UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A Bayesian perspective on severity: risky predictions and specific hypotheses

van Dongen, N.; Sprenger, J.; Wagenmakers, E.-J.

DOI
10.3758/s13423-022-02069-1

Publication date
2023

Document Version
Final published version

Published in
Psychonomic Bulletin and Review

License
CC BY

Link to publication

Citation for published version (APA):
van Dongen, N., Sprenger, J., & Wagenmakers, E-J. (2023). A Bayesian perspective on severity: risky predictions and specific hypotheses. *Psychonomic Bulletin and Review*, *30*(2), 516–533. https://doi.org/10.3758/s13423-022-02069-1

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:31 Aug 2023

Psychonomic Bulletin and Review (2023) 30:516–533
https://doi.org/10.3758/s13423-022-02069-1

THEORETICAL/REVIEW

A Bayesian perspective on severity: risky predictions and specific hypotheses

Noah van Dongen[1] · **Jan Sprenger**[2] · **Eric-Jan Wagenmakers**[1]

Accepted: 8 February 2022 / Published online: 15 August 2022
© The Author(s) 2022

Abstract

A tradition that goes back to Sir Karl R. Popper assesses the value of a statistical test primarily by its *severity*: was there an honest and stringent attempt to prove the tested hypothesis wrong? For "error statisticians" such as Mayo (1996, 2018), and frequentists more generally, severity is a key virtue in hypothesis tests. Conversely, failure to incorporate severity into statistical inference, as allegedly happens in Bayesian inference, counts as a major methodological shortcoming. Our paper pursues a double goal: First, we argue that the error-statistical explication of severity has substantive drawbacks; specifically, the neglect of research context and the specificity of the predictions of the hypothesis. Second, we argue that severity matters for Bayesian inference via the value of specific, risky predictions: severity boosts the expected evidential value of a Bayesian hypothesis test. We illustrate severity-based reasoning in Bayesian statistics by means of a practical example and discuss its advantages and potential drawbacks.

Keywords Statistical test · Bayes factors · Null hypothesis significance testing · Severity · Error statistics · Karl Popper · Deborah Mayo

> *What, then, is the end of an explanatory hypothesis? Its end is, through subjection to the test of experiment, to lead to the avoidance of all surprise and to the establishment of a habit of positive expectation that shall not be disappointed.* C.S. Peirce (1931)[CP 5.197]

The Bayesian framework for statistical inference—expressing one's uncertainty about hypotheses and parameters by means of subjective probabilities and updating them through Bayes' theorem—is increasingly popular in psychology (e.g., Rouder et al., 2009; Lee & Wagenmakers, 2013; Vandekerckhove et al., 2018). This popularity is easy to explain: Bayesian inference is based on a general theory of uncertain reasoning, its basic principles are simple and easily remembered, and statistical evidence is quantified by means of relative predictive performance (i.e., Bayes factors); moreover, Bayesian inference avoids the problems with the interpretation of p-values, and statistical inferences are connected to our beliefs and the practical consequences of our decisions (e.g, Jeffreys, 1961; Jeffrey, 1971; Savage, 1972; Bernardo & Smith, 1994; Lindley, 2000; Howson & Urbach, 2006; Evans, 2015; Morey et al., 2016; Sprenger & Hartmann, 2019).

A major objection to Bayesian inference consists in its apparent neglect of the role of severe hypothesis tests in scientific inference. For champions of severe testing like Sir Karl R. Popper (1963, 1959/2002), hypotheses are not confirmed by their predictive performance or their agreement with available data. Rather, they count as confirmed only if they have survived repeated and stringent attempts to prove them wrong. What matters for the status of a hypothesis is its *probative value*, or, in other words, whether it has "proved its mettle" (Popper, 1959/2002, p. 264). Via the work of methodologists and statisticians such as Ronald A. Fisher (1935, 1956) and Paul Meeh (1978, 1986, 1990a, 1990b), Popper's idea has also exerted a profound influence on statistical practice in social science.

Recently, Deborah Mayo (1996, 2018) has developed a philosophy of statistical inference called *error statistics*, based on the concept of severe testing. Mayo states explicitly that data should only count as evidence for a claim C if they resulted from a real and severe test; otherwise we

✉ Noah van Dongen
nnnvandongen@gmail.com

[1] University of Amsterdam, Amsterdam, Netherlands

[2] University of Turin, Turin, Italy

have "BENT: Bad Evidence—No Test" (p. 5 ; Mayo, 2018). Mayo's account deals primarily with controlling (Type I and II) error rates in inference and is therefore frequentist in nature (cf. Neyman & Pearson, 1933; 1967; Neyman, 1977; Mayo, 1996). One of the key objectives of this paper is to provide a critical analysis of Mayo's account and to assess the prospects of error statistics for application in social science.

The other objective consists in outlining how Bayesians can implement the ideas of severe testing and error control into their inference framework. Prima facie, neither the posterior probability of a hypothesis nor the Bayes factor seems to depend on the severity of a test, or the extent to which one has tried to prove the theory wrong. Specifically, our paper addresses the following questions:

Q1 Should Bayesians care about severity?
Q2 How can Bayesians account for the value of severity in inference?
Q3 Is severity more naturally accounted for in frequentist than in Bayesian inference?

We will defend an affirmative answer to the first question, sketch a constructive answer to the second question and give a negative answer to the third question. However, note that none of these positions is self-evident.

With respect to Q1, many Bayesians deny that severity should matter at all in inference. They refer to the *Likelihood Principle*: all the evidence that an experiment provides about an unknown quantity is expressed by the likelihood function of the various hypotheses on the observed data (Birnbaum, 1962; Edwards et al., 1963; Berger & Wolpert, 1984). "Consequently the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood" (Jeffreys, 1961, p. 57). Therefore, the severity of a test, as expressed by whether the hypothesis of interest stood a risk of being refuted (e.g., whether unobserved data could have proven it wrong), cannot enter the statistical evaluation of an experiment. According to this line of response, Popper, Mayo and other defenders of severe testing are just mistaken when they believe that severity should enter the (post-experimental) assessment of a theory. Much of the "statistics wars" Mayo (2018, p. xi) between Bayesians and frequentists have revolved around this controversy (see also Mayo & Kruse, 2001; Mayo, 2010). Our paper, by contrast, acknowledges Popper's and Mayo's argument that severity needs to be accounted for by an adequate logic of scientific, and statistical, inference.

With respect to Q2, Bayesians have only recently started to care about severity and to explain its role in Bayesian inference (Vanpaemel, 2010; Lee & Vanpaemel, 2018; Vanpaemel, 2019; 2020; Dienes, 2021). In particular, Vanpaemel (2020) argues that severity does not (only) consist in making precise predictions: a severe test has to rule out *plausible* outcomes, that is, outcomes with a high prior predictive probability (we engage with Vanpaemel's positions in a later section). Our own account highlights how the specificity of a hypothesis and its predictions boost severity by raising the expected evidential value of the experiment, and how error control can be embedded naturally into Bayesian inference. Specifically, we construe severity within the Bayesian framework as the *specificity of the prediction of a hypothesis in relation to the potential data that could be observed*. The value of severity is reflected in the *expected evidential value* of a test, quantified by the expected absolute log-Bayes factor (Lindley, 1956; Good, 1979; Cavagnaro et al., 2010; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019). In addition, we show, in line with Vanpaemel, how prior probability distributions connect scientific theory with a statistical model, and how priors contribute to the severity of a test (Vanpaemel, 2010; 2019).

Finally, with regard to Q3, we argue that the error-statistical explication of severity faces considerable conceptual challenges. Since error statistics enjoys increasing popularity in the psychological science community (see, e.g., Haig 2020 or the Shiny App by Richard Morey), we believe that this negative argument is an important contribution to the debate. In other words, we show that a statistical practitioner who cares about severe testing need not be a frequentist. Severity is a concept that can be integrated equally well, or even better, into the framework of Bayesian inference.

Theoretical background: scientific theories and severe tests

Scientific theories are imperfect descriptions and explanations of reality. They allow us, with a certain degree of accuracy, to predict future observations or to uncover structure in available data. According to Popper (1959/2002), the value or *empirical content* (p. 96) of a scientific theory lies in the combination of its *universality* and its *precision* (i.e., generality to what it pertains and specificity of what it predicts/explains; pp. 105–106). A theory has high empirical content when it has a vast scope in which many observations are possible in principle (e.g., all planetary motions in the universe); yet only a few of these possible observations are consistent with the theory (e.g., ellipses around a center of gravity). This is consistent with the common intuition that a successful risky prediction is more impressive than a successful vague prediction.

Meehl (1978) expressed this intuition with an example of two meteorological theories that both make predictions

on next month's weather. Theory A predicts: in Turin it will rain on three days in the next month. Theory B predicts: in Turin it will rain on the third, fourth, and the seventh day of next month. We expect that most people agree with Meehl that a success of Theory B's prediction is more impressive than the success of Theory A's prediction. This difference can be made explicit in terms of its *falsifiablity* (Popper, 1959/2002, p. 96): Theory B's prediction is only correct in $\binom{30}{1} = 1$ out of 2^{30} possibilities, while Theory A's prediction is correct in $\binom{30}{3} = 4060$ out of 2^{30} possibilities. What is important for the standing of a scientific theory is not its ability to fit the data, but how specific it is with respect to all possible data (see also Roberts & Pashler, 2000). Popper (1959/2002) gave the examples of Freudian psychoanalysis and Marxist sociology as theories that could fit any pattern of data, thus being low in empirical content and even unscientific. For these 'theories', both the presence and absence of a personality trait or both presence and absence of worker unrest could be considered as consistent with the theory.[1]

Competing theories are falsifiable to the extent that they make sharply contrasting predictions. The outcome of a test should either favor one theory and contradict the other, or vice-versa (i.e., strong inference: Platt 1964 or *experimentum crucis* as defined by Hooke: Lohne 1968):

> The scientific method is always comparative and there are no absolutes in the world of science. It follows from this comparative attitude that a good theory is one that enables you to think of an experiment that will lead to data that are highly probable on [the theory], highly improbable on [the complement of the theory], or vice versa, so that the likelihood ratio is extreme and your odds substantially changed. (p. 209 ; Lindley, 2006)

In modern mathematical language, this means that one needs to maximize the expected information gain regarding the assessment of the competing theories (Oaksford & Chater, 1994; Myung & Pitt, 2009; Myung et al., 2013). A prime example is Eddington's 1919 test of Einstein's General Theory of Relativity (GTR) versus Newton's classical theory of gravity (Dyson et al., 1920). Both theories make sharp mutually exclusive predictions about the degree to which passing massive bodies, like our sun, bend light from distant sources, like other stars. The predictions of GTR were famously verified by Eddington during the 1919 solar eclipse.

The benefits to science from increasing the falsifiability of theories and riskiness of predictions have been presented in numerous critiques of social science methodology (e.g., Meehl, 1978; 1986; 1990a; 1990b; Roberts & Pashler, 2000). In the case of GTR vs. Newtonian mechanics, the specificity of the predictions of a theory is tightly linked to the *capacity for testing it severely* (and thus, to its falsifiability). The higher the proportion of possible outcomes that are consistent with a theory, the less informative the theory becomes. For example, it is hard to falsify the rather uninformative theory that it will rain in Amsterdam on some days in the next year. On the opposite side, we have highly informative theories, with a limited set of parameters whose values are fixed and interrelated. An excellent example of such a theory is GTR. This theory makes specific predictions about time dilatation in GPS satellites relative to the earth's surface, improving navigational accuracy. The more informative a theory is, the more falsifiable it is, because fewer possible observations are consistent with it. Only such informative theories can be severely tested:

> We may say that to make predictions with great accuracy increases the probability that they will be found wrong, but in compensation they tell us much more if they are found right. [...] The best procedure, accordingly, is to state our laws as precisely as we can, while keeping a watch for any circumstances that may make it possible to test them more strictly than has been done hitherto. (pp. 39–40; Jeffreys, 1973)

We now investigate how severity manifests itself in various statistical frameworks, starting with Mayo's error statistics.

Severity in error statistics

At first glance there is a striking resemblance between Popper's emphasis on the severe testing of scientific hypotheses and *null hypothesis significance tests* (NHST). Typically, at the center of NHST there is a *point null hypothesis* postulating that a parameter takes a precise value (e.g., the mean of a population, $H_0 : \mu = \mu_0$)—corresponding to absence of a causal effect in an experimental intervention, equality of two medical treatments, and so on. When the measured divergence from the null hypothesis exceeds a given threshold (e.g., the observed test statistic falls in the most extreme 5% of the probability density function), the null hypothesis is *rejected* and the statistical analysis reports *statistically significant* results against the null. Otherwise, no conclusion is drawn.

At first glance, the NHST methodology squares well with Popper's falsificationism. First, the point null hypothesis is maximally falsifiable, because it commits to a single point value for the effect. Second, the idea of submitting the point null to a severe test and rejecting it when it fails to explain the data, has a distinct Popperian note. However, the

[1]For an argument that unverifiable hypotheses are essential for scientific progress see e.g., Poincaré (1913).

scientific theory that we would like to submit to a severe test is usually *not* the null hypothesis (see also Rouder et al., 2009; Gallistel, 2009; Sprenger & Hartmann, 2019, ch. 9). Rather, we would like to severely test the hypothesis that there *is* a meaningful effect: that a medical drug is better than a placebo, that playing a musical instrument makes people happier, that video games improve reasoning skills, and so on. This "alternative" hypothesis usually takes the generic form $H_1 : \mu > \mu_0$ (or even $H_1 : \mu \neq \mu_0$). It does not commit the scientist to precise predictions; rather it states that the unknown parameter lies in a *range* of values. NHST does not explain how such hypotheses can be tested severely; nor does it give a general definition of severity in statistical inference.

Mayo's (1996, 2018) *error-statistical approach* addresses this problem and explicates how parametric hypotheses can be severely tested in general. Error statistics follows Popper's perspective on theory testing as the ability to detect and control error between data and hypothesis. The standing of a theory is primarily determined by the severity of the tests it has survived. The extent to which the theory has proved its mettle constitutes statistical evidence. Mayo (2018, p. 14) states this explicitly in her Strong Severity Principle:

> **Severity Principle (strong):** We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of findings flaws or discrepancies from C, and yet none or few are found, the passing result, x, is evidence for C.

Prima facie, two important elements of severe testing are missing in the Severity Principle (in this formulation). First, Mayo's perspective does not make explicit essential elements of severe testing that Popper stressed: the universality of the tested hypothesis H and the specificity of the predictions it makes (Popper, 1959/2002, p. 266). Nowhere in Mayo's severity principle or in the specifications discussed below is the riskiness or specificity of a hypothesis, and the predictions it makes, incorporated as a requirement for a severe test (Meehl, 1978; 1986; 1990b; 1990a; Roberts & Pashler, 2000).[2]

Second, without explicit reference to alternatives to the claim, it is not clear how a claim can survive "stringent scrutiny" or when a test is "highly capable" of finding differences. For example, a stringent scrutiny of the claim C: "90% of all swans are white" requires only a single swan if the alternative claim is "all swans are black". When the alternative is that "fewer than 80% of all swans are white", however, the observation of a single white swan is relatively uninformative, and a large sample of swans is needed to discriminate between both hypotheses. If C encompasses all that is possible (except ¬C), no test —no matter how probative it usually is— will be capable of severely testing C. This suggests that the Severity Principle should not be interpreted in absolute, but in *relative terms*, comparing the claim to the most plausible and most severely tested competing hypotheses (e.g., as done by Eddington in the 1919 solar eclipse experiment). This contrastive aspect of severe testing is missing in Mayo's formalization.

This problem comes into sharper focus when we consider Mayo's operationalization of the Severity Principle Mayo (2018, p. 92):

> **Severity Requirement:** for data to warrant hypothesis H requires not just that
> (S-1) H agrees with the data (H passes the test), but also
> (S-2) with high probability, H would not have passed the test so well, were H false.

Suppose that the hypothesis of interest is $\mu \geq \mu_0 + \delta$, where $\delta > 0$ expresses the effect size relative to the default value μ_0. Mayo calls the hypothesis of interest the 'Claim' and indicates it with C. The application of (S-1) is then similar to the Neyman-Pearson (and NHST) approach (see Mayo 2018, p. 142): C passes a statistical test when $H_0 : \mu \leq \mu_0$ is rejected at the pre-specified Type I error level α.[3] This happens when the probability of the observed data, or data deviating more extremely from H_0, is lower than α (i.e., $p(d(X) \geq d(x); H_0) < \alpha$), where the statistic d measures divergence of the data from H_0).

To assess the evidence in favor of C, Mayo applies (S-2) and computes a severity function (SEV: see the corresponding Shiny App by Richard Morey). The severity function takes as arguments the statistical test, the observed data, and the target hypothesis C. The severity function outputs the probability of obtaining the observed data x or data closer to H_0 and deviating more from C, if C were false. As a representative of the negation of C, we choose the point hypothesis in C that is closest to H_0, which is $\mu = \mu_0 + \delta$. If this hypothesis is rejected with test T based on data x, the severity with which C has passed the test is defined as

$$\begin{aligned} SEV(T, x, C) &= p(d(X) \leq d(x); \neg C) \\ &= p(d(X) \leq d(x); \mu = \mu_0 + \delta), \end{aligned}$$

[2] We recognize that in principle it is easier to scrutinize a precise prediction than it is to scrutinize a vague prediction. However, precision is not a requirement for a stringent test as Mayo (2018) defines it.

[3] Actually, H_0 is rejected in favor of the standard alternative hypothesis $H_a : \mu > \mu_0$, to keep the parameter space neatly partitioned, but this does mean that the data are consistent with C.

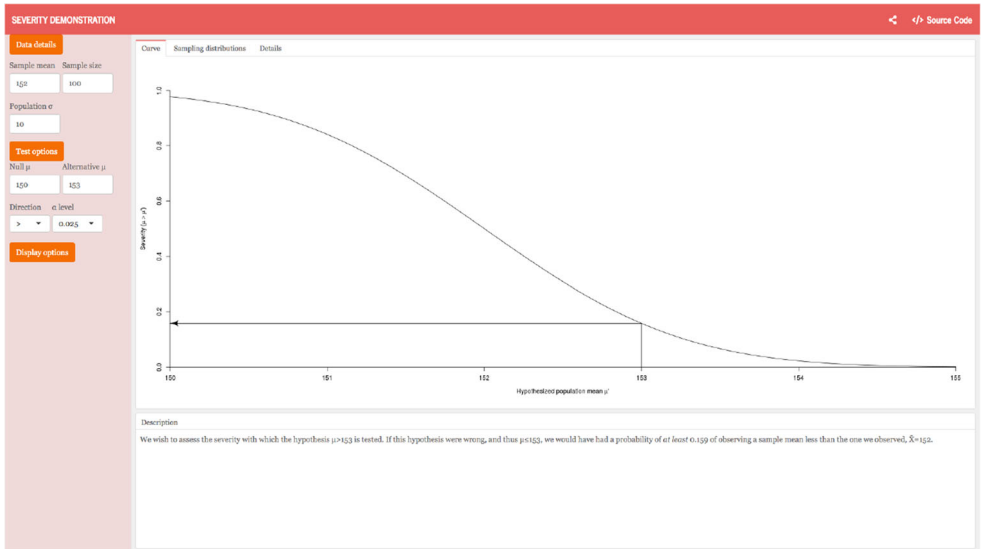

Fig. 1 SEV results for the original water plant example. In this case, $H_0 : \mu = 150$, $C : \mu = 153$, $\overline{x} = 152$, and $SEV = 0.159$. This image is a screenshot from the *Severity Demonstration application*. This Shiny App was developed by Morey (2020) and can be accessed via https://richarddmorey.shinyapps.io/severity

which is (a lower bound on) the probability of observing the actually observed data, or data more deviating from C, in the direction of H_0 if C were false.[4]

To show how the severity function works in practice, we borrow a simple example from (Mayo, 2018, pp. 142–144). The example concerns the scenario of an "accident at a water plant" (p. 142) where a leak in the cooling system is discharging water into the ecosystem. The cooling system is "meant to ensure that the mean temperature of discharged water stays below the temperature that threatens the ecosystem, perhaps not much beyond 150 degrees Fahrenheit."

The question is, whether the temperature of the water high enough to constitute an ecological disaster that requires counter measures beyond mere repair of the cooling system. When the cooling system is working properly, the temperature of the water discharged into the ecosystem is around 150 degrees Fahrenheit and a temperature of 153 or higher is considered a "full-on emergency for the ecosystem" (p. 143). A series of 100 water measurements are taken at random time periods and their sample mean $\overline{x}$ is computed. The standard deviation is known and when the cooling system is working properly, the distribution of means from samples of 100 measurements is captured by $\overline{X} \sim N(\mu = 150, 1)$. Mayo supposes that initially the test concerns "$H_0 : \mu \leq 150$ vs. $C : \mu > 150$" (p. 142). She sets the Type I error rate to 2.5% ($\alpha = 0.025$), thus rejects "H_0 (infer there's an indication that $\mu > 150$) iff $\overline{X} \geq 152$" (p. 142).[5]

In this scenario, we observe a sample mean of 152 degrees Fahrenheit and thus reject H_0, fulfilling (S-1) of the Severity Requirement. According to the severity rationale, we are warranted in concluding that $\mu > 150$, because "[w]ere the mean temperature no higher than 150, then over 97% of the time their method would have resulted in a lower mean temperature than observed" (p. 143). However, we are primarily concerned with the extent to which these data indicate that the water has reached devastating temperatures of 153 Fahrenheit or higher ($\mu \geq 150 + \delta$; $\delta = 3$). According to the Severity Requirement (S-2), we need the probability of observing our mean temperature of 152 or lower if the actual temperature is equal the critical threshold of 153 degrees. This can be calculated with the SEV function, which is $p(\overline{X} \leq 152; \mu = 153) \approx 0.16$. From these results, one can conclude that one does not have evidence for the claim that the water has reached temperatures of 153 degrees Fahrenheit or higher. Namely, the "severity principle blocks $\mu > 153$ because it is fairly probable (84% of the time) that the test would yield an even larger mean temperature than we got, if the water samples came from a body of water whose mean temperature is 153" (p. 144). Thus, the results do not meet requirement (S-2) and the hypothesis that the temperature of the water is 153 degrees or higher does not pass a severe test (see Fig. 1).

With this depiction, a consequence of the second problem described above comes into view; the severity principle operationalized as the SEV function implies a curious,

[4]For convenience, the SEV function is evaluated at $\mu = \mu_0 + \delta$, because the probability value that it produces will be even greater for $\mu < \mu_0 + \delta$ (for detailed explanation, see Mayo 2018, p. 144).

[5]For convenience, Mayo rounds 151.96 up to 152.

perhaps outright problematic concept of statistical evidence. In the water plant example, the observed mean temperature of 152 degrees Fahrenheit is not considered evidence that the water has reached the dangerous temperature of 153 degrees ($SEV(T, \overline{x}, \mu > 153) \approx 0.16$). However, this result is *independent of what is considered the normal or default state of affairs*. In Mayo's example, this is 150 degrees Fahrenheit, but this value is irrelevant for her SEV function as long as the observed mean is sufficiently high to reject the null hypothesis. Suppose that we had observed the same mean temperature of 152 degrees Fahrenheit, though the normal mean temperature is 100 degrees. Then one would again reject H_0 and meet requirement (S-1); one would again not meet (S-2) and draw the same conclusion as before (see Fig. 2).

However, when normal temperatures are around 100 degrees, you observe a mean temperature of 152 with a standard error of 1, and an "full-on emergency for the ecosystem" is imminent when the temperature is 153, it is clear that counter measures are acutely required. Evidence must be evaluated in context and relative to the competing or default hypotheses. The error-statistical analysis of the case goes against any intuition one might have about the concept of severity, treatments of this concept by other philosophers and methodologists (e.g., Popper, 1959/2002; Meehl, 1990b; 2005; Roberts and Pashler, 2000), and even against a reasonable interpretation of Mayo's Severity Principle (Mayo, 2018, p. 14).

For this reason, we disagree that Mayo's Severity Principle and Severity Requirement (as explicated by the Severity Function) are good operationalizations of the function of severity in scientific inference and statistical testing. To recapitulate, our main criticism are (1) any theory of severe testing must be sensitive to what is considered the normal or default state of affairs, and Mayo's theory does not; (2) more generally, Mayo's theory neglects that the severity of a test is a function of the degree of similarity between the competing hypotheses, and the specificity of their predictions. Mayo's explications of severity in statistical inference are decoupled from the scientific context in which the data are collected.

Severity in Bayesian inference

In Bayesian inference, an observation x supports hypothesis H if and only if x raises the subjective probability of H (e.g., Carnap, 1950; Horwich, 1982; Howson & Urbach, 2006; Evans, 2015; Sprenger & Hartmann, 2019):

$$p(H \mid x) > p(H).$$

Equivalently, in the case of a null hypothesis H_0 and an alternative H_1, x is evidence for H_1 if and only if x is better predicted under H_1 than under H_0, i.e., if the *Bayes factor* BF_{10} exceeds 1:

$$\mathrm{BF}_{10}(x) := \frac{p(x \mid H_1)}{p(x \mid H_0)} > 1.$$

The higher the Bayes factor, the stronger the evidence for H_1, and the closer it is to zero, the stronger the evidence for H_0. See Table 1 for a conventional classification of Bayes factors.

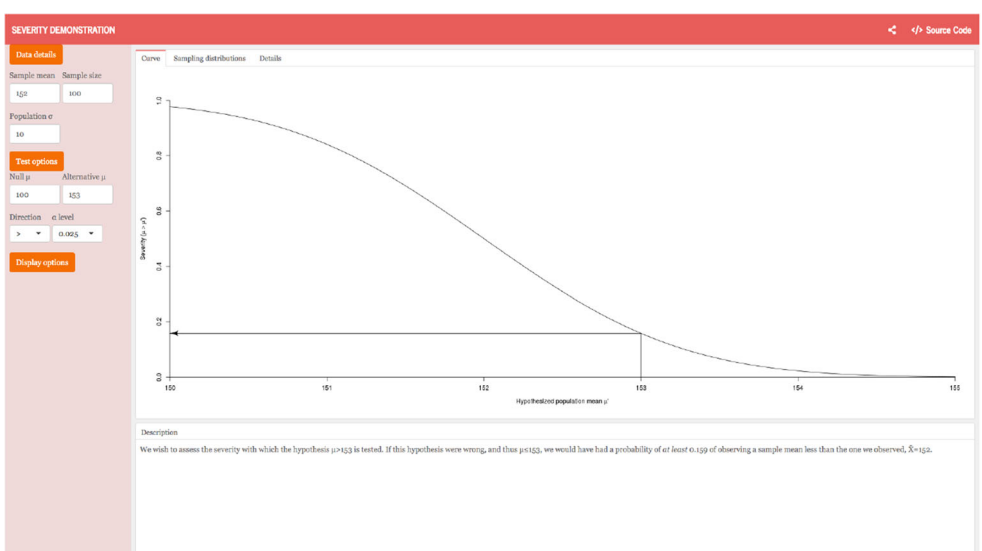

Fig. 2 SEV results for the original water plant example. In this case, $H_0 : \mu = 100$, $C : \mu = 153$, $\overline{x} = 152$, and $SEV = 0.159$. This image is a screenshot from the *Severity Demonstration application*. This Shiny App was developed by Morey (2020) and can be accessed via https://richarddmorey.shinyapps.io/severity

Table 1 Classification of Bayes factors according to Lee and Wagenmakers (2013), adjusted from Jeffreys (1961)

Bayes Factor BF_{10}	Interpretation
>100	Extreme evidence for H_1
30–100	Very strong evidence for H_1
10–30	Strong evidence for H_1
3–10	Moderate evidence for H_1
1–3	Anecdotal evidence for H_1
1	No evidence for either hypothesis
1/3–1	Anecdotal evidence for H_0
1/3–1/10	Moderate evidence for H_0
1/10–1/30	Strong evidence for H_0
1/30–1/100	Very strong evidence for H_0
<1/100	Extreme evidence for H_0

The posterior probability of a theory H_0 given data x is calculated as $p(H_0 \mid x) = p(H_0) \cdot p(x \mid H_0)/p(x)$. This allows us to write the posterior odds in favor of H_0 over H_1 as

$$\frac{p(H_0 \mid x)}{p(H_1 \mid x)} = \frac{p(H_0)}{p(H_1)} \cdot \frac{p(x \mid H_0)}{p(x \mid H_1)}.$$

Thus, keeping the prior probability $p(H_0)$ fixed, the posterior probability of H_0 will be the larger (1) the better H_0 predicts x and (2) the more surprising (worse predicted) x is under H_1 (e.g., Roberts and Pashler, 2000; Howson & Urbach, 2006).

By construction, the Bayes factor is relative to an explicit choice of context, i.e., the alternative hypothesis. Quite often, this alternative hypothesis will depend on the prior distribution of the parameter of interest, e.g., when we test a point null hypothesis $H_0 : \mu = \mu_0$ against an unspecific alternative $H_0 : \mu \neq \mu_0$. However, as argued by Vanpaemel (2010), this property of Bayes factors is a virtue rather than a vice: the prior distribution expresses, after all, our theoretical expectations and predictions. A scientist using Bayesian inference needs to think about the prior distribution in advance, strengthening the link between scientific theorizing and statistical analysis whose absence has often been named as a cause of the lack of reliability and replicability of psychological research (compare Meehl, 1967; Ioannidis, 2005; Dienes, 2021).

At first glance, severity-related aspects appear to be lacking in the Bayesian paradigm: the Bayes factor only depends on the probability of the data in light of the two competing hypotheses. As Mayo emphasizes (e.g., Mayo and Kruse, 2001; Mayo, 2018), the Bayes factor is insensitive to variations the sampling protocol that affect the error rates, i.e., optional stopping of the experiment. The Bayes factor only depends on the actually observed data, and not on whether they have been collected from an experiment with fixed or variable sample size, and so on. In other words, the Bayesian ex-post evaluation of the evidence stays the same regardless of whether the test has been conducted in a severe or less severe fashion.

We agree with this observation, but we believe that the proper place for severity in statistical inference is in the choice of the tested hypotheses (are they specific? are they sufficiently contrastive?), and in the experimental design. Bayesians cash out severity by ensuring that an experiment gives positive answers to the two questions below:

Evidential Value Is it probable to *obtain strong, discriminatory evidence* from this particular experiment?

Error Control Does the design of the experiment *limit the probability of finding misleading evidence*? That is, does it limit the probability of making an erroneous inference?

These properties also depend on the experimental design as a whole (e.g., the chosen sample size), but specifically on the choice of the tested hypotheses:

Specificity Do the competing hypotheses H_0 and H_1 make *specific predictions*? That is, do they rule out, or make implausible, a large proportion of possible outcomes?

Contrastivity Are the predictions of hypothesis H_0 and H_1 *sufficiently different* that the experiment can discriminate between them?

In other words, the specificity and contrastivity of a hypothesis contributes to the severity of a test by increasing its evidential value and controlling error in inference. We will now discuss these concepts in detail.

Specificity and evidential value in Bayesian inference

For a test to be severe, the tested hypothesis needs to *impose substantial restrictions on the range of potential data* that are consistent with it. This is a basic ingredient of severe testing that has been retained both by philosophers such as Popper and psychological science methodologists. For example, Roberts and Pashler (2000, p. 359) argue that a good fit between data and model does not have to be convincing as evidence when the parameter values of the model are fully adjustable to accommodate the data:

> Theorists who use good fits as evidence seem to reason as follows: if our theory is correct, it will be able to fit the data; our theory fits the data; therefore it is more likely that our theory is correct. *However, if a theory did not constrain possible outcomes, the fit is meaningless.* (Italics added for emphasis)

Figure 3 shows for a theory concerned with the relation between the values of two measures that the data provide

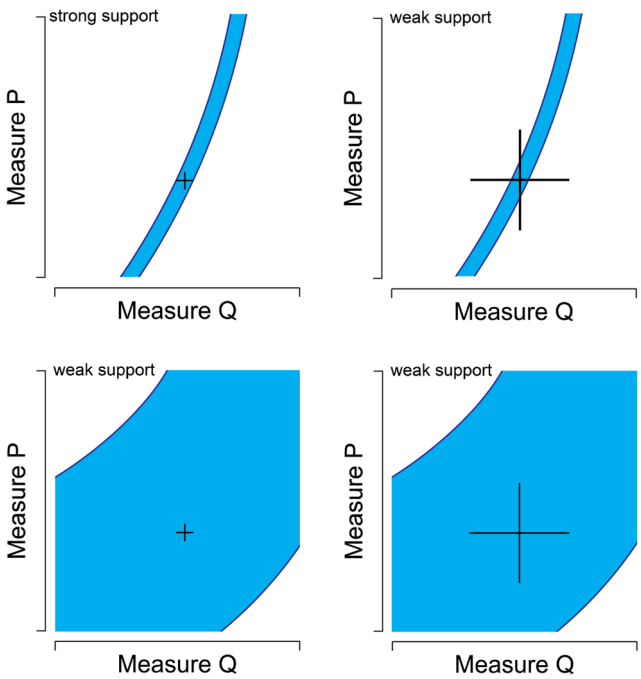

Fig. 3 Four possible relations between data and theory. Measures P and Q are both measures of some observable. The axes cover the range of possible values. The highlighted areas indicate the outcomes that are consistent with the theory. Standard errors of the observation are indicated by the error bars. In this example, the theory fits the data, though only when both theory and data are sufficiently constraint (upper left) does this provide significant evidence for the theory (this figure is published under CC-BY 4.0 and is adapted from: Roberts & Pashler, 2000, p. 360)

strong support only if (a) a narrow range of possible values is consistent with the theory; (b) the data fall in this narrow range; (c) the experimental measurements were precise. These three criteria can be summarized as specificity of predictions, fit and measurement precision.

Vanpaemel (2010, 2020) points out that the Bayesian can express criterion (a) by means of the prior predictive distribution $\int p(x|\theta, H)\, p(\theta|H)\, d\theta$: all outcomes that fall outside the "core predictions" of a hypothesis H (e.g., the data points with the 95%, 99% or 99.99% prior probability mass) are judged to be inconsistent with the theory.[6] In particular, Vanpaemel (2020) demands that the tested hypothesis do not only rule out possible, but also *plausible* outcomes. This means that possible outcomes that fall outside the X% highest density region as predicted by the hypothesis have some non-negligible probability based on, for instance, prior results (for a more detailed explanation, see Vanpaemel, 2020). Only in this case can a hypothesis be tested severely.[7]

Two comments or amendments on Vanpaemel's account are in order. First, Vanpaemel defines the "plausible" in the requirement of excluding plausible outcomes as independent of the models that are entertained. However, he does not explain how this understanding squares with, or should be reconciled with, the definition of "plausible" via the data prior in the model that encompasses the tested hypotheses. Neither does Vanpaemel offer a clear specification of *how many* plausible outcomes, relative to the overall data space, need to be excluded.

Second, Vanpaemel argues that the threshold for counting outcomes as inconsistent with a model depends on the context: in his running example of the 2020 paper, he chooses 99.99%, but he adds that in other contexts, other thresholds may be possible. Also here, we agree, but we would like to add that we (still) cannot test a hypothesis H in isolation, in the sense of "if the actual data fall outside the core predictions, the theory is H refuted". This would amount to relying on Fisher's infamous disjunction (Fisher, 1956; Cohen, 1994)—either a very improbable event has occurred, or the tested hypothesis must be wrong—and replicating the inverse probability fallacy: since $p(E|H)$ is low, and since we want to explain events in a systematic way, we infer that $p(H|E)$ is low. *A theory must always be tested against a competitor.* Indeed, this aspect of severe testing is underappreciated in Vanpaemel's account, just as it is missing in Pashler and Roberts's and Mayo's account.

From a Bayesian perspective, specificity contributes to severity by means of increasing the diagnosticity or *expected evidential value* of a hypothesis test. This concept can be operationalized as the expected absolute log-Bayes factor[8] of the experiment (Good, 1950; 1975; 1983; 1979; Lindley, 1956; Nelson, 2005; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019):

$$\mathbb{E}_y\left[|\log \mathrm{BF}_{10}|\right] = \int p(y)|\log \mathrm{BF}_{10}(y)|\,\mathrm{d}y, \tag{1}$$

[6]This idea is similar to using credible intervals based on regions of highest posterior density (ROPE) in order to draw inferences from Bayesian hypothesis tests (Kruschke, 2018; Kruschke & Liddell, 2018).

[7]For Vanpaemel, the prior probability distribution for the model parameters is an indispensable element of evaluating the severity of a tests, since the data prior depends on the prior over the parameter space: "Wide data priors make for lenient tests; narrow data priors imply strong tests" (142 ; Vanpaemel, 2020). Since narrow data priors are usually obtained by means of choosing specific hypotheses that occupy only a small range of the parameter space, severe testing is also connected to specificity on his account.

[8]Nelson (2005) objects to the use of the expected log-Bayes factor as a measure of diagnostic value on a variety of reasons, of which we recapitulate the two most important ones: (1) it takes infinite value if an outcome is *logically excluded* by H_0; (2) the lack of sensitivity to the prior probability of the hypotheses is not psychologically adequate, as he seeks to demonstrate using a series of cognitive experiments.

where p is the prior probability density function over the sample space that encompasses both H_0 and H_1.[9] The expected absolute log-Bayes factor quantifies the amount of evidence one can expect, for and against combined, where values close to $\mathbb{E}_y|\log \mathrm{BF}_{10}| = 0$ constitute uninformative tests (i.e., the hypotheses make similar predictions).[10]

When a hypothesis is highly specific, its predictions will in general differ notably from the alternative, and this means that we can expect strong evidential support for either one or the other hypothesis. Furthermore, the expected evidential value is higher for the more restricted hypothesis: it is easier to find evidence against it when it is false, and one obtains more easily evidence for it when it is true. Testing simple, restrictive hypotheses is thus valuable from a Bayesian perspective. Typically—but not always—the most simple, and most severely testable hypothesis is a point null hypothesis H_0 which fixes the parameter of interest to a single number. More generally, there is a systematic connection between the degree to which a hypothesis restricts its parameters, and the degree to which it restricts the range of data consistent with it.[11] We will now illustrate this reasoning with a simple binomial model.

An example of specific hypothesis testing

Let us clarify this specification of the Bayes factor with a simple fictitious example. For this example, we have adapted the *informative hypothesis testing* approach (Klugkist & Hoijtink, 2005; Klugkist et al., 2005; Hoijtink, 2011). The Bayes factors reported in this section are calculated with respect to an encompassing model with a uniform prior. We have added the mathematical explication as an Appendix.

In this example, we imagine that military veterans receive either one of two treatments to address post traumatic stress disorder (PTSD). Treatment A is regular psychotherapy (e.g, Cognitive Processing Therapy; Monson et al., 2006). Treatment B is the same psychotherapy but enhanced with regulated dosages of the drug 3,4-Methylene-dioxymethamphetamine (Sessa, 2017).

Based on the theory and previous research, success rates of these treatments should fall in a particular range. The more uncertainties there are in the theory about what could influence the success rate (e.g., inter-patient psychological variability) and the more measurement error one can expect (e.g., low reliability of PTSD assessment), the wider the interval around these expected values that are still considered consistent with the theory.

Specifically, assume that either treatment has a true success rate, θ_A and θ_B. Under an encompassing model these rates can have any value between 0 and 1, $H_e : \theta_A = [0.0, 1.0],\ \ \theta_B = [0.0, 1.0]$. For simplicity, we consider these values equally likely and adopt a uniform prior over θ_A and θ_B in the encompassing model. The scenario can be modelled as two independent binomial distributions. We start treating participants, randomly divided in equal number to each treatment, and count the number of successful treatments. In this case, for a treatment we would have a specific number of participants N_i, where $i \in \{A, B\}$, and the number of successes for this treatment S_i would be determined by the actual value of θ_i and sampling error: $S_i \sim \mathrm{Bin}(N_i, \theta_i)$.

We use four scenarios to allow comparison with the four cases in Fig. 3 identified by Roberts and Pashler (2000). Specifically, we consider two hypotheses, one vague and one specific, and two different datasets, one with few and one with many patients. In the first and second scenario, corresponding to the bottom panels in Fig. 3, the hypothesis makes vague predictions: $H_v : \theta_A = [0.00, 0.50],\ \theta_B = [0.50, 1.00]$.[12] Thus, H_v takes up 25% of the encompassing model and it predicts that the success rates for Treatment A and B will very likely be below and above 0.5, respectively (see Fig. 4). From Vanpaemel's perspective, a severe test of this (vague) hypothesis is possible, because plausible outcomes fall outside the highest density region. Specifically, between 14 and 20 successes for treatment A and between 0 and 6 successes for treatment B fall outside the 99% highest density region.

In the third and fourth scenario, corresponding to the two top panels in Fig. 3, the hypothesis makes specific predictions and strongly restricts the possible parameter values: $H_s : \theta_A = [0.20, 0.30],\ \theta_B = [0.70, 0.80]$. Thus, H_s takes up 1% of the encompassing model and it predicts that success rates between 0.2 and 0.3 for Treatment A and between 0.7 and 0.8 for Treatment B are most probable while success rates above and below those values are increasingly unlikely (see Fig. 5). From Vanpaemel's perspective, a severe test of this (specific) hypothesis is possible, because plausible outcomes fall outside the highest density region. Specifically, 0 and between 12 and 20 successes for treatment A and between 0 and 9 and 20 successes for treatment B fall outside the 99% highest density region.[13]

[9] This expression does not depend on the choice of H_0 vs. H_1 since $|\log BF_{10}| = |\log BF_{10}|$.

[10] The idea of taking an expectation over the sample space may look alien to Bayesian inference, but it is a well-established part of the modern Bayesian's toolbox: see de Finetti (1974, pp. 90-91), Pérez and Berger (2002) and O'Hagan and Forster (2004, pp. 166-167).

[11] Popper notes that a hypothesis is the more testable, and the more restrictive on the data, the simpler it is, where simplicity is defined as paucity of parameters (Popper, 1959/2002, pp. 126–128 and 392–394).

[12] Locations of these hypotheses are chosen for symmetry; not much changes when different locations are used.

[13] A case could be made that 0 and 20 successes are implausible outcomes, because they fall outside the highest density regions for both treatments. However, this is irrelevant for deciding whether or not the hypothesis can be tested severely according to Vanpaemel (2020).

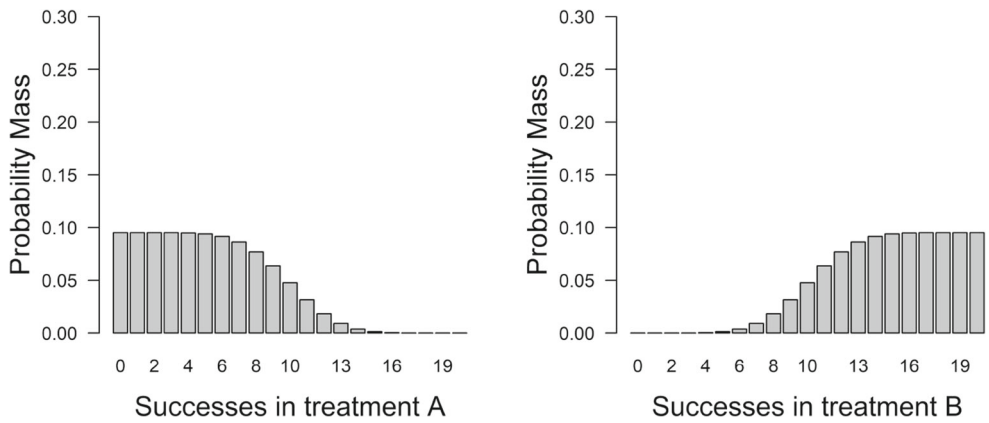

Fig. 4 Data predictions for the vague hypothesis. If twenty patients were to be tested in both Treatment A and Treatments B, then according to H_v we would expect to see these numbers of success with these probabilities

These hypotheses are compared on two data sets. A small data set of four participants per treatment, $N_A = N_B = 4$ represents the two right panels of the Roberts and Pashler (2000) quartet. A relatively large data set of twenty participants per treatment, $N_A = N_B = 20$, represents the two left panels of the Roberts and Pashler (2000) quartet. If the data align well with each hypothesis (e.g., $S_A = \frac{1}{4} N_A$ and $S_B = \frac{3}{4} N_B$), the specific hypothesis H_s is better supported than the vague hypothesis H_v. It is even the case that the specific hypothesis is better supported by the small data set ($\mathrm{BF}_{se} = 4.37$) than the vague hypothesis is supported by the large data set ($\mathrm{BF}_{ve} = 3.89$). As shown in Fig. 6, this scenario qualitatively reproduces the figure of Roberts and Pashler (2000).

However, the benefit of a specific hypothesis when right comes at a cost when wrong. As is visible in Fig. 7, H_s only has a small amount wiggle room and the Bayes factor quickly drops towards zero when the number of successes deviates from the predicted range, while H_v has a large area to move in with respect to possible success rates that support it.

This is also reflected in the expected evidential value of the experiment in the four situations of the example. As explained in the previous section (see Eq. 1), the expected absolute log-Bayes factor describes the evidential value one can expect either for or against the specific hypothesis in relation to the encompassing model, where the expectation is taken over all possible data y:

$$\mathbb{E}\left[|\log \mathrm{BF}_{se}|\right] = \int p(y)|\log \mathrm{BF}_{se}(y)|\,\mathrm{d}y$$

where $p(y)$ is

$$p(y) = p(H_s)p(y \mid H_s) + p(H_e)p(y \mid H_e)$$

and $p(y \mid H)$ is

$$p(y \mid H) = \int p(y \mid \theta, H)g(\theta \mid H)\,\mathrm{d}\theta$$

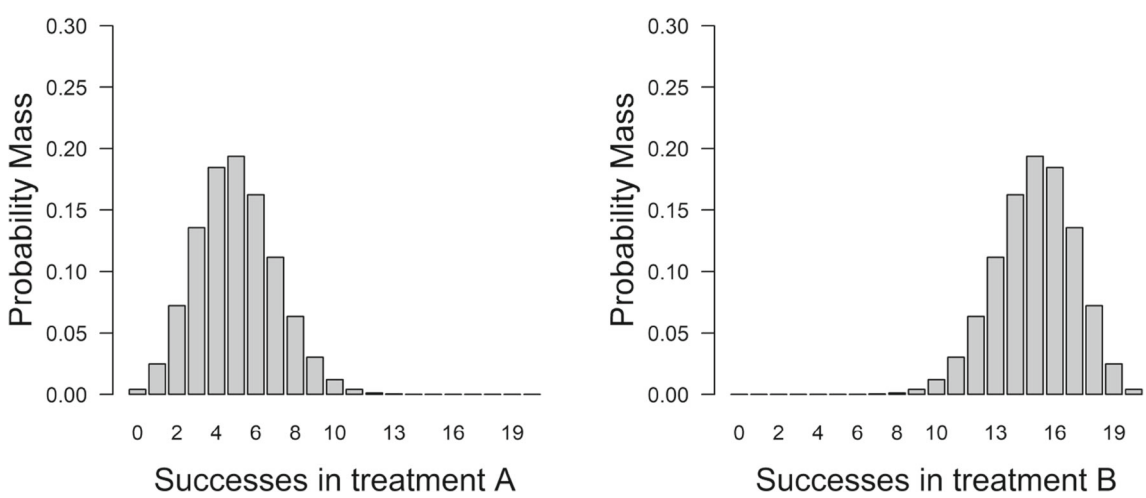

Fig. 5 Data predictions for the specific hypothesis. If twenty patients were to be tested in both Treatment A and Treatments B, then according to H_s we would expect to see these numbers of success with these probabilities

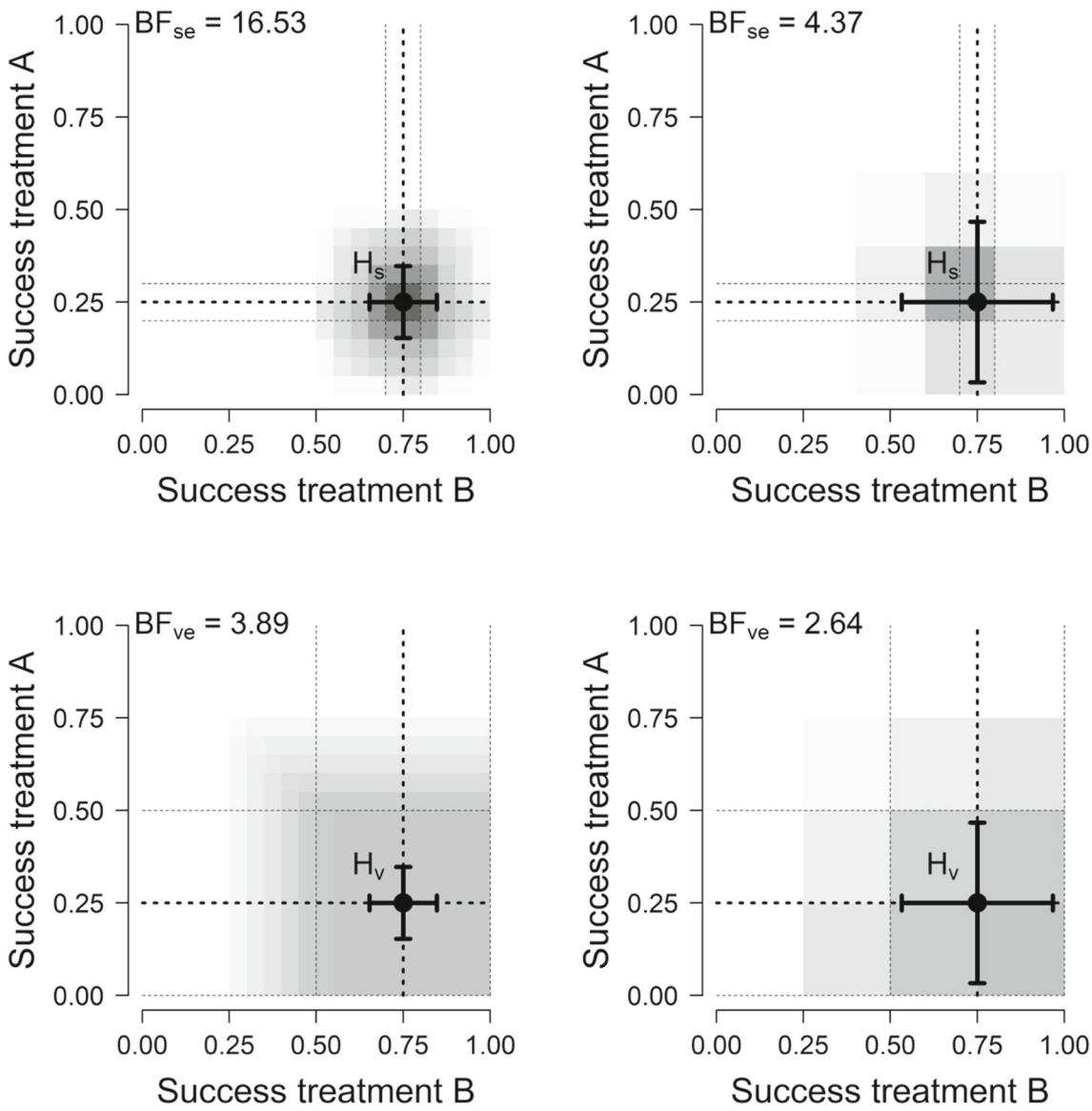

Fig. 6 Four relations between hypothesis and data. The top two graphs show the results of specific hypothesis H_s. The bottom two graphs show the results of the vague hypothesis H_v. The gradient gray areas depict the probabilities mass with respect to the hypotheses' predicted outcomes (a top view of Figs. 5 and 4). The dotted-lines in gray visualize the hypotheses' restrictions on the parameter values. The point estimates and standard errors are visualized as black crosses. The graphs in the left column display the results of the large data set ($N_i = 20$) and the graphs in the right column display the results of the small data set ($N_i = 4$). In this example, the data align perfectly with the hypotheses. The evidential support for the hypotheses in comparison to the encompassing model is quantified as Bayes factors (top-left of each plot). From top-left to bottom-right, these Bayes factors are 16.53, 4.37, 3.89, and 2.64 respectively

In this example, the computation of $\mathbb{E}\left[|\log \mathrm{BF}_{se}|\right]$ comes down to taking the average of absolute log Bayes factors over all possible treatment outcomes for both Treatment A and Treatment B, weighted by the marginal probability of these treatment outcomes according to H_s and the encompassing model. Analogously for $\mathbb{E}\left[|\log \mathrm{BF}_{ve}|\right]$.

As can be gleaned from Fig. 7, there is substantial similarity between the vague hypothesis and the encompassing model, in the sense that a quite large set of possible outcomes is explained to a similar degree by either hypothesis. This property decreases the expected evidential value of the experiment. When we move to specific hypothesis, on the other hand, there are striking differences between the outcomes expected under the hypothesis and the encompassing model. This is also reflected in the expected evidential values: $\mathbb{E}_y\left[|\log \mathrm{BF}_{se}|\right] > \mathbb{E}_y\left[|\log \mathrm{BF}_{ve}|\right]$. The expected evidential value is highest for the specific hypothesis with the predetermined sample size of 20 per treatment, $\mathbb{E}[|\log \mathrm{BF}_{se}|] = 6.32$. This is followed by the vague hypothesis with $N_i = 20$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{ve}|] =$

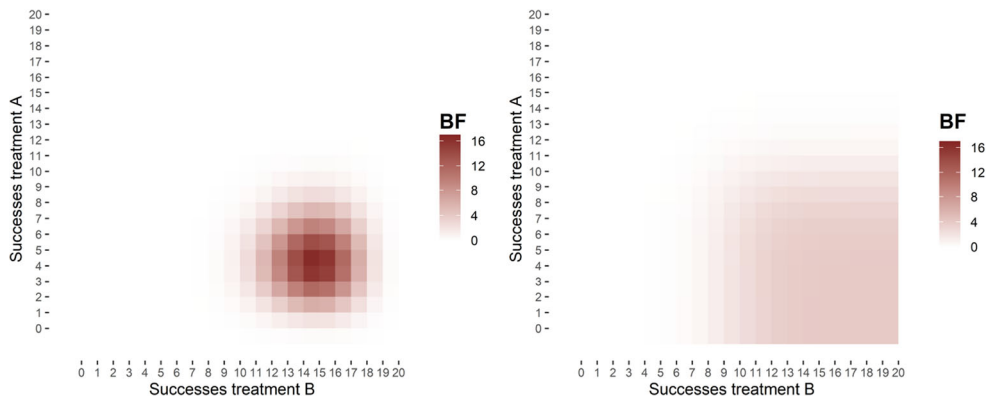

Fig. 7 Specific hypotheses can yield more evidence than vague hypotheses, though in most situations this is evidence against the hypothesis. The graphs displays the size of the Bayes factor with respect to possible combinations of outcomes for the specific hypothesis H_s (left) and the vague hypothesis H_v (right). The x-axis and y-axis indicate, for treatment A and treatment B respectively, the number of successful treatments (S_i) out of the 20 participants that are treated ($N_i = 20$). The values within the lattice indicate the evidence for the hypothesis that results from the particular combination of S_A and S_B. This evidence is quantified in terms of the Bayes factor for the hypothesis (left: specific, right: vague) with respect to the encompassing model. The size of the Bayes factor is also indicated as the intensity in color

3.14; the specific hypothesis with $N_i = 4$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{se}|] = 1.61$; and the vague hypothesis with $N_i = 4$ per treatment, $\mathbb{E}[|\log \mathrm{BF}_{ve}|] = 1.25$.[14]

Summing up, the evidence is much more telling and discriminatory when testing specific hypotheses (cf. Etz et al., 2018). From a Bayesian perspective, these numbers show why severe testing is important for inference and decision-making, and what specific hypotheses with contrastive predictions contribute to the value of an experiment. We have examined the case of testing a nested hypothesis against an encompassing model, but the case generalizes straightforwardly to other statistical models, and specifically, to models with contrastive, mutually exclusive hypotheses.

Error control and contrastive hypotheses in Bayesian inference

For the Bayesian, it is essential that theory testing, and scientific method as a whole, are *comparative*. Experiments can be probative and severe only when contrasting a hypothesis H_0 to an explicit alternative H_1. And this alternative must be specified in the light of our best scientific knowledge, i.e., in the light of our prior expectations. Specifically, the Bayesian answers the question "... what constitutes a good theory?" (Lindley, 2006, p. 196) in terms of its potential to yield high likelihood ratios in testing the theory against its negation:

> [...] A good theory is one that makes lots of predictions that can be tested, preferably predictions that are less probable were the theory not true. [...] what is wanted are data that are highly likely when the theory is true, and unlikely when false. A good theory cries out with good testing possibilities (Lindley, 2006, p. 197, notation changed).

That experiments should be as contrastive as possible has already been highlighted in the previous section. It is especially important because the more two hypotheses make contrasting predictions, the better can we limit the *probability of misleading evidence*, along the lines of the Lindley) (see also Royall 2000, for a semi-Bayesian, likelihoodist treatment), and assess whether the test was severe.

One Bayesian tool for making this assessment is a Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers 2018, see also shinyapps.org/apps/BFDA/). Before collecting data, the experimenter decides on a threshold for when a Bayes factor counts as evidence for a hypothesis, e.g. $\mathrm{BF}_{10} > 3$ or $\mathrm{BF}_{10} < 1/3$. Of course, one can set more stringent requirements as one desired (e.g., $\mathrm{BF}_{10} = 10$, the threshold for strong evidence). Suppose the alternative hypothesis H_1 is in fact true. The BFDA then calculates

- the probability of obtaining positive evidence for H_1 (e.g., $\mathrm{BF}_{10} > 3$);
- the probability of obtaining inconclusive evidence (e.g., $1/3 \leq \mathrm{BF}_{10} \leq 3$); and
- the probability of obtaining misleading evidence for H_0 (e.g., $\mathrm{BF}_{10} < 1/3$).

Identically, one can define the probability of obtaining positive evidence for H_0, and misleading evidence for H_1 (when H_0 is true). A graphical illustration is given in Fig. 8

[14]The example suggests that point hypotheses are easier to test severely than composite hypotheses, but this holds only *ceteris paribus* and cannot be generalized without qualification.

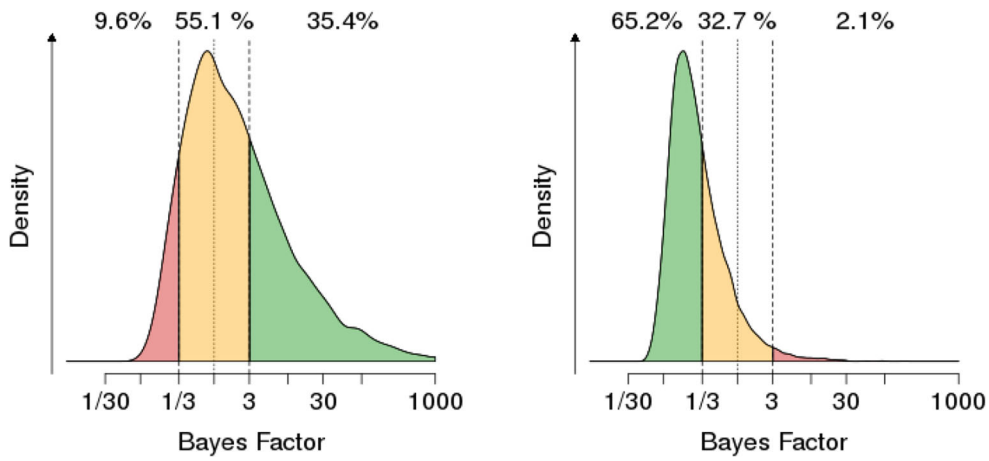

Fig. 8 Graphical illustration of the distribution of Bayes factors and probability of misleading evidence in the Bayesian framework for N=36, a two-sided t-test and hypothesized effects of $d = 0$ and $d = 0.4$. Figure produced with the BFDA app https://shinyapps.org/apps/BFDA/ based on Schönbrodt and Wagenmakers (2018)

for a two-sided t-test with a moderate group size $N = 36$ in both groups. The probability of finding misleading evidence for H_0 when H_1 is true is relatively large: 9.6%. Moreover, the most likely result is actually to obtain inconclusive evidence (55.1%). Certainly such an experiment controls error rates (here defined not as Type I or Type II errors, but in the above sense of finding misleading evidence) poorly and cannot be counted as a severe test of the competing hypotheses.

When the sample size increases, the impact of sampling variability is reduced and this facilitates the control of the probability of misleading evidence. With $N = 190$, for example, these rates drop to 0,2% and 1%, respectively. See Fig. 9.

This shows that error rates are no unique feature of frequentist statistics: the Bayesian has analogous tools to calculate the probability of misleading evidence and to construct his (or her) experiment in a way to control for the probability of misleading evidence. This is exactly what Mayo alludes to in her Severity Principle, which we quote again for convenience:

> **Severity Principle (strong):** We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of findings flaws or discrepancies from C, and yet none or few are found, the passing result, x, is evidence for C.

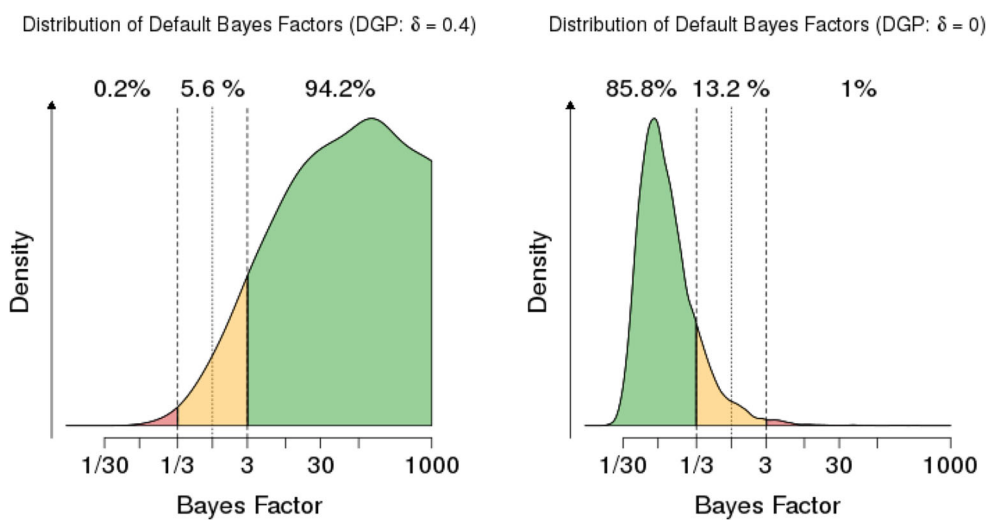

Fig. 9 Graphical illustration of the distribution of Bayes factors and probability of misleading evidence in the Bayesian framework for N=190, a two-sided t-test and hypothesized effects of $d = 0$ and $d = 0.4$. Figure produced with the BFDA app https://shinyapps.org/apps/BFDA/ based on Schönbrodt and Wagenmakers (2018)

Reinterpreting Mayo from a Bayesian viewpoint, we have evidence for the claim *C* if and only if (a) we observe a Bayes factor in favor of *C* beyond a context-sensitive threshold, and (b) the probability of finding misleading evidence for *C* (under the assumption that *C*'s competitor is true) is low. The Shiny apps and the R package for a Bayes Factor Design Analysis make the implementation of this planning explicit and easy to handle for an experimenter. There is thus no reason why a Bayesian needs to give up an appreciation of severity in terms of error control. What is more, in the light of the challenges inherent to Mayo's own operationalization of her Severity Principle, one could even draw the conclusion that Bayesian inference is suited best to implement her philosophical stance into experimental practice.

Discussion

This article presents a Bayesian proposal on how to accommodate the concept of severity when testing statistical hypotheses. It provides a translation of Popper's (1959/2002) falsificationist philosophy and the intuitive impressiveness of risky predictions in terms of a test of specific hypotheses. The specificity of a hypothesis is defined by the degree to which predictions are spread out across the sample space, while evidence is defined by relative predictive performance, that is, how much probability mass is allocated to the data under the competing hypotheses (i.e., the Bayes factor).

A complex or vague hypothesis spreads out its predictive mass across a wide range of options, and by hedging its bets will lose out against a more restrictive hypothesis that makes a more precise (and accurate) prediction. The more precise the predictions from the competing hypotheses, the higher the expected diagnosticity, and the more severe the test. This approach clearly quantifies Popper's idea of evaluating theories on the basis of their empirical content and degree of falsifiablity (Popper, 1959/2002). In the specific case of a parameter space whose regions correspond to different statistical hypotheses, the specificity of a hypothesis can often be measured by the proportion of the parameter space it occupies (weighted with the prior probability density; cf. Appendix).

Popper and Bayes can thus be reconciled: the evaluation of hypotheses in terms of Bayes factors is influenced by their specificity and Bayesian inference has the conceptual resources to reward specific predictions. Notably, obtaining this conclusion does *not* require any blending of Bayesian and frequentist inference; our account stays faithful to the principles of subjective Bayesianism. While error statistics postulates severe testing as a key virtue of statistical inference, our approach also *explains* why severity matters for conducting efficient hypothesis tests and making good decisions.

Advantages of specific hypothesis testing

There are several clear advantages to testing specific hypotheses in a Bayesian framework. First and foremost, it is to the researchers' benefit to make more specific predictions. As outlined earlier, the more specific the hypothesis is, the more evidence one can expect: either more evidence is obtained from the same amount of data, or less data are needed for the same amount of evidence. This is in stark contrast with orthodox frequentist methods where "more evidence" requires more data if the discrepancy between null hypothesis and alternative hypothesis is fixed. In experimental psychology, demonstrations of the benefit of informed/specific Bayes factor hypothesis tests includes Vohs et al. (in press), Gronau et al. (2017), Ly et al. (2019) and in particular the work of Zoltan Dienes (e.g., Dienes, 2008; 2011; 2014; 2016; 2019). For cognitive models, the value of informed prior distributions has been highlighted by Vanpaemel and Lee (2012) and Lee and Vanpaemel (2018).[15] All in all, unlike frequentists, Bayesian can directly integrate theoretical expectations into the premises of a statistical inference and use them to test hypotheses severely.

Second, our account makes explicit why vague predictions lack diagnostic value. This is most clearly expressed in cases where (almost) everything is consistent with the theory. In such cases where all parameter values are allowed by the theory, one can never expect to obtain strong evidence either in favor or against the theory. Only when the theory is restrictive, strong confirming or undermining evidence can plausibly be expected. Consequently, evidence is limited by the specificity of the hypothesis. Intuitively, the possible evidence in favor of a theory should be limited by its strength and how well it is tested. The specific hypothesis testing approach provides this feature, because the maximum amount of evidence in terms of the Bayes factor is one divided by the specificity of the hypothesis.

Third, our approach extends the treatment of severity by Vanpaemel (2019; 2020), who argues that predictions are risky, and tests severe, only if the tested hypothesis rules out plausible outcomes a priori. Vanpaemel's approach is so far qualitative, and the exclusion of plausible outcomes can be

[15] Vohs et al. (in press) assign a positive-only Gaussian prior for effect size with mean .30 and standard deviation .15 – the Vohs prior. Gronau et al. (2017) assign to effect size a positive-only *t*-distribution with location 0.350, scale 0.102, and three degrees of freedom – the Oosterwijk prior, see (Gronau et al., 2020). Finally, Ly et al. (2019) outline a test to assess replication success, where the posterior distribution of the original experiment functions as the prior distribution of the replication experiment; see also Verhagen and Wagenmakers (2014).

taken as a minimal requirement of a severe test as we did in the example discussed in the previous section.

Fourth, our approach to severity can be extended to the research planning phase. Specifically, the specificity of one's hypothesis, or the restrictiveness of one's theory, can inform the researcher about the sample size required to be fairly certain of strongly supporting or contradicting results. Vice versa, in the case where the sample size is predetermined, the researcher can infer how specific the tested hypothesis needs to be in order to be fairly certain of obtaining high degrees of confirmation or disconfirmation. Explicitly, our approach could be integrated into the Bayes Factor Design Analysis discussed in the previous section (BFDA; SchÖnbrodt & Wagenmakers, 2018; Stefan et al., 2019). Given background assumptions concerning the rival models and the data-generating process, BFDA provides either the distribution of Bayes factors for a fixed sample size, or the distribution of sample sizes to obtain a fixed Bayes factor. The use of more specific models will affect the BFDA such that the same data will become more diagnostic: a fixed sample size is projected to yield more evidence, and a given level of evidence is reached with smaller samples.

Similarly, in Adaptive Design Optimization (ADO; e.g., Ahn et al., 2020; Myung et al., 2013) the next stimulus to be presented is determined by maximizing expected information gain. As the observations accumulate, the rival hypotheses become increasingly specific (i.e., their constituent posterior distributions become more peaked), and therefore easier to discriminate. Our research thus connects to the growing literature that considers Bayesian inference not only a method of hypothesis evaluation, but also uses it in planning and designing experiments that are both reliable and efficient.

Disclaimers and conclusion

Our approach is susceptible to any objection that may be raised against Bayesian statistics as a whole (e.g., Moyé, 2008; Senn, 2011; Mayo, 2010; 2018). Our claim here is just that the epistemic value of severity is not a compelling argument for preferring frequentist to Bayesian statistics. Severity is a notoriously elusive and hard-to-quantify concept in the frequentist paradigm, too, and the state-of-the-art explications are arguably unsatisfactory. At the same time, it is possible to give a Bayesian account of the evidential value of severity. We hope that our approach provides a useful tool that can fill an empty spot in the researcher's toolbox of statistical methods, specifically with regard to the role of severe testing Bayesian inference. At the very least, we expect that this paper will inspire criticism, which might stimulate fruitful debate on how theories can show their probative value irrespective of the preferred statistical paradigm.

Appendix: specific hypothesis testing with encompassing priors

The Bayesian *informative hypothesis testing* approach (e.g., Klugkist et al., 2005; Klugkist & Hoijtink, 2005) can be used to illustrate how Bayesian inference incorporates the concept of severity as the ability of an experimental design to discriminate the hypotheses at hand. Within this approach, each hypothesis to be tested is derived from a more general *encompassing* model. This model consists of all pertinent parameters free from any constraints (e.g., $H_e : \mu_a, \ \mu_b$); all specific hypotheses are nested in this model (e.g., $H_0 : \mu_a = \mu_b$, $H_1 : \mu_a > \mu_b$, $H_2 : \mu_a < \mu_b$). Only a prior for the parameters in the encompassing model needs to be specified and the priors from the nested hypotheses follow from this encompassing prior. Specifically, the hypotheses occupy particular sections of the parameter space, allotting each hypothesis a segment of the encompassing prior (for further explanation of the encompassing prior, see Klugkist & Hoijtink, 2005).

We can denote the encompassing model as H_e and the encompassing prior as $g(\boldsymbol{\theta} \mid H_e)$, where $\boldsymbol{\theta}$ is the vector of parameters of interest. The prior distribution of any hypothesis H_i nested in H_e can be obtained from the encompassing prior by restricting the parameter space according to the limits of the hypothesis, given by

$$g(\boldsymbol{\theta} \mid H_i) = \frac{g(\boldsymbol{\theta} \mid H_e) I_{H_i}(\boldsymbol{\theta})}{\int g(\boldsymbol{\theta} \mid H_e) I_{H_i}(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Here, $I_{H_i}(\boldsymbol{\theta})$ is an indicator function which equals 1 if the parameter value is within the limits of H_i and 0 if it is outside the limits of H_i.

The evaluation of the hypothesis is based on the Bayes factor. Specifically, support provided by the data for one versus another hypothesis is quantified as the ratio of marginal likelihoods of hypotheses (Kass & Raftery, 1995). For data x and hypotheses H_i and H_e, the Bayes factor (BF) is

$$\mathrm{BF}_{ie} = \frac{p(x \mid H_i)}{p(x \mid H_e)},$$

that is, the quotient of the marginal likelihoods for the hypotheses H_i and H_e. Due to Bayes' Theorem

$$p(\boldsymbol{\theta} \mid x, H) = \frac{p(x \mid \boldsymbol{\theta}, H) g(\boldsymbol{\theta} \mid H)}{p(x \mid H)},$$

the marginal likelihood can, for any value of the unknown parameter $\boldsymbol{\theta}$ consistent with H, also be written as

$$p(x \mid H) = \frac{p(x \mid \boldsymbol{\theta}, H) g(\boldsymbol{\theta} \mid H)}{p(\boldsymbol{\theta} \mid x, H)}.$$

In this formulation, the denominator is the posterior density of $\boldsymbol{\theta}$ under model H_i and the numerator is the product of the likelihood function and the prior distribution.

As a consequence, the Bayes factor of H_i and H_e can be written as

$$\mathrm{BF}_{ie} = \frac{p(x \mid H_i)}{p(x \mid H_e)} = \frac{p(x \mid \boldsymbol{\theta}, H_i)g(\boldsymbol{\theta} \mid H_i)/p(\boldsymbol{\theta} \mid x, H_i)}{p(x \mid \boldsymbol{\theta}, H_e)g(\boldsymbol{\theta} \mid H_e)/p(\boldsymbol{\theta} \mid x, H_e)}. \tag{2}$$

If we then suppose, more specifically, that the value $\boldsymbol{\theta}$ is allowed by H_i, (2) delivers

$$\mathrm{BF}_{ie} = \frac{g(\boldsymbol{\theta} \mid H_i)p(\boldsymbol{\theta} \mid x, H_e)}{g(\boldsymbol{\theta} \mid H_e)p(\boldsymbol{\theta} \mid x, H_i)}.$$

Through H_i being nested in H_e, the densities $g(\boldsymbol{\theta} \mid H_i)$ and $p(\boldsymbol{\theta} \mid x, H_i)$ can be rewritten as follows:

$$\begin{aligned} g(\boldsymbol{\theta} \mid H_i) &= c_i \times g(\boldsymbol{\theta} \mid H_e) \\ p(\boldsymbol{\theta} \mid x, H_i) &= f_i \times p(\boldsymbol{\theta} \mid x, H_e), \end{aligned}$$

where c_i and f_i are constants. Thus, H_i's prior and posterior densities can be given in terms of the densities of the encompassing model H_e. $1/c_i$ denotes how much of the prior distribution of H_i is in agreement with the encompassing model H_e, and $1/f_i$ denotes the same quantity for the posterior distribution. $1/c_i$ is therefore a natural measure of the complexity of the nested model H_i, $1/f_i$ of its post hoc fit. Since H_i is nested within H_e, we can write the Bayes factor simply as

$$\mathrm{BF}_{ie} = \frac{1/f_i}{1/c_i} = \frac{\text{fit}}{\text{complexity}}$$

From this specification of the Bayes factor, it follows that good model fit is not enough as evidence for one's hypothesis. Near perfect fit is easily reached when close to all possible parameter values and functional forms are allowed. Only when one's hypothesis is restrictive enough that it occupies only a small fraction of possible parameter values and functional forms a priori, though still is in accordance with the data a posteriori, does one have evidence in favor of the hypothesis under consideration when supporting observations are made.

Acknowledgements We would like to thank Riet van Bork and Fabian Dablander for their helpful comments on earlier versions of the paper.

Author Contributions The authors have contributed equally to the manuscript.

Funding The research by NvD and JS was supported by Starting Investigator Grant No. 640638 ("OBJECTIVITY—Making Scientific Inferences More Objective") of the European Research Council (ERC). The research by EJW was supported by ERC Advanced Grant No. 743086 ("UNIFY—A Unified Framework for the Assessment and Application of Cognitive Models") and Vici Grant No. 016.Vici.170.083 ("Monitoring Evidential Flow: New Bayesian Methods for Medicine and Psychology") by the Netherlands Organization for Scientific Research (NWO).

Code Availability The *R* code for calculating the Bayes factors is available at https://osf.io/3cdyx/.

Declarations

Conflict of Interests There are no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahn, W.-Y., Gu, H., Shen, Y., Haines, N., Hahn, H. A., Teater, J. E., . . . , Pitt, M. A. (2020). Rapid, precise, and reliable measurement of delay discounting using a Bayesian learning algorithm. *Scientific Reports*, *10*, 12091.

Berger, J. O., & Wolpert, R. L. (1984). *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, *53*, 259–326.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

de Finetti, B. (1974). *Theory of probability*. New York: Wiley.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave MacMillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psycholology*, *5*, 781.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psycholology*, *72*, 78–89.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, *2*, 364–377.

Dienes, Z. (2021). Testing theories with Bayes factors. Accessed on 9 June 2021.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of may 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *220*(571–581), 291–333.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1*, 281–295.

Evans, M. (2015). *Measuring statistical evidence using relative belief*. Boca Raton: CRC Press.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Human Genetics*, *6*, 391–398.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. New York: Hafner.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

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

Good, I. J. (1975). Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, *30*, 39–73.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, *66*, 393–396.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t*-tests. *The American Statistician*, *74*, 137–143.

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138.

Haig, B. D. (2020). What can psychology's statistics reformers learn from the error-statistical perspective? *Methods in Psychology*, *2*, 100020.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: Chapman & Hall/CRC.

Horwich, P. (1982). *Probability and evidence*. Cambridge: Cambridge University Press.

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach, 3rd edn.* Chicago: Open Court.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Jeffrey, R. C. (1971). The Logic of Decision, (2nd ed.). Chicago and London: University of Chicago Press 1983.

Jeffreys, H. (1961). *Theory of probability*, (3rd ed.). Oxford: Oxford University Press.

Jeffreys, H. (1973). *Scientific inference*, (3rd ed.). Cambridge: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Klugkist, I., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*(4), 477–493.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*, 270–280.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta–analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178–206.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge: Cambridge University Press.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986–1005.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–337.

Lindley, D. V. (2006). *Understanding uncertainty*. Hoboken: Wiley.

Lohne, J. A. (1968). Experimentum crucis. *Notes and Records of the Royal Society of London*, *23*(2), 169–199.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*, 2498–2508.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In Mayo, D. G., & Spanos, A. (Eds.) *Error and inference: Recent exchanges on experimental reasoning, reliability and the objectivity and rationality of science, chapter 3*, (pp. 305–314). Cambridge: Cambridge University Press.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the science wars*. Cambridge: Cambridge University Press.

Mayo, D. G., & Kruse, M. (2001). Principles of inference and their consequences. In Cornfield, D., & Williamson, J. (Eds.) *Foundations of Bayesianism*, (pp. 381–403). Dordrecht: Kluwer Academic Publishers.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(1), 806–834.

Meehl, P. E. (1986). What social scientists don't understand. In Fiske, D. W., & Shweder, R. A. (Eds.) *Metatheory in social science: Pluralisms and subjectivities, chapter 14*, (pp. 315–338). Chicago: University of Chicago Press.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Meehl, P. E. (2005). Cliometric metatheory II: Criteria scientists use in theory appraisal and why it is rational to do so. *Psychological Reports*, *91*(6), 339–404.

Monson, C. M., Schnurr, P. P., Resick, P. A., Friedman, M. J., Young-Xu, Y., & Stevens, S. P. (2006). Cognitive processing therapy for veterans with military-related posttraumatic stress disorder. *Journal of Consulting and clinical Psychology*, *74*(5), 898–907.

Morey, R. (2020). Severity demonstration. Accessed on 27 August 2020. https://richarddmorey.shinyapps.io/severity.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Moyé, L. A. (2008). Bayesians in clinical trials: Asleep at the switch. *Statistics in Medicine*, *27*, 469–482.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of Mathematical Psychology*, *57*(3-4), 53–67.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*(3), 499–518.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131.

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

Neyman, J., & Pearson, E. S. (1967). *Joint statistical papers*. Berkeley: University of California Press.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference, 2nd edn*. London: Arnold.

Peirce, C. S. (1931). *The collected papers of Charles Sanders Peirce* Vol. I–VI. Cambridge: Harvard University Press.

Pérez, J. M., & Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89.

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353.

Poincaré, H. (1913). *The foundations of science (G. B. Halsted Trans.)* New York: The Science Press.

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

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Royall, R. (2000). On the probability of observing misleading statistical evidence (with discussion). *Journal of the American Statistical Association*, *95*, 760–780.

Savage, L. J. (1972). *The foundations of statistics*, (2nd ed.). New York: Wiley. Originally published in 1954.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Senn, S. (2011). You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets and Morals*, *2*, 48–66.

Sessa, B. (2017). MDMA and PTSD treatment: "PTSD: from novel pathophysiology to innovative therapeutics". *Neuroscience Letters*, *649*, 176–180.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford: Oxford University Press.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior research methods*, *51*(3), 1042–1058.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2019). The really risky registered modeling report: Incentivizing strong tests and honest modeling in cognitive science. *Computational Brain & Behavior*, *2*(3), 218–222.

Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, *127*, 136–145.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A., Wagenmakers, E.-J., & Albarracín, D. (in press). A multi–site preregistered paradigmatic test of the ego depletion effect. Psychological Science.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.