac{p(e|h_1) - p(e|h_i)}{p(e|h_1) + p(e|h_i)} \\
= \; & (\omega_0 + \omega_1) \frac{p(e|h_0) - p(e|h_1)}{p(e|h_0) + p(e|h_1)} + \sum_{h_i \in \mathcal{H} \setminus \{h_0, h_1\}} \omega_i \\
& \cdot \frac{(p(e|h_0) - p(e|h_i))\,(p(e|h_1) + p(e|h_i)) - (p(e|h_0) + p(e|h_i))\,(p(e|h_1) - p(e|h_i))}{(p(e|h_0) + p(e|h_i))\,(p(e|h_1) + p(e|h_i))} \\
= \; & (\omega_0 + \omega_1) \frac{p(e|h_0) - p(e|h_1)}{p(e|h_0) + p(e|h_1)} + \sum_{h_i \in \mathcal{H} \setminus \{h_0, h_1\}} 2\omega_i \, \frac{p(e|h_0) p(e|h_i) - p(e|h_1) p(e|h_i)}{(p(e|h_0) + p(e|h_i))\,(p(e|h_1) + p(e|h_i))} \\
= \; & (p(e|h_0) - p(e|h_1)) \left[\frac{\omega_0 + \omega_1}{p(e|h_0) + p(e|h_1)} + \sum_{h_i \in \mathcal{H} \setminus \{h_0, h_1\}} \frac{2\omega_i \, p(e|h_i)}{(p(e|h_0) + p(e|h_i))\,(p(e|h_1) + p(e|h_i))} \right] \\
= \; & (p(e|h_0) - p(e|h_1)) \sum_{h_i \in \mathcal{H}} 2\omega_i \, \frac{p(e|h_i)}{(p(e|h_0) + p(e|h_i))\,(p(e|h_1) + p(e|h_i))}
\end{aligned}$$

□

References

Bernardo, J.M. (2012): "Integrated objective Bayesian estimation and hypothesis testing", in J.M. Bernardo et al. (eds.): *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*, 1–68 (with discussion). Oxford: Oxford University Press.

Bovens, L., and S. Hartmann (2003): *Bayesian Epistemology*. Oxford: Oxford University Press.

Brössel, P. (2013): "The Problem of Measure Sensitivity Redux", *Philosophy of Science* 80, 378–397.

Brössel, P. (2014): "On the Role of Explanatory and Systematic Power in Scientific Reasoning". Forthcoming in *Synthese*.

Carnap, R. (1950): *Logical Foundations of Probability*. Chicago: The University of Chicago Press.

Chase, W., and F. Brown (2000): *General Statistics*. New York: Wiley.

Cohen, J. (1994): "The Earth is Round ($p < .05$)", *American Psychologist* 49, 997–1001.

Crupi, V. (2014): "Confirmation", in: E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/entries/confirmation/`, retrieved on June 14, 2014.

Crupi, V., and K. Tentori (2014): "Confirmation theory", in: A. Hájek and C. Hitchcock (eds.), *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press (forthcoming).

Crupi, V., Chater, N., and K. Tentori (2013): "New Axioms for Probability and Likelihood Ratio Measures", *The British Journal for the Philosophy of Science* 64, 189–204.

Crupi, V., Tentori, K., and M. González (2007): "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues", *Philosophy of Science* 74, 229–252.

Díez, J. (2011): "On Popper's strong inductivism (or strongly inconsistent anti-inductivism)", *Studies in the History and Philosophy of Science A* 42, 105–116.

Edwards, A.W.F. (1972): *Likelihood*. Cambridge: Cambridge University Press.

Eells, E., and B. Fitelson (2002): "Symmetries and Asymmetries in Evidential Support", *Philosophical Studies* 107, 129–142.

Fidler, F. (2013): *From Statistical Significance to Effect Estimation: Statistical Reform in Psychology, Medicine and Ecology*. London: Routledge.

Fisher, R.A. (1956): *Statistical Methods and Scientific Inference*. New York: Hafner.

Fitelson, B. (2001): *Studies in Bayesian Confirmation Theory*. PhD thesis, University of Wisconsin/Madison.

Fitelson, B. (2002): "Putting the Irrelevance Back Into the Problem of Irrelevant Conjunction", *Philosophy of Science* 69, 611–622.

Frigg, R., and S. Hartmann (2006): "Models in Science", in: E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/entries/models-science/`, retrieved on June 14, 2014.

Gaifman, H., and M. Snir (1982): "Probabilities Over Rich Languages, Testing and Randomness", *Journal of Symbolic Logic* 47, 495–548.

Gelman, A., and C. Shalizi (2012): "Philosophy and the practice of Bayesian statistics in the social sciences", in: Harold Kincaid (ed.), *Oxford Handbook of the Philosophy of the Social Sciences*, 259–273. Oxford: Oxford University Press.

Gelman, A., and C. Shalizi (2013): "Philosophy and the practice of Bayesian statistics (with discussion)", *British Journal of Mathematical and Statistical Psychology* 66, 8–18.

Gemes, Ken (1998): "Hypothetico-Deductivism:The Current State of Play", *Erkenntnis* 49, 1–20.

Gillies, D. (1971): "A Falsifying Rule for Probability Statements", *British Journal for the Philosophy of Science* 22, 231–261.

Good, I.J (1952): "Rational Decisions", *Journal of the Royal Statistical Society B* 14, 107–114.

Good, I.J. (1960): "Weight of Evidence, Corroboration, Explanatory Power, Information and the Utility of Experiments", *Journal of the Royal Statistical Society B* 22, 319–331.

Good, I.J. (1968): "Corroboration, Explanation, Evolving Probability, Simplicity and a Sharpened Razor", *The British Journal for the Philosophy of Science* 19, 123–143.

Good, I.J. (1982): "A Good Explanation of an Event Is Not Necessarily Corroborated by the Event", *Philosophy of Science* 49, 251–253.

Good, I.J. (1983/2009): *Good Thinking: The Foundations of Probability and Its Applications*. Minneapolis: University of Minnesota Press. Reprint 2009. New York: Dover.

Hempel, C.G. (1960): "Inductive inconsistencies", *Synthese* 12, 439–469.

Hawthorne, J., and B. Fitelson (2004): "Re-solving Irrelevant Conjunction with Probabilistic Independence", *Philosophy of Science* 71, 505-514.

Howson, C. and P. Urbach (2006): *Scientific Reasoning: The Bayesian Approach*. Third Edition. La Salle: Open Court.

Huber, F. (2008): "Hempel's Logic of Confirmation", *Philosophical Studies* 139, 181–189.

Joyce, J. (2008): "Bayes' Theorem", in: E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. `http://plato.stanford.edu/entries/confirmation/`, retrieved on June 14, 2014.

Kass, R. and A. Raftery (1995): "Bayes Factors", *Journal of the American Statistical Association* 90, 773–790.

Kemeny, J.G., and P. Oppenheim (1952): "Degrees of factual support", *Philosophy of Science* 19, 307–324.

Lele, S. (2004): "Evidence Functions and the Optimality of the Law of Likelihood (with discussion)", in: M. Taper and S. Lele (eds.), *The Nature of Scientific Evidence*, 191–216. Chicago & London: The University of Chicago Press.

Milne, P. (1996): "log[p(h/eb)/p(h/b)] is the One True Measure of Confirmation", *Philosophy of Science* 63, 21–26.

Neyman, J., and E. Pearson (1933): "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society A* 231, 289–337.

Popper, K.R. (1934/2002): *Logik der Forschung*. Berlin: Akademie Verlag. Translated as *The Logic of Scientific Discovery*. Routledge: London.

Popper, K.R. (1954): "Degree of Confirmation", *The British Journal for the Philosophy of Science* 5, 143–149.

Popper, K.R. (1956/83): *Realism and the Aim of Science*. Totowa/NJ: Rowman and Littlefield.

Popper, K.R. (1957): "A Second Note on Degree of Confirmation", *The British Journal for the Philosophy of Science* 7, 350–353.

Popper, K.R. (1958): "A third note on degree of corroboration or confirmation", *The British Journal for the Philosophy of Science* 8, 294–302.

Popper, K.R. (1963): *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Harper.

Popper, K.R. (1979): *Objective knowledge: an evolutionary approach*. Oxford: Clarendon Press.

Rowbottom, D.P. (2011): *Popper's Critical Rationalism: A Philosophical Investigation*. London: Routledge.

Rowbottom, D.P. (2013): "Popper's Measure of Corroboration and P(h | b)", *The British Journal for the Philosophy of Science* 64, 739–745.

Royall, R. (1997): *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Schupbach, J., and J. Sprenger (2011): "The Logic of Explanatory Power", *Philosophy of Science* 78, 105–127.

Sober, E. (2008): *Evidence and Evolution*. Cambridge: Cambridge University Press.

Sprenger, J. (2009): "Evidence and Experimental Design in Sequential Trials", *Philosophy of Science* 76, 637–649.

Sprenger, J. (2014): "Bayesianism vs. Frequentism in Statistical Inference", in: A. Hájek and C. Hitchcock (eds.), *Handbook of the Philosophy of Probability*. Oxford: Oxford University Press (forthcoming).

Wasserman, L. (2004): *All of Statistics*. New York: Springer.

Williamson, J. (2010): *In defense of objective Bayesianism*. Oxford: Oxford University Press.

Ziliak, S.T., and D.N. McCloskey (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor: University of Michigan Press.