# Degrees of Corroboration: An Antidote to the Replication Crisis*

Jan Sprenger†

May 24, 2019

**Abstract**

The replication crisis poses an enormous challenge to the epistemic authority of science and the logic of statistical inference in particular. Two prominent features of Null Hypothesis Significance Testing (NHST) arguably contribute to the crisis: the lack of guidance for interpreting non-significant results and the impossibility of quantifying support for the null hypothesis. In this paper, I argue that also popular alternatives to NHST, such as confidence intervals and Bayesian inference, do not lead to a satisfactory logic of evaluating hypothesis tests. As an alternative, I motivate and explicate the concept of corroboration of the null hypothesis. Finally I show how degrees of corroboration give an interpretation to non-significant results, combat publication bias and mitigate the replication crisis.

*I would like to thank the audience at PSA2018 in Seattle and various other places for constructive discussion, and Mattia Andreoletti and Felipe Romero for detailed feedback on the full paper. The research was supported by the European Research Council (ERC) through Starting Investigator Grant No. 640638 "Making Scientific Inferences More Objective".

†Contact information: Center for Logic, Language and Cognition (LLC), Department of Philosophy and Educational Science, Università degli Studi di Torino, Via Sant'Ottavio 20, 10124 Torino, Italy. Email: jan.sprenger@unito.it. Webpage: www.laeuferpaar.de.

# 1 Introduction

Various scientific disciplines are currently undergoing a **replication crisis**: researchers struggle to reproduce the results of previous experiments when copying the original experimental design. Studies that try to assess the extent of the crisis in a systematic way lead to sobering outcomes: the rate of statistically significant findings drops dramatically and the observed effect sizes are often much lower (for the fields of psychology, experimental economics and cancer biology, respectively: Open Science Collaboration 2015; Camerer et al. 2016; Nosek and Errington 2017).

The causes of the replication crisis are a subject of vivid debate. Popular explanations cite adverse effects of social and structural factors in academia, such as the pressure to publish novel and ground-breaking results, or the presence of questionable research practices (e.g., Bakker, Wicherts and van Dijk 2012; Romero 2017). According to these authors, we need to **change the incentive structure of the scientific enterprise**. Other explanations relate the replication crisis to methodological shortcomings in prevalent methods of statistical inference, in particular Null Hypothesis Significance Testing (NHST). The well-known criticisms of NHST gain renewed importance in the context of the replication crisis: the focus on statistically significant results leads to exaggerated effect size reports and suppresses statistically non-significant, but scientifically valuable findings (e.g., Cohen 1994; Goodman 1999a; Ioannidis 2005; Ziliak and McCloskey 2008). In this way, NHST promotes publication bias and contributes to a higher rate of replication failures. In response, many authors suggest **statistical reforms** such as lowering the statistical significance threshold (Benjamin et al. 2018), replacing NHST by Bayesian inference (Goodman 1999b; Lee and Wagenmakers 2013) or abandoning hypothesis testing altogether and replacing it by inference with confidence intervals (Cumming 2012).

This paper contributes to the statistical reform project, but it draws less radical conclusions. It retains that hypothesis tests are an essential tool for scientific inference which cannot be fully replaced by the estimation-centered perspective of confidence intervals. It demonstrates that Bayesian inference, though in many ways improving upon NHST, is limited in the set of inferential questions it can ask. In particular, many hypothesis tests in science should be conceptualized as *asymmetric* tests and Bayesian inference captures this idea only in a limited way. Finally it develops a method of hypothesis testing, based on the concept of **degree of corroboration**, that is able to quantify support for the null hypothesis and combats publication bias and the

replication crisis from within the (broadly) frequentist paradigm.

The paper is structured as follows. Section 2 shows why NHST cannot quantify evidence in favor of the null hypothesis and how this feature promotes publication bias. Section 3 expounds alternatives to NHST—confidence intervals, equivalence testing and Bayesian statistics—and explains their limitations. Section 4 motivates and explicates the concept of degree of corroboration as a way of interpreting non-significant test results and stating support for the null hypothesis. The final Section 5 explains how this proposal mitigates the replication crisis.

## 2 NHST and Support for the Null Hypothesis

NHST is based on severely testing a precise hypothesis $H_0$—the "null" or default hypothesis. Usually it denotes the absence of an effect in an experimental manipulation: for example, a medical drug is no better than a placebo treatment. The idea is that before inferring to the presence of an effect, and taking it for granted in decisions we make (e.g., patient treatment), we must have found strong evidence against the default hypothesis that such an effect is absent. NHST is applied across all domains of science, but is especially prominent in psychology and medicine.

Evidence against the null hypothesis is usually expressed by means of ***p*-values**. They quantify the probability of obtaining a result that diverges from the null hypothesis at least as much as the actual data.[1] Generally, the $p$-value is the smaller, the more the actual result diverges from the hypothesized null value.

Conventionally, $p$-values smaller than .05 are classified as "statistically significant evidence" against the null hypothesis, $p$-values smaller than .01 as "highly significant evidence", and so on. Since the null hypothesis could explain such low observed $p$-values only by reference to a very unlikely event, and since chance is no good explanation in scientific reasoning, a statistically significant $p$-value counts as evidence for the alternative hypothesis $H_1$—the presence of a (causal) effect.

Explicit evidence *for* the null hypothesis $H_0$ is, however, impossible to obtain since the logic of NHST is based on the idea of falsification and disconfirmation:

[1] A concrete example: suppose that our parameter of interest is the unknown mean $\mu$ of a population, which the null hypothesis claims to take the value $\mu_0$. For a divergence-measuring function such as $z(\bar{X}) = (\bar{X} - \mu_0)/s$ with sample mean $\bar{X}$ and estimated standard deviation $s$, the $p$-value is calculated as $p = p_{H_0}(|z(\bar{X})| \geq |z(\bar{x})|)$. It states the probability that given $H_0$, the sample mean is as least as far away from $\mu_0$ as the actually observed mean $\bar{x}$. See Romeijn 2014 for a detailed review.

> "[...] it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." (Fisher 1935/74, 16)

In particular, $p$-values greater than .05 do not have an evidential interpretation, or license a judgment of support in favor of the null hypothesis—even when they are close to unity. Current statistics textbooks and encyclopedias mirror this attitude when they classify $p$-values greater than .05 by phrases such as "little or no evidence against $H_0$" (Wasserman 2004) or "insufficient to support a conclusion" (Wikipedia).

This approach neglects that null hypotheses are, due to their precision and high testability, often of significant theoretical and practical importance (Gallistel 2009). They may express the independence of two factors in a causal mechanism, postulate that performance difference between two groups is due to chance, or claim that a generic medical drug is equally effective as the originally patented drug. Due to their salience in theoretical inference and decision-making, it is imperative that a framework of statistical analysis be able to quantify evidence in their favor.

Since NHST cannot express evidence in favor of the null, it contributes to publication bias: when only statistically significant results with $p$-values smaller than .05 are evidentially interpretable, it is not easy to relate bigger $p$-values to general scientific conclusions and to package the experimental results into a convincing narrative. Such experiments are therefore more likely to end up in the proverbial file drawer (Rosenthal 1979) and not to be shared with the scientific community, although their methodological quality is not inferior to experiments with statistically significant outcomes. This effect is reinforced by the common tendency to focus attention on statistically significant outcomes, and to identify them with scientifically relevant findings (Cohen 1994; Ziliak and McCloskey 2008). As a consequences, effect sizes in published research tend to be inflated and many null hypotheses are underestimated with respect to their empirical support. Also the converse may happen: in experiments with small sample sizes (e.g., a rare medical condition), a judgment of non-significance may conceal a modest, but scientifically meaningful effect (Aczel et al. 2018).

Either way, we need a principled method of determining when and how non-significant results translate into support for the null hypothesis. Purely methodological reform proposals, such as compulsory data sharing and pre-registration of experiments, do not answer this question—we need a proper statistical reform, too.

# 3 Solution Proposals: Confidence Intervals, Equivalence Testing and Bayesian Inference

Before presenting my own approach, I would like to briefly discuss three well-known statistical reform proposals. While the discussion has to remain at the surface, it will help us to identify criteria for a satisfactory solution.

First, we may decide to give up hypothesis testing altogether and to evaluate experimental findings by means of **confidence intervals**. By replacing $p$-values with interval estimates, we get rid of the concept of statistical significance and the bias toward suppressing non-significant results (Cumming 2012).

This proposal has two major problems: an epistemic and a methodological one. The epistemic problem is that a 95% confidence interval cannot be interpreted as a range of plausible or probable values of the unknown parameter $\mu$. Rather it indicates those parameter values $\mu'$ where, if $\mu = \mu'$ were adopted as a null hypothesis, the actually observed data would lead to a non-significant outcome ($p > .05$). Thus, confidence intervals rely on the logic of NHST and share their limitations, too. In particular, they cannot provide a judgment of evidential support for the null hypothesis.

The methodological problem is that pure estimation techniques struggle to address crucial statistical inference problems in science, such as contrasting models with different parameter spaces (e.g., a linear vs. a quadratic model) or theories with different ontological assumptions (e.g., the existence of the Higgs Boson). Hypothesis tests are indispensable tools for such inferences. Interval estimators are useful in a variety of contexts, but they cannot replace the search for a proper evidential interpretation of hypothesis tests.

Second, we could try to quantify support for the null by means of **equivalence testing** (Lakens, Scheel and Isager 2018): conducting two one-sided hypothesis tests based on the minimal scientifically meaningful effect size $\varepsilon > 0$. If we reject the two novel null hypotheses $H_0' : \mu \geq \mu_0 + \varepsilon$ and $H_0'' : \mu \leq \mu_0 - \varepsilon$, we have found evidence for the original null hypothesis of no effect $H_0 : \mu = \mu_0$. However, equivalence testing does not (yet) contain a quantitative dimension: it is not clear how the $p$-values from two one-sided tests should be aggregated into an overall judgment of evidential support for the null.

The third and most discussed solution proposal—switching to **Bayesian hypothesis testing**—possesses the required conceptual vocabulary. Evidence for or against the null hypothesis $H_0$ is quantified by means of the **(logarithmic) Bayes factor**, that is,

the degree to which $H_0$ and $H_1$ can account for the observed data E:

$$\log \mathrm{BF}_{01}(\mathrm{E}) = \log \frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_1)}. \qquad \text{(Log-Bayes Factor)}$$

This quantity is positive if and only if the null hypothesis explains the data better than the alternative. Unlike in the case of $p$-values, the evidence can be quantified *in both directions*, from strong evidence for the alternative hypothesis to strong evidence for the null. See Table 1 for a conventional interpretation of Bayes factors based on the natural logarithm.

| Range of $\log \mathrm{BF}_{01}$ | | |
|---|---|---|
| Support for $H_0$ | Support for $H_1$ | Interpretation |
| $>5$ | $< -5$ | Very strong evidence for $H_0$/ $H_1$ |
| 3 to 5 | -3 to -5 | Strong evidence for $H_0$/ $H_1$ |
| 1 to 3 | -1 to -3 | Moderate evidence for $H_0$/ $H_1$ |
| 0 to 1 | 0 to -1 | Not worth more than a bare mention |
| 0 | 0 | No evidence for either hypothesis |

Table 1: Classification of log-Bayes factors adapted from Kass and Raftery (1995, 777).

However, this symmetric conceptualization of hypothesis tests does not square well with the asymmetric rationale of NHST. The Bayes factor represents an alternative hypothesis such as $H_1 : \mu \neq \mu_0$ as the *average* of the performance of all (precise) alternatives to $H_0 : \mu = \mu_0$, weighted by the prior distribution of effect size. Thus, its performance can be distorted severely by poor performance of extreme alternatives (e.g., hypotheses that postulate huge effect sizes). If these hypotheses are "heavy enough"—that is, if the prior distribution is sufficiently spread out—, the null hypothesis will be supported even if there is a theoretically meaningful hypothesis of small effect size that explains the data better. It seems mistaken to classify such cases as support for the null hypothesis. Rather, the null hypothesis has been outperformed by a relevant competitor and not stood up to the test. While Bayes factors are a sound measure of evidence in symmetric hypothesis tests, they do not capture typical scientific reasoning in asymmetric testing problems (i.e., with more than one relevant alternative).

## 4 Degrees of Corroboration

As explained above, a null hypothesis often competes against various alternative hypotheses: for example, when a linear model is contrasted to a quadratic or cubic model. Moreover, the null hypothesis is usually not considered a candidate for literal truth. After all, most experimental manipulations will have *some* minuscule (positive or negative) effect, though it may be practically meaningless. For example, we do not consider a diet cure effective if it reduces weight on average by half a kilo. Rather, NHST aims at finding out whether $H_0$ is a *good proxy* for the general statistical model (e.g., all possible hypotheses about the effect of the diet). In other words, the question is whether the precise null hypothesis is a reasonable idealization of the general statistical model that comprises $H_0 : \mu = \mu_0$ as well as $H_1 : \mu \neq \mu_0$.

An answer is given by the concept of **corroboration**, introduced by Karl R. Popper (1979, 818): "a concise report evaluating the state [...] of the critical discussion of a theory. Corroboration [...] is thus an evaluating report of past performance. Like preference, it is essentially comparative." This concept, originally developed for testing deterministic hypotheses, can be transferred to statistical reasoning: it expresses whether the hypothesis in question has survived tests against all relevant competitors, and whether it can stand in as a proxy for a complex statistical model. In other words, corroboration judgments indicate whether the point null hypothesis strikes a good tradeoff between precision/testability and empirical fit (see also Popper 1959/2002, ch. 10; Rowbottom 2011; Sprenger 2018). We now explicate Popper's informal description in the context of statistical inference.

First, we do not conceptualize the alternative hypothesis as the negation of the null hypothesis, but as a *partition* $\mathcal{H}$, that is, a grouping of the parameter space into mutually disjoint subsets. Typically, the elements of these partition are effect size ranges that correspond to different, scientifically meaningful alternatives (see also Good 1968). Suppose we are interested in whether frequent chessplaying improves student exam performance. The null hypothesis states that chessplaying does not affect student performance. Corroborating such a null hypothesis would be both interesting and surprising since the skills trained by studying chess (e.g., mental visualization, logical analysis, endurance) are plausibly useful in school exams, too. We partition the alternative hypotheses into effect size ranges that correspond to different practical conclusions: collecting further data (0–10% difference), funding a research project on brain patterns activated by chessplaying (10–20% difference), or taking chess educa-

tion into school curricula (>20% difference). Also for other research problems, such as studying the effects of a medical drug, we can easily imagine how different effect sizes correspond to different treatment and prescription policies. The choice of the partition of alternative hypotheses is thus context-sensitive and depends on the relevant research problem.

In agreement with Popper's rationale, the null counts as corroborated if and only if it has survived a test against all relevant competing hypotheses. Thus I define corroboration as the weight of evidence in favor of $\mathrm{H}_0$ with respect to the best-performing alternative in $\mathcal{H}$, or in other words, as the minimal weight of evidence in favor of $\mathrm{H}_0$:

**CA1: Corroboration = Weight of Evidence** For a null hypothesis $H_0$ and a (possibly infinite) partition of alternative hypotheses $\mathcal{H} = \{H_1, H_2, \ldots\}$, the degree of corroboration that observation E provides for $\mathrm{H}_0$ relative to $\mathcal{H}$ is defined as

$$\zeta_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) \;=\; \min_{\mathrm{H_i} \in \mathcal{H}} \omega(\mathrm{H}_0, \mathrm{H_i}, \mathrm{E}), \tag{1}$$

where $\omega(\mathrm{H}_0, \mathrm{H}_i, \mathrm{E})$ quantifies the weight of evidence that E provides for $\mathrm{H}_0$ and against the specific alternative $\mathrm{H}_i$.

CA1 leaves open what the weight-of-evidence function looks like. In agreement with widespread explications of evidential favoring (e.g., Good 1950; Sober 2008), I propose that this function should only depend on the predictive performance of the competing hypotheses, as measured by their likelihoods on the observed data E.

**CA2: Predictive Performance** For hypotheses $\mathrm{H}_0$ and $\mathrm{H}_1$ and observation E, there is a continuous function $g\colon [0;1]^2 \to \mathbb{R}$, increasing in the first and decreasing in the second argument, such that

$$\omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) \;=\; g\big(p(\mathrm{E}|\mathrm{H}_0), p(\mathrm{E}|\mathrm{H}_1)\big) \tag{2}$$

Finally, I demand that weight of evidence be additive with respect to independent and identically distributed (i.i.d.) observations. This requirement allows us to aggregate evidence from various experiments with identical design in a convenient manner. Moreover, irrelevant observations should not change the overall weight of evidence.

**CA3: Independent and Irrelevant Evidence** If for two observation E and E′ and any $\mathrm{H} \in \{\mathrm{H}_0, \mathrm{H}_1\}$, $p(\mathrm{E} \wedge \mathrm{E}'|\mathrm{H}) = p(\mathrm{E}|\mathrm{H}) \times p(\mathrm{E}'|\mathrm{H})$, then

$$\omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E} \wedge \mathrm{E}') \;=\; \omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) + \omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}') \tag{3}$$

Specifically, if it is also the case that $p(\mathrm{E}'|\mathrm{H}_0) = p(\mathrm{E}'|\mathrm{H}_1) > 0$, then $\omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E}) = \omega(\mathrm{H}_0, \mathrm{H}_1, \mathrm{E} \wedge \mathrm{E}')$.

These three constrains, each natural and plausible, jointly determine a measure of corroboration that is unique up to scaling properties:

**Theorem 1.** *CA1, CA2 and CA3 uniquely determine the corroboration measure*

$$\zeta_{\mathcal{H}}(\mathrm{H}_0, \mathrm{E}) = k \min_{\mathrm{H}_i \in \mathcal{H}} \log \frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_i)} \qquad \textit{for } k > 0. \tag{4}$$

The base of the logarithm and the scalar $k$ may be chosen *ad libitum*; but in order to keep the scale consistent with logarithmic Bayes factors I suggest the natural logarithm and $k = 1$. Positive degree of corroboration entails that $\mathrm{H}_0$ outperforms all relevant alternatives. It has thus survived the test with a certain degree of positive support, in line with Popper's informal characterization. In particular, degree of corroboration is measured along the same scale as Bayes factors and so we can use Table 1 for interpreting the evidential support in favor of $\mathrm{H}_0$.

We now return to our chessplaying example. Simplifying a bit, we adopt a Binomial with null hypothesis $H_0 : \mu = .6$ (=60% base rate of students who achieve all learning objectives) and $N = 200$ chessplaying students in the dataset. To show the dependence of $\zeta_{\mathcal{H}}$ on the choice of the partition $\mathcal{H}$, we consider three different partitions of the space of alternative hypotheses: (1) a maximal partition $\mathcal{H}_{\text{max}} = [0; 1]$ that treats every parameter value as an alternative in its own right, (2) a medium-sized partition $\mathcal{H}_{\text{med}} = \{\ldots, (0.6; 0.7], (0.7; 0.8], (0.8; 0.9], \ldots\}$ where the alternatives are effect size ranges corresponding to different practical conclusions (see the motivation of the example); (3) a minimal partition $\mathcal{H}_{\text{min}} = \{[0; 1]\}$ with only one alternative hypothesis, that is, the weighted average of all values of $\mu$. The partition $\mathcal{H}_{\text{min}}$ thus leads to a Bayesian hypothesis test.[2]

Figure 1 plots the degree of corroboration of the null hypothesis for different possible results and partitions. Corroboration strongly depends on the partition: The minimal partition $\mathcal{H}_{\text{min}}$ equates degree of corroboration to logarithmic Bayes factors and concludes support for the null hypothesis in the entire non-significant range. The medium-sized partition $\mathcal{H}_{\text{med}}$ motivated by considerations of effect size relevance has the cutoff already at a 63,5% success rate. The maximal partition $\mathcal{H}_{\text{max}}$ states positive corroboration only if exactly a 60% success rate is observed. This dependence of

[2]We use a logistic weighting function for calculating $p(\mathrm{E}|\mathrm{H}_1)$ when $\mathrm{H}_1$ is an effect size range rather than a point hypothesis, but this choice barely matters for the outcome.

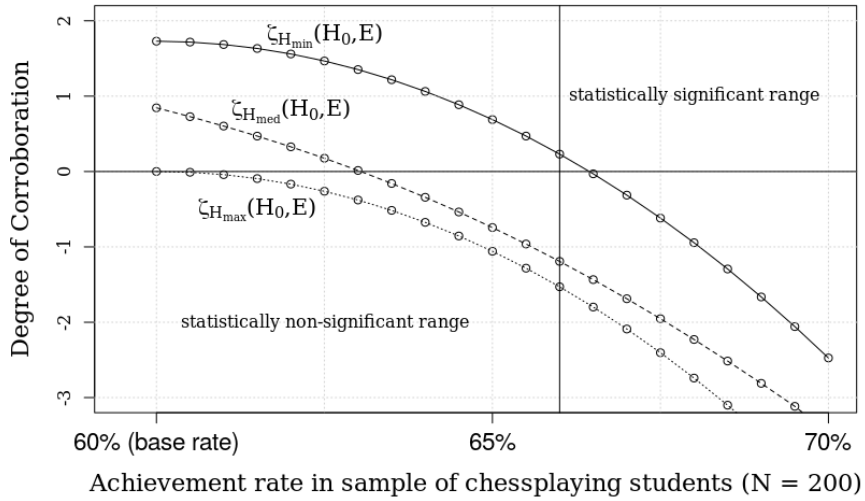


Figure 1: Degree of corroboration of the null hypothesis $H_0$, as a function of the number of chessplaying students who achieve all learning objectives ($N = 200$, base rate: 60%). Solid line on top: partition $\mathcal{H}_{\text{min}}$, dashed line in the middle: $\mathcal{H}_{\text{med}}$, dotted line at bottom: $\mathcal{H}_{\text{max}}$. The vertical line at $x = 66\%$ demarcates the statistically significant and non-significant range.

corroboration on the conceptualization of the relevant alternatives is arguably an advantage of the proposed approach: it underlines that context and practically relevant effect sizes are essential elements of designing and interpreting hypothesis tests. In this way, the proposed approach also vindicates well-known criticisms of NHST.

## 5 Discussion: Back to the Replication Crisis

How do degrees of corroboration help to address the replication crisis? The most obvious improvement over the classical NHST perspective and $p$-values is that they **quantify support for the null hypothesis** in a meaningful way: conceptually similar to Bayes factors and likelihood ratios, but without giving up the idea of trying to falsify a precise default hypothesis against the background of various scientifically relevant alternatives. Integrating degrees of corroboration into the asymmetric rationale of NHST is a relatively minor methodological change compared to a full-scale switch to Bayesian reasoning, increasing the chances of this proposal to be adopted in scientific practice.

In particular, degrees of corroboration enable researchers to give an evidential interpretation to non-significant test results, to formulate a convincing narrative in terms of positive support for the null hypothesis, and to disseminate such findings in the scientific community. By doing so, the corroboration measure acts as an **antidote to the file drawer effect and publication bias** (and thus, the replication crisis). In particular, it familiarizes researchers with the idea that support for null hypotheses can be expressed in a positive manner, rather than just stating "failure to reject the null hypothesis". This will help them to overcome the pernicious and unfortunately still widespread idea that only a study with significant results *against* the null hypothesis counts as a successful experiment.

Finally, the corroboration-based approach forces a researcher to make up her mind about theoretically meaningful alternatives to the null hypothesis at the stage of experimental design—that is, *before* the experiment is conducted. This specification is helpful in various ways. First, it leads to **predictively more powerful and testable alternative hypothesis** than just considering the negation of the null hypothesis. Second, specificying alternatives beforehand **disentangles statistical and scientific significance**: a statistically significant result against the null hypothesis is scientifically significant only if the null hypothesis fails to be corroborated and there is evidence for a specific competitor. Third, it is much more explicit than NHST with respect to **powering experiments adequately**. Specifying a precise set of alternatives allows researchers to calculate the chances of misleading evidence—that is, the probability of corroborating the null when it is substantially mistaken—and to calibrate the experimental design accordingly (see also Royall 2000; Schönbrodt and Wagenmakers 2018). In this way, corroboration-based hypothesis testing combats underpowered experiments as one of the causes of the replication crisis (e.g., Ioannidis 2005; Szucs and Ioannnidis 2017).

All in all, the concept of degree of corroboration has great innovative and unifying potential in statistical inference. It cures methodological shortcomings of NHST that have contributed to the replication crisis, without giving up hypothesis tests as a cornerstone of scientific method. It generalizes Bayesian inference to asymmetric testing contexts while keeping Bayes factors as a special case of degrees of corroboration. While more needs to be said, all this should suffice to establish this proposal an attractive middle ground between sticking to tradition and calls for radical statistical reform.

# A Appendix: Proof of Theorem 1

Suppose the conditions of CA3 are satisfied for observations E, E′ and E″, and in particular, $p(\mathrm{E}'|\mathrm{H}_0) = p(\mathrm{E}'|\mathrm{H}_1)$. Then we infer, using CA2, that

$$g\big(p(\mathrm{E}\wedge\mathrm{E}'|\mathrm{H}_0),\, p(\mathrm{E}\wedge\mathrm{E}'|\mathrm{H}_1)\big) \;=\; g\big(p(\mathrm{E}|\mathrm{H}_0)\,p(\mathrm{E}'|\mathrm{H}_0),\, p(\mathrm{E}|\mathrm{H}_1)\,p(\mathrm{E}'|\mathrm{H}_0)\big)$$

and moreover

$$g\big(p(\mathrm{E}|\mathrm{H}_0),\, p(\mathrm{E}|\mathrm{H}_1)\big) \;=\; g\big(p(\mathrm{E}\wedge\mathrm{E}'|\mathrm{H}_0),\, p(\mathrm{E}\wedge\mathrm{E}'|\mathrm{H}_1)\big)$$

Combining these two equations, we infer that the function $g$ satisfies the general equality $g(a,b) = g(ax,bx)$ for $x \in (0,1]$ and therefore, also $g(a,b) = g(a/b,1)$ for $a < b$. In other words, $\omega(\mathrm{H}_0,\mathrm{H}_1,\mathrm{E})$ depends on the ratio of $p(\mathrm{E}|\mathrm{H}_0)$ and $p(\mathrm{E}|\mathrm{H}_1)$ only; in the remainder of the proof we call this function $f$. We then obtain

$$\omega(\mathrm{H}_0,\mathrm{H}_1,\mathrm{E}\wedge\mathrm{E}'') \;=\; f\left(\frac{p(\mathrm{E}\wedge\mathrm{E}''|\mathrm{H}_0)}{p(\mathrm{E}\wedge\mathrm{E}''|\mathrm{H}_1)}\right) \;=\; f\left(\frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_1)}\cdot\frac{p(\mathrm{E}''|\mathrm{H}_0)}{p(\mathrm{E}''|\mathrm{H}_1)}\right) \tag{5}$$

Moreover, we know that $\omega(\mathrm{H}_0,\mathrm{H}_1,\mathrm{E}) = f\left(\frac{p(\mathrm{E}|\mathrm{H}_0)}{p(\mathrm{E}|\mathrm{H}_1)}\right)$ and $\omega(\mathrm{H}_0,\mathrm{H}_1,\mathrm{E}'') = f\left(\frac{p(\mathrm{E}''|\mathrm{H}_0)}{p(\mathrm{E}''|\mathrm{H}_1)}\right)$. Combining these equations with Equations (3) and (5) using the variables $x := p(\mathrm{E}|\mathrm{H}_0)/p(\mathrm{E}|\mathrm{H}_1)$ and $y := p(\mathrm{E}''|\mathrm{H}_0)/p(\mathrm{E}''|\mathrm{H}_1)$, we can then derive the general equality $f(x\cdot y) = f(x) + f(y)$. This equality is only satisfied by functions of the form $f(x) = k\log_a x$. The rest of the proof follows by plugging in the weight-of-evidence function into CA1 which immediately yields Equation (4). □

# References


Aczel, Balazs, Bence Palfi, Aba Szollosi, et al. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science 1*, 357–366.

Bakker, Marjan, Jelte Wicherts, and Annette van Dijk (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science 7*, 543–554.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, and et al. (2018). Redefine statistical significance. *Nature Human Behaviour 2*, 6–10.

Camerer, Colin F., Anna Dreber, Eskil Forsell, et al. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*. https://doi.org/10.1126/science.aaf0918.

Cohen, Jacob (1994). The Earth is Round ($p < .05$). *Psychological Review 49*, 997–1001.

Cumming, Geoff (2012). *Understanding the New Statistics*. New York: Routledge.

Fisher, R. A. (1935/74). *The Design of Experiments*. New York: Hafner Press. Reprint of the ninth edition from 1971. Originally published in 1935 (Edinburgh: Oliver & Boyd).

Gallistel, C. R. (2009). The Importance of Proving the Null. *Psychological Review 116*, 439–453.

Good, I. J. (1950). *Probability and the Weighting of Evidence*. London: Griffin.

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

Goodman, Stephen N. (1999a). Toward Evidence-Based Medical Statistics 1: The *P* value Fallacy. *Annals of Internal Medicine 130*, 995–1004.

Goodman, Stephen N. (1999b). Toward Evidence-Based Medical Statistics 2: The Bayes Factor. *Annals of Internal Medicine 130*, 1005–1013.

Ioannidis, John P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine 2*. https://doi.org/10.1371/journal.pmed.0020124.

Kass, Robert E. and Adrian E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association 90*, 773–795.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science 1*, 259–269.

Lee, Michael D. and Eric-Jan Wagenmakers (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Nosek, Brian A. and Timothy M. Errington (2017). Reproducibility in cancer biology: Making sense of replications. *eLife 6*, e23383.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science 349*. Retrieved from http://science.sciencemag.org/content/349/6251/aac4716.full.pdf.

Popper, Karl R. (1959/2002). *The Logic of Scientific Discovery*. London: Routledge. Reprint of the revised English 1959 edition. Originally published in German in 1934 as "Logik der Forschung".

Popper, Karl R. (1979). *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

Romeijn, Jan-Willem (2014). Philosophy of Statistics. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/statistics/.

Romero, Felipe (2017). Novelty vs. Replicability: Virtues and Vices in the Reward System of Science. *Philosophy of Science 84*, 1031–1043.

Rosenthal, Robert (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin 86*, 638–641.

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

Royall, Richard (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 760–768.

Schönbrodt, Felix D. and Eric-Jan Wagenmakers (2018). Bayes factor design analysis: Planning for Compelling Evidence. *Psychonomic Bulletin & Review 25*, 128–142.

Sober, Elliott (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge: Cambridge University Press.

Sprenger, Jan (2018). Two Impossibility Results for Popperian Corroboration. *British Journal for the Philosophy of Science 69*, 139–159.

Szucs, Denes and John Ioannnidis (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology 15*, e2000797.

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

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