Obtaining evidence for no effect

Article (Accepted Version)

Obtaining evidence for no effect

Zoltan Dienes

School of Psychology, University of Sussex, Brighton, UK

dienes@sussex.ac.uk

Abstract

Obtaining evidence that something does not exist requires knowing how big it would be were it to exist. Testing a theory that predicts an effect thus entails specifying the range of effect sizes consistent with the theory, in order to know when the evidence counts against the theory. Indeed, a theoretically relevant effect size must be specified for power calculations, equivalence testing, and Bayes factors in order that the inferential statistics test the theory. Specifying relevant effect sizes for power, or the equivalence region for equivalence testing, or the scale factor for Bayes factors, is necessary for many journal formats, such as registered reports, and should be necessary for all articles that use hypothesis testing. Yet there is little systematic advice on how to approach this problem. This article offers some principles and practical advice for specifying theoretically relevant effect sizes for hypothesis testing.

Introduction

If there is no way for data from a study to count against a theory, the putative test of the theory is not a test at all. When a research program continues to operate under conditions where it is impossible to get evidence against at least one theory motivated by the program, then we are not doing science (Lakatos, 1978; Popper, 1963). Thus, this article examines what is needed to have results count against a theory claiming an effect. The use of non-significance to indicate no effect is shown to be misleading. An overview is provided of how to think about the problem, using well established tools including Bayes factors or inference by intervals. This paper addresses what might be called the pragmatics of statistical inference: How theory relates to inferential statistics in order to severely test a theory. Statisticians often work on the mathematics of statistics; scientists often work on the substantial area of science they investigate. The pragmatics of linking the two domains is thus often ignored, although it is a crucial link in the chain of reasoning. How does statistical hypothesis testing test a scientific theory? In a simple case, a theory predicting an effect in one direction can be falsified by finding an effect in the other direction. But there may not be any meaningful effect at all.

How can we get evidence that something does not exist? We can look for it. Consider a time in the future when the snow leopard is regarded as extinct in the Himalayas. We have a quick poke around one afternoon and do not find a snow leopard. But that would scarcely constitute good evidence that there was not one left in the Himalayas. Simply looking and not finding is not good enough in itself to provide evidence of something not existing. Yet that is what people do when they declare the absence of an effect based only on a non-significant result.

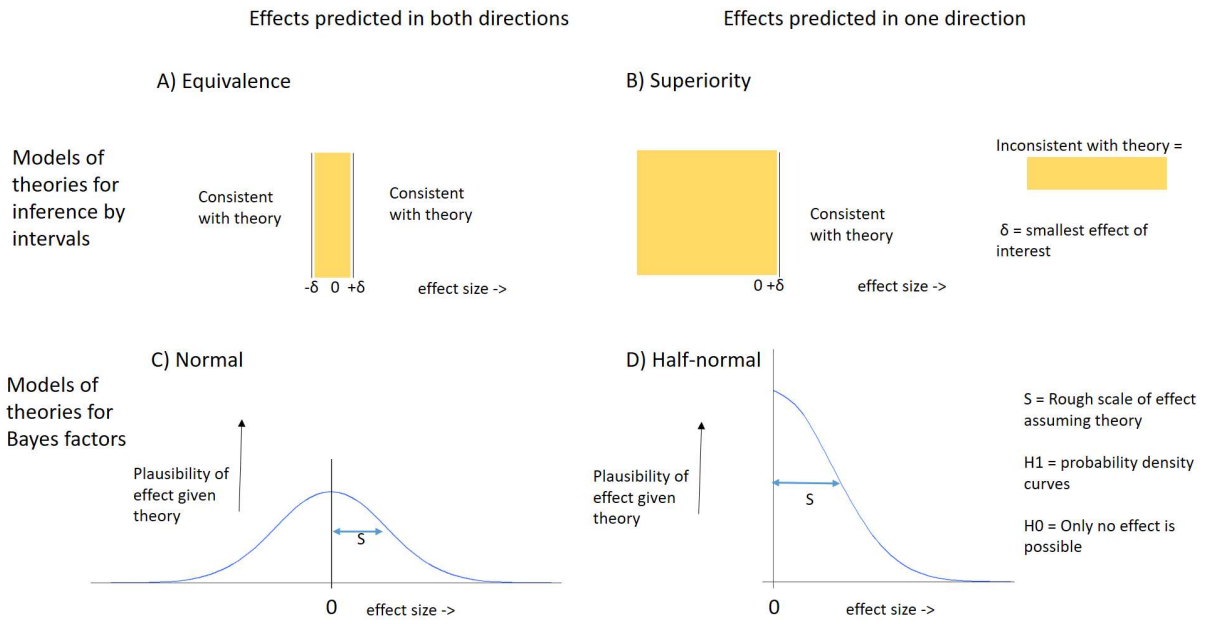
One needs to look closely enough. How closely? That depends on how big the thing is that one is looking for. If asked “Is there an animal in the room now?”, the strength of evidence against there being an animal present depends on what animal it may be. If a rhino, you have good evidence there is not one, based on glancing around. If an ant, you have no evidence one way or the other just by glancing around. One can only quantify the evidence for something not being there if given the size of the thing looked for. That is a truism no matter what philosophy of statistics is used. One can only get evidence for there not being a signal if one knows the size of the possible signal relative to the noise. A p-value is calculated without reference to the possible size of an interesting effect. Therefore, a p-value (using the H0 of no effect) can never provide evidence for an effect not being there. And by the same token, since what makes an effect interesting is relative to a particular scientific context – are we investigating ants or rhinos? - neither can evidence for no theoretically interesting effect be provided by AIC, BIC, a default Bayes factor, or any other measure that does not involve specifying, for scientific reasons, what effect is of interest. (For a defence of AIC for

other reasons see Burnham & Anderson, 2004; a defence of BIC, Raftery, 1999; and a defence of default Bayes factors, Ly, Verhagen, & Wagenmakers, 2016.)

There turn out to be two ways of specifying an effect of interest. First, one can specify the smallest effect that is plausible and just interesting for theoretical or practical reasons. For example, if considering an intervention to lose weight, any effect less than 0.5 kg over one month might be regarded as too small to be of clinical interest. A "raw effect size" is an effect in a unit of measurement, as in this case, kg. That is typically what is of practical or theoretical interest. A person interested in losing weight is not interested in how much noise is in the scales measuring them (the noise affects standardized effect sizes); just in the kg's lost over a period of time (Ziliak & McCloskey, 2008). The smallest effect of interest is what is needed for power (Dienes, 2008), for equivalence testing or inference by intervals in its various Bayesian (Freedman, & Spiegelhalter, 1983; Greenwald, 1975 Krushcke, 2014; Kruschke & Lidell, 2018) or frequentist forms (Lakens et al, 2018; Meyners, 2012; Rogers, Howard, & Vessey, 1993; Westlake, 1972); or for very similar approaches such as frequentist severity testing (Mayo, 2018). Second, one can specify the rough scale of effect with a plausible range around it. For example, if a previous intervention, similar to the current one according to a theory, produced a loss of 3 kg over a month, one might expect the current intervention to have about the same effect, 3 kg (or at least in the range of, say, 0 to 6kg). The rough scale of effect is what is needed for Bayes factors (Dienes, 2014; Jeffreys, 1939; Kass & Raftery, 1995; Kass & Wasserman, 1996; Rouder, Morey, Verhagen, Province, et al., 2016; Rouder, Morey, & Wagenmakers, 2016; Vanpaemel & Lee, 2012). Figure 1 illustrates the two different approaches.

Figure 1

Different models for hypothesis testing



The different models for testing a theory predicting a difference via a smallest effect of interest (A and B), using inference by intervals (e.g. equivalence testing); or a rough scale of effect (C and D), using Bayes factors. The scale of effect, S, implies a rough range of plausible effects of -2S to +2S in C, and of 0 to +2S in D. The equivalence region in A is typically symmetric though need not be. Note for every test two, models are specified, one for H1 and one for H0 (in the diagram, an interval H0 for inference by intervals and a point H0 for the Bayes factors).

Because the inference that there is no effect depends crucially on the possible size of the effect, that size should be specified for objective reasons: That is, the numerical value for the effect should come from a public place one can point to (for example, data), so that other people can criticize the reason for choosing that value. Numbers pulled out of the air ("a Cohen's d of 0.5 because it is medium") are not based on reasons for a particular scientific situation (e.g. Funder & Ozer, 2019). A committee deciding on a smallest effect of interest is useful in so far as the committee gives its reasons so they can be criticized and the estimate improved. Thus, black boxes such as committees, or an expert's opinion, push back the problem of what would constitute a good reason.

One can evade the problem by only reporting p-values and not interpreting non-significant results as evidence for H0. But then one can never get evidence against a theory that predicted a bidirectional effect. A large p-value in a test with H0 of no effect can never in itself provide evidence against a theory predicting an effect. And if we cannot get evidence against a theory, we would not be doing science. So what is urgently needed is an account of what constitutes objective reasons for specifying a minimally interesting effect size or a rough scale of effect expected. This paper will present some ideas for solving this problem (see also Dienes, 2021a). First, we will introduce methods that use the smallest effect of interest (power, or variants of inference by intervals) and then provide examples of heuristics for specifying a smallest effect size of interest. Then we will introduce Bayes factors and then provide examples of heuristics for specifying a rough expected scale of effect. Specifying effect sizes allows hypothesis testing to test theories. After each of these sections the problem of making the test a severe test of a theory will be discussed.

Smallest effect size of interest.

When using equivalence testing or severity testing, after data are in one may determine if the obtained effect is smaller than specified values (Lakens, Scheel et al, 2018), or is plausibly contained within tight enough bounds (Greenwald, 1975; Kruschke & Lidell, 2018). That is, one defines a null interval, a region within which the effect is too small to be interesting. The result would only relate to theory or practice if the bounds were those specified as just interesting according to the theoretical or practical concerns. Having made a commitment to a null interval H0, consistency demands that one rejects H0, only when one can reject the whole null interval by one's decision procedure (Greenwald, 1975; Dienes, 2014; Kruschke, 2014). That is, having defined an interval H0, an effect significantly different from the point 0 does not entail rejecting the interval H0. One only rejects the interval H0 when the effect is significantly larger in magnitude than the minimally interesting effect. In sum, using an interval H0 is a different procedure than typical significance testing even in the case of accepting (or getting evidence for) H1.

In frequentist (Neyman-Pearson) equivalence testing, one uses a one-tailed α % significance test of the sample mean (or other parameter) against the upper bound of the null interval (is the mean lower than the upper bound of the equivalence region, the name of the null interval in equivalence testing?); and another such test against the lower bound (is the mean higher than the lower bound of the null interval?). If both tests are significant, the mean is asserted as being within the null interval (equivalence region) (Lakens et al, 2018). So long as the α level (significance level) of each test is the same as the other, this procedure amounts to determining if a $(1 - 2 \times \alpha)$ confidence interval is within with null interval (Berger & Hsu, 1996). For example, if the two tests

are at the one-tailed 5% level, equivalence can be asserted if the 90% confidence interval lies within the null interval. Similarly, equivalence can be rejected if the 90% CI lies outside the null interval[1]. Kruschke (2014) defines a very similar Bayesian procedure: Determine if the Bayesian credibility interval (i.e. highest density region) is inside or outside of a null interval (which he calls the Region Of Practical Equivalence, ROPE). Similarly, Mayo's (2018) error statistics method is to determine what set of parameter values have been severely tested (i.e. for her, can be rejected at the α % level), and see if the remaining values are sufficiently small to be theoretically uninteresting. Despite radically different philosophies, these Neyman-Pearson, Bayesian and error statistical procedures can be very similar practically: They all revolve around the relation of post-data implausible or rejected values to a null interval.

When calculating power in advance of collecting data, in order to specify a stopping rule (for a fixed number of subjects, or a sequential design, Lakens, 2014), the aim is to design a decision procedure that if repeatedly used, would not often miss effects that are of interest (Anderson, Kelley, & Maxwell, 2017). If power were calculated with respect to an effect found in a previous study or a pilot, the Type II error rate is controlled with respect to that effect; but not with respect to any smaller effect. But surely if a previous study found an effect of 200 ms, one would still find a smaller effect, say 150 ms, interesting. So power calculated with respect to a previous obtained effect size does not control error rates with respect to all interesting effects that are plausible (cf Gelman & Carlin, 2014). Thus, a study powered with respect to the mean effect of a previous study, if non-significant, would not count against a theory that predicted an effect. In order for a study to be compelling in itself, power must be calculated with respect to the smallest effect that is plausible and still just interesting for theory or application (Dienes, 2008; see also Albers & Lakens, 2018). We now consider a non-exhaustive list of ways for obtaining a smallest effect of interest.

Heuristics for obtaining a smallest effect size of interest.

i) The judgment of an end user. In applied research what matters is whether the outcome is good enough for the end user (cf King, 2011; Lakens, McLatchie et al., 2020; also see Dienes, 2021a, for further examples). Consider the example of detecting deception. Often in an interview (by e.g. police, human resources, or airport security) it is useful to determine if a person is telling the truth

[1] Using an X% CI for both asserting H0 and rejecting H0 means conclusions follow from a simple rule: Is the X% CI inside or outside the null region? However, from the perspective of Type I error rates, a 90% CI involves a familywise error rate of 5% for the two tests that allow the assertion of equivalence; and of 10% for the one test that allows the assertion of superiority. The leniency of the latter might be thought of as a good tradeoff for testing against a minimally interesting effect size rather than against 0.

or not. Sandham et al. (2020) instructed a police interviewer to use two interview methods: The interviewer holding back crucial evidence (call it the hold back method), or drip feeding known evidence throughout the interview (drip feed method). Interviewees had previously been instructed to lie or tell the truth 50% of the time. Thirty police observers watched the videos of the interview and judged whether the interviewee was truth telling or deceptive. The percentage accuracy for the hold back method was 51%, 90% CI [46%, 56%], and for the drip feed method 68%, 90%, CI [62%, 74%] (90% CIs estimated only approximately from the p value limits given in the paper). Whether or not the hold back method can be asserted to provide no meaningful accuracy above chance (50%) depends on what a minimal interesting effect size would be in the context of a police investigation. And presumably this is for police to decide, given their needs. The same applies for concluding whether the drip feed method can be asserted to provide meaningful accuracy; that would depend on whether 62% is greater than a minimally interesting effect size from a police perspective. For example, if hypothetically 60% was a minimal interesting effect size as elicited by interviewing police about what accuracy would be just useful, the hold back method could be asserted to provide no meaningful accuracy, and the drip feed method could be asserted to provide meaningful accuracy.

Based on results such as those of Sandham et al. (2020), and others from the same research group, Diane Sweeney at the University of Sussex wished to establish interviewing methods for determining honesty in the context of job recruitment. In order to obtain a null interval, Sweeney sought to obtain a minimally interesting effect size by asking the end user. Thus, Sweeney (personal communication, 4 Jan 2021) in an initial pilot, asked eight managers or company owners who were responsible for hiring people, “With very few exceptions, deception research and meta-analyses of the last 50+ years have failed to discriminate truth-tellers from liars much above a chance level of 50%. With that in mind, what is the minimum effect that would make it worthwhile changing current interview practice?” The mean was 74% with a range of 60-80%. Did the respondents really take into account the objective costs of failing to detect deception (e.g. of taking on someone who did not have the claimed expertise) and thus the benefits of even a small increase in accuracy given a small change in interview practice? Sweeney is following up with a more thorough questionnaire. Eliciting plausible or interesting effects from experts can take some care (cf. O’Hagan et al., 2006).

There will always be a spread of opinion amongst end users, as shown in this example. In interviewing cancer clinicians, Freedman and Spiegelhalter (1983) noted a wide range in elicited minimal effects of interest for a putative new cancer treatment. The simplest way of dealing with such variability is to use the mean elicited minimal effect as a single estimate of the minimal effect. Freedman and Spiegelhalter recommended using a grey interval between the interval H0 and the interval H1. A grey area allows some flexibility in drawing conclusions: If the CI lies mainly in the

H0 interval and the remaining minority only in the grey interval, one could accept H0; similarly, if the CI lies mainly in the H1 interval and the remaining minority only in the grey interval, accept H1; otherwise more data are needed[2].

While not specifically recommended by Freedman and Spiegelhalter, one could, for example, use the interquartile range of the end user's assessments as the grey interval, the null interval reaching to the bottom of the grey interval and the H1 interval extending from the top of it. Thus for example, if an interquartile range for the stated minimally interesting effect size elicited from recruitment personnel is 55-65% accuracy for detecting deception, the null interval could be [0, 55%] (corresponding to the hypothesis that the interview is not of use), the grey interval [55, 65%], and H1 [65, 100%] (corresponding to the hypothesis that the interview is of use). If the 90% CI were [45, 61%], H0 would be accepted.

ii) Calibration. Dienes (2014; supplemental data- Appendix 1, example 2) showed how one measure, for which we do not have a relevant interesting effect size, can be regressed against another, for which we do, in order to calibrate the former. For (an imaginary) example, a researcher explores different styles of interaction between a therapist and a client, and wishes to examine which makes the client happier without having the client reflecting on their happiness (and maybe thereby just responding to demand characteristics; or interrupting the flow of the interaction). Anvari and Lakens (2019) found that 0.3 Likert units (on a scale from 1 "not at all" to 5 "extremely"; the PANAS scale; Watson et al., 1988) is appreciated by people as a noticeable change in affect. But asking people to rate their affect explicitly is what the clinical researchers wished to avoid. So first they ran a norming study that manipulated mood so that there was a range of happiness; they measured both activation of the muscles responsible for smiling and also rated positive effect. They regressed muscle activation against Likert ratings[3]. The change in muscle activation corresponding to a change of 0.3 units of rated affect can be read off from the raw regression line. This change in muscle activation may be taken as a minimally interesting effect in that it would correspond to a change in positive affect the client appreciated as a change.

Note that the regression assumes that the Likert ratings were measured without error. If the predictor is measured with error, the expected sample regression line is flattened compared to the population regression line that occurs with perfect measurement of the predictor. An estimate of the

[2] In terms of the frequentist properties of such a rule, the rate of falsely asserting equivalence will not be higher than that achieved when the joint H0 interval and grey interval are treated as the equivalence region in equivalence testing. Likewise, the rate of falsely asserting non-equivalence will not be higher than that achieved when the H0 interval is treated as the equivalence region.

[3] Note that the smaller the raw regression slope, the smaller the difference in muscle activations that corresponds to a difference of 0.3 Likert units in ratings. Thus, the more participants would be needed to provide a severe test of a theory predicting a difference in happiness between approaches (see below). Ensuring a severe test motivates researchers to use predictors that correlate highly with the variable they are calibrating against.

population regression line can be obtained if one has an estimate of the measurement error for the predictor (Malejka et al., in press; Matzke et al. 2017). In the case of the10-item positive affect scale of PANAS, the Cronbach's α is 0.9 (Díaz-García et al., 2020), and no correction would be necessary using the PANAS scale as a predictor (see Malejka et al., in press, for effects of correction for different reliabilities). For somewhat smaller reliabilities, see Malejka et al. (in press) for correcting the slope with hierarchical modelling.

iii) Checking the lower limit of a confidence interval is still theoretically relevant. For conceptual and direct replications using the same dependent variable, where there have been past studies, one can look at the lower limit of 95% Confidence Interval of the raw effect (cf. Perugini, Gallucci, & Costantini, 2014). Check if it is it still theoretically or practically interesting. If so, this is the smallest interesting value not rejected by the confidence interval. This may be used as a minimal interesting effect so long as the theory claims the phenomena in the new paradigm is the same as that explored in the previous ones analysed, with no theoretical reasons for why it should be stronger or weaker. For (an imaginary) example, previous research using English and English participants has shown that syntactic incongruity produces a change in a certain ERP component. You wish to replicate in Hungarian using Hungarian participants. Check if the lower limit of the 95% CI for the past research is still interesting. This may be a difficult decision to make. If you have reasons for making such a decision, this heuristic is useful. Let us say the lower limit is 10 μV. The theory explored is that the ERP reflects a more general linguistic incongruity detector. Other research has shown the ERP changes by 8 μV to semantic incongruities; thus, 10 μV is meaningful from the system's point of view, given its postulated theoretical role. To use this heuristic one just has to assess if the lower limit is meaningful; not that it is only just meaningful. Thus, a difficult judgment (only just meaningful) is turned into one slightly less difficult (meaningful) by this heuristic. In the example just described, 10 μV could be used as the smallestinteresting effect that is plausible.

iv) Checking whether modelling assumptions are satisfied. Assumptions for statistical tests are typically explored by simulating the smallest violation that is just consistent with adequate performance of the statistic. That is, assumptions are typically shown to be satisfied well enough when the population violation is smaller than a minimal amount. For example, Flores and Orcana (2018; table 1) indicate what minimal size the ratio of variances for a two-group t-test should be for different scenarios.

Severe testing

Specifying a relevant effect size links the statistical test to a theory; namely the theory to which the effect size is relevant. The substantial theory is the broadest claim that could be falsified by the study (assuming the background assumptions linking theory to predictions are safe). For example, a substantial theory is: drip feeding crucial evidence known only to event participants rather than holding that evidence back increases the capacity of interviewers to detect deception (see above). The statistical hypothesis is a particular instantiation of that theory, that is, with specific dependent and independent variables (and sometimes a named population which is a sub-set of the entire population to which the theory applies). For example, the dependent variable may be percentage correct and a two-alternative forced choice with 50% cases of deception. The independent variable could be two standardized interviewing techniques. A model is a mathematical representation of the statistical hypothesis (e.g. a model of H0 could be the interval [-s, +s] in units of the dependent variable). The smallest effect size of interest instantiates a value in that model (in this case, the bounds of the null interval). It is in this way the theory itself can be tested: The model represents predictions that data can count for or against. If background assumptions are safe, any failure of predictions count against the theory.

In testing theory, the test would ideally be severe; that is, as defined by Popper (1963), if the theory were false, the test is likely to count against the theory. The ideal of a severe test of a theory applies regardless of the philosophy of statistics which is being used (e.g. Dienes, 2008, 2021b; Mayo, 2018; Vanpaemel, 2020). Consider a theory that predicts a difference. The smaller the minimally interesting effect, m, the easier to obtain a 90% CI outside the null interval [-m, +m]; but the harder to obtain a 90% CI inside the null interval. Assume that the decision rule is that the null interval hypothesis will be accepted if the 90% CI lies within the null interval. If the theory predicts a difference, a test of the theory is only severe if, given H0, the probability of obtaining a 90% CI in the null interval is high. Thus, the null interval must be wide enough to allow severe testing. For example, in planning a study one may work out a sample size such that if the null region hypothesis were true, with the number of participants used, at least 90% of the time the 90% CI would fall within the null region.

The rough scale of effect predicted

A Bayes factor compares how probable the data are on one model (e.g. H1) compared to another model (H0) (Jeffreys, 1939; Wagenmakers, Verhagen, Ly, Matzke et al., 2017; see Dienes, 2020; JASP, 2020; Morey, 2021; or Rouder, 2020 for software for calculating Bayes factors). If H1

represents the predictions of a theory, then the Bayes factor measures the evidence for the theory as opposed to H0 (Morey, Romeijn, & Rouder, 2016). The model of H1 is a probability density function indicating how plausible different possible population effects are given the theory. The task of the researcher is to represent theoretical predictions using assumptions that are informed by the scientific context and are also simple. One simple representation is a normal or Cauchy distribution centred on zero (Dienes, 2008; Rouder, Speckman, Sun, Morey, & Iverson, 2009; van Doorn, van den Bergh, Bohm, Dablander et al. 2019). Such a distribution indicates that smaller effect sizes are more likely than larger ones. Peaking the most predicted effect sizes around zero (or chance) may seem strange, but in that the model of H1 then makes similar predictions as the model of H0, it typically becomes harder to distinguish them. And given the exact shape of the distribution will be theoretically arbitrary, making a choice that biases against making a discrimination means that when a discrimination is made, it is despite our choices not because of them. Bear in mind that a Bayes factor can use any models that seem worthwhile to compare (Etz, Haaf, Rouder, & Vandekerckhove, 2018). But for the sake of argument consider a normal distribution centred on zero (Figure 1 C). If the theory predicts a direction, the half of the distribution below zero can be removed (thus, by convention, a positive effect is defined as in the direction predicted by theory) (Jeffreys, 1948; Wagenmakers, 2020) (Figure 1 D). The half-normal distribution requires its standard deviation being set; the standard deviation defines how steeply the curve drops off and thus sets the scale of effect predicted. In sum, the half-normal model of H1 (introduced by Dickey, 1973) assumes that the rough scale of effect is that given by its standard deviation, that smaller effects are more likely than larger ones, and that a rough maximum effect expected is about twice the standard deviation (see Dienes & McLatchie, 2018, for justification of why these assumptions may widely hold). If these assumptions appear to represent one's theory adequately, then the task simplifies to specifying the rough scale of effect. Note the half-normal distribution indicates that the scale is not a point prediction; the true effect could be anything from zero to roughly twice the scale.

Typically, H0 is represented as a point prediction; for example, that of no effect. In that case, no minimal interesting effect need be specified; this can be approximated as close to zero, which will be a good enough approximation if the true minimal interesting effect is smaller than the standard error of the effect (otherwise a null interval can be specified: Morey & Rouder, 2011; Palfi & Dienes, 2019; see Skora et al., 2020, for actual use of a null interval hypothesis). Thus, the Bayes factor shifts the burden from postulating a minimal interesting effect to postulating the scale of effect predicted.

A Bayes factor of 1 means the data were equally well predicted by H1 as H0; thus the data are not evidential and do not discriminate the two models. If the Bayes factor is expressed in terms

of evidence for H1 rather than H0, then numbers between 1 and 0 are progressively more evidence for H0 rather than H1; and numbers between 1 and ∞ are progressively more evidence for H1 rather than H0. If a normal or half-normal distribution with a mode of zero is used to model H1, then when the sample effect is about that predicted, a Bayes factor of 3 roughly corresponds to a significance level of 5% (Jeffreys, 1939); a Bayes factor of 6 to 2% significance; and of 10 to 1% significance, though there is no monotonic relation between Bayes factors and p values (Lindley, 1957; Morey, 2018). The thresholds should serve only as rough benchmarks; the Bayes factor itself is a continuous measure of evidence, with no special bumps at 3, 6, or 10, or anywhere else. Also note that a Bayes factor can indicate evidence for H0 over H1 (by being sufficiently far below 1); a p value, no matter how high, cannot indicate evidence for H0 over H1.

Heuristics for obtaining a predicted scale of effect.

We now consider a non-exhaustive number of ways of deriving a scale of effect relevant to a theory to be tested.

(i) Replication. In attempting to replicate a study, the effect found in the original study can be used as the standard deviation of a half-normal distribution for the model of H1 for the replication attempt (see e.g. Dienes & McLatchie, 2018, for worked examples). Verhagen and Wagenmakers (2014) and Ly, Etz, Marsman, and Wagenmakers (2019) present related methods. The theory tested is that implied by any empirical paper: That the methods of the original study (as given in the Methods section) describe a procedure for obtaining the sort of effect obtained (as reported in the Results section).

A study may be a conceptual replication of an original study. In this case, one may often take the effect from the original study as the scaling factor, just as in the case of the direct replication. The theory being tested is that the phenomena studied in both experiments belong to the same class.

For either a direct or conceptual replication, there is uncertainty in the estimate of the effect of the original study. This is in the first place accounted for by the model of H1 having a spread of plausible population values around the value given as the scale factor. When a half-normal distribution is used in the model of H1, the spread in plausible parameter values is roughly from 0 to twice the scale factor. Further, one can report a Robustness Region, RR, which is the set of scale factors (e.g. SDs of a half-normal) for which the same conclusion holds as achieved by one's chosen scale factor (Dienes, 2019). For example, one may have decided (or a journal decided for you) to accept H1 for $B > 6$, and accept H0 for $B < 1/6$. The original study may have had an estimate of the effect of 60 ms, 95% CI [10, 90ms]. Thus, H1 may be modelled as a half-normal

with SD = 60ms (and a mode of 0 ms), represented as $B_{HN(0,60ms)}$. Your study may obtain a 50ms effect, SE = 21 ms, $B_{HN(0,60\ ms)}$= 8.14, $RR_{B>6}$ = [23 ms, 103 ms]. That is, the scale factor (SD) could be between 23 ms and 103 ms, and the Bayes factor would still exceed 6. Given the 95% CI for the original study, and the spread around the scale factor in any of the models of H1, the robustness region indicates that the conclusion is robust. The conclusion of robustness does not reflect a formal or conventional rule; it is a judgment given available information.

The heuristic of using the effect size of an original study is useful for Bayes factors but not for power or inference by intervals; the effect obtained in a previous study does not in itself provide a minimally interesting effect size. A true population effect smaller than an original effect is typically theoretically very interesting.

(ii) Basic effect heuristic. When investigating whether an intervention moderates an effect, the size of the basic effect itself may be judged to provide a scale appropriate for expecting how much the effect could be altered (Dienes, 2019). In an experiment investigating the effect of the personality of the psychodynamic therapist on therapeutic outcome, past studies indicating the effect of that therapy can be taken as the basic effect. If there are no past studies estimating a basic effect a pilot study may be run not of the full experiment, but just the basic effect. Having estimated the basic effect, the full intervention study may be run, with the basic effect used as the rough scale of effect for the intervention in modelling H1.

Dong et al. (2019) investigated whether English speakers could learn the four linguistic tones of Mandarin more effectively if they were trained with only one speaker (low variability) or four speakers (high variability). Both groups were tested with novel speakers and, as in the training phase, had to pick one of two pictures that matched a spoken word in which only the tone gave information about the which picture was correct. Analysis was with a logistic mixed effects model. The authors judged no previous research sufficiently similar to motivate a scale factor to model H1 for their new paradigm, for the difference in accuracy for high versus low variability groups. For the contingency table correct vs incorrect by low vs high variability, the dependent variable was the log odds, which is 0 for equal accuracy. The authors argued that the overall mean effect, that is, overall accuracy (as measured by log odds), is a useful reference. Their argument was as follows. The maximum difference between the groups would be obtained if the low variability group were at chance; then the difference between groups would be twice the overall mean. Thus, the overall mean is a useful SD for a half-normal in modelling H1. In other words, the theory tested was that high variability will be superior to low variability, to such an extent that that latter may even be at chance. The difference between groups was a log odds of 0.13, SE = 0.228, $B_{HN(0,\ 1.71)}$ = 0.219,

$RR_{B<1/3}$ [1.11, ∞]. That is, surprisingly, variability in training, under the conditions of the study, may have no effect on learning Chinese tones[4].

The basic effect heuristic provides a measure of the scale of the effect; it does not provide a minimally interesting effect so it is not relevant for power or inference by intervals. Modifications of a basic effect are typically very interesting even if they are smaller than the effect itself.

(iii) Calibration. Regression of one dependent variable, for which a relevant effect size is not known, on another, for which the relevant effect size is known, may be useful in order to determine a relevant effect size for the former (Dienes, 2014; Skora et al., 2020). For example, a researcher may be interested in the effect of a mindfulness intervention on pain reduction (see Lovell & Dienes, 2021, for a closely related example). She wishes to have a plausible control group to take account of expectancy effects. Her theory is that mindfulness has an effect on pain that goes beyond the placebo effect. Thus, the researcher wishes to use a control that accounts for all the placebo effect that may occur. On the hypothesis that the suggested control does fully account for the placebo effect, the pre-treatment expectancy that the control will reduce pain must be the same as the expectancy for the mindfulness intervention itself. What scale of difference in expectation could there be, if expectation differences were to explain at least some of any difference in treatment effect on pain between the active control and the mindfulness intervention?

The answer to that question depends on two quantities. First, an estimate of the effect of the treatment on pain ratings. Second, the expectation of pain relief that corresponds to that treatment effect on pain is needed; that is, we need to calibrate the two quantities against each other[5]. In terms of the first quantity, imagine a previous study, using a pain scale of 0 = ‘‘no pain to 10 = ‘‘the most intense pain imaginable”, finding that pain was reduced from say 5 units in a control group to 3 units in the mindfulness group, that is, by about 2 units compared to the control. In terms of the second quantity, a study may be conducted, or already have been conducted, measuring both actual pain after an analgesic is given, and the expected pain the analgesic will result in, by asking “What pain intensity do you expect?” on the same scale. Using the same scale for both pain (i.e. 0 to 10, labelled as before) and expectation of pain (the same 0 to 10 scale) means one has a straightforward calibration between the two scales. But expectancy may not produce the very amount of pain expected; that is, in a placebo response, pain may shift by less than expected. One may discover the relation empirically by regressing actual pain after treatment on expected pain. The raw regression

[4] The authors conducted additional tests, including some tests conducted pre- and post- training where they were interested in whether there was a greater change in test performance for the high variability than low variability condition (i.e an interaction between test-session and learning condition). Using a similar logic, they informed the estimate of H1 for this interaction based on the main effect of test-session, finding generally similar results.

[5] See https://psyarxiv.com/yc7s5/ version 2 for the same example worked out with real data

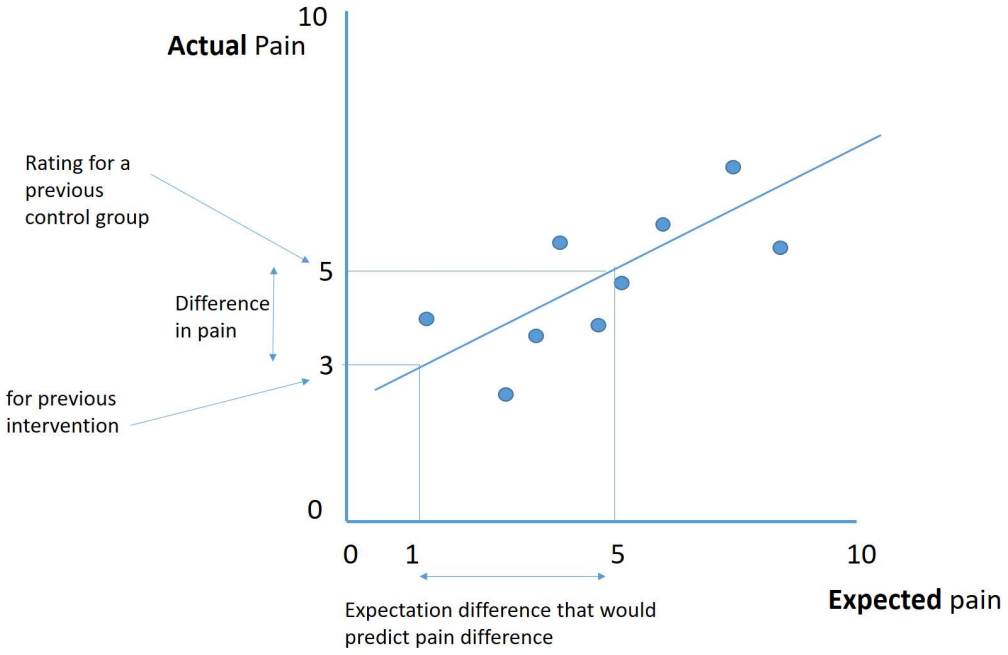
slope of pain against expected pain may, for example, be found to be 0.5 pain units per expectation unit[6]. (See Figure 2.)

Now the two quantities can be combined. As stated earlier, the actual difference in pain produced by the previous mindfulness intervention was 2 units; let us assume the follow up study will induce pain in the same way and use the same mindfulness manipulation, as well as using a plausible control. Given a slope of 0.5 pain units per expectation unit, the corresponding expectation difference is 2 pain units ÷0.5 pain units per expectation unit ≈ 4 expectation units on the 0 to 10 scale. Thus, a model of H1, for the difference in expectation between mindfulness treatment and the control, given the theory that expectation accounts for at least part of any treatment difference between mindfulness treatment and control, could be a half-normal with a scale factor (SD) of 4 expectation units. Note this scale is based on a theory and relevant background information. Conversely, there is no guarantee that a default scale factor would have been relevant. Similarly, a non-significant difference in expectation between the mindfulness and control group would in itself be uninformative.

.

Figure 2

Calibration



[6] Regression assumes there is no measurement error in the predictor, as mentioned before; if the reliability of the predictor is known, a Bayesian correction can be determined (see Malejka et al., 2021).

The plot shows (hypothetical) data of a study by Smith, plotting actual pain after an intervention against what people predicted their pain would be after the intervention (expected pain), where 0 = “no pain” and 10 = “the most intense pain imaginable.” Assume the theory that the difference found between an intervention and a control group arises from differences in expectations (i.e. the theory that the treatment is a placebo). It may have been found by Jones that an intervention has lower pain than a control, as illustrated. You plan to run the same intervention as Jones again, but this time controlling for expectancy differences between the intervention and a plausible control. If you replicate the effectiveness of the intervention Jones found, you predict a 2-unit difference in actual pain between groups. The study by Smith allows you to determine the difference in expectancy that would be needed to explain the 2-unit pain difference: Reading off the graph, a difference of 4 units in expectation predicts a 2-unit difference in actual pain.

In calibrating, which variable should act as predictor and which as the dependent variable? This question is a scientific question and therefore depends on what makes scientific sense (Dienes, 2014). In regressing pain against expectations, as we did earlier, the smaller the relationship, the larger the expectation difference that would be used as a scale factor, and thus the easier it would be find evidence for H0 in testing treatment against control in levels of expectation. If there were no relation between pain and expectation, there would be no need to control for expectation - and the Bayes factor would in fact automatically favour H0[7].

(iv) Heuristics for when there are no prior studies. Dienes (2019) presents a set of heuristics for setting predicted scales of effects when there are no past studies, for ANOVA, regression and mediation designs (see also Dong et al., 2019, just discussed; also Gallistel, 2009). For example, the ratio-of-means heuristic can be used for deriving a rough regression slope expected, given a theory that two variables will go to zero together. For example, in an fMRI study, a theory might predict that blood flow difference between a learning and control condition will correlate with the perceptual discriminability of one of the learning stimuli (e.g. the theory claims that one stimulus

[7] In this case, the theory predicts equivalence, and a severe test would need to readily find evidence for H1 if H1 were true. This requirement motivates using an expectancy rating with a strong relation to pain (cf. footnote 2: Calibration is the other way round, but so is whether the theory predicts H0 or H1). The greater the slope, the smaller the expectation corresponding to a given amount of pain reduction, and hence the smaller the scale factor of the half-normal distribution used for modeling H1. When the scale factor is infinite, there will be infinite evidence for H0. As the scale factor comes down to the magnitude of the standard error, the evidence for H1 increases for any actual difference bigger than the standard error.

was responsible for learning and not another one.) In predicting the slope of the regression of the contrast against discriminability, take as a given the point (mean contrast, mean discriminability). If the theory were perfectly true, the theory predicts another point: (0,0). The line between these two points is the theoretically predicted regression line of contrast against discriminability, from which the predicted slope can be obtained (and used in the model of H1 for testing the existence of the regression slope).

Severe testing

Popper (1963, p 526) formalized the notion of a severe test of a theory by representing it as the ratio of the probability of the predicted outcome given the theory to the probability of the outcome assuming the theory were false (and assuming the rest of background knowledge). That is, a severe test is one that can generate an extreme Bayes factor (i.e. one that is very large or very small). Thus, for a test to be severe using Bayes factors, enough participants must be run to allow an extreme Bayes factor to occur. Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017) and Schönbrodt & Wagenmakers (2018) indicate how one can simulate how many participants are needed in order to achieve, for example, a 90% probability of obtaining a certain Bayes factor threshold (e.g. either greater than 6 or less than 1/6). Palfi and Dienes (2019; version 3 Table 2) provide very quick heuristics for also working this probability out, without needing simulations. In both cases, one uses the expected scale of effect, not the minimally interesting effect size. While these probabilities are in some respects like frequentist “power”, they serve very different functions. For Bayes factors, these “power” calculations serve the function of indicating if you likely have the resources to obtain good enough evidence either way, and so test the theory severely. But once the data are in, the Bayes factor itself indicates the strength of evidence for H1 over H0; the “power” calculations do not modify that interpretation.

A severe test according to Popper (1963) requires that the prediction of the theory be implausible given background knowledge. Vanpaemel (2020) uses this principle to evaluate whether a test of a model is a severe one: Determine whether the predictions of the model are implausible, using a “data prior”, representing which outcomes can be considered plausible in the light of background knowledge.”. One could formalize the predictions of background knowledge by modelling a simple theory and determining whether a large Bayes factor could be obtained for the theory being tested against the alternative simple theory derived from background knowledge. For example, could demand characteristics make similar predictions as an alternative substantial theory (cf Dienes, Lush & Palfi, in press)? Lush (2020; Orne, 1962) illustrates a way of obtaining quantitative predictions from the theory of demand characteristics which could be compared to a theory one wished to test.

Discussion

There are two approaches for obtaining evidence for no effect: Either by specifying a minimally interesting effect size, and determining whether the evidence supports whether the true effect is smaller or larger than that (Blume et al., 2018; Kruschke, 2014; Lakens, Scheel, & Isager, 2018; Mayo, 2018); or specifying a rough scale of effect and using Bayes factors to determine the evidence for a model predicting that scale of effect versus a model predicting no effect (Dienes, 2014; Jeffreys, 1939; Wagenmakers et al., 2017). Either way one must determine a relevant effect size based on scientific context (Morey, Homer, & Proulx, 2018; Vanpaemel, 2011, 2016; Lee & Vanpaemel, 2018, consider the same question for testing cognitive models). This paper offers some guidelines on determining what one's theory predicts - in order to severely test it. Which approach one chooses in part depends on whether a minimally interesting effect size or the rough scale of effect is easier to justify using publicly available reasons. Those reasons will not pinpoint an exact effect size; this issue is partly addressed when using Bayes factors by modelling a range of predicted effect sizes. Still, that range is not precisely determined. Thus, for both approaches, one should indicate the robustness of the conclusion to different estimated effect sizes; Dienes (2019, in press) provide a method and notation for indicating robustness.

Bayes factors have been criticized for being sensitive to their prior, i.e. the way H1 is modelled (the model representing the predictions of the theory concerning relevant effect sizes) (e.g. Kruschke, 2013). However, any method of obtaining support for the claim of no effect needs to model relevant effect sizes; inference by intervals achieves this by specifying the null region. Thus, all methods for obtaining support for no effect are sensitive to priors (even if not explicitly coded as such, for example in AIC). So the problem of specifying a "prior" has to be confronted and cannot be avoided if one wishes to potentially be able to obtain evidence for no effect. Inference by intervals requires inference to depend exclusively on the minimally interesting effect size, which can be hard to pinpoint with objective reasons. Hopefully this paper gives some ways of approaching the problem.

Inference by intervals and Bayes factors typically do not test exactly the same models, and so the two approaches answer different questions. They will sometimes therefore give apparently different answers. For example, Dienes (2016) and Linde et al. (2020) found that inference by intervals is often unable to give a definitive answer when Bayes factors can provide support for the null model. The Bayes factor in modelling H1 as a probability distribution, uses more information about the theory (that predicts a difference), and hence can often draw more definite conclusions, where there

is relevant information to be used. (Often the information is about a plausible maximum; this is often the most influential aspect of a normal or uniform distribution in determining the value of a Bayes factor, Dienes 2015; a plausible maximum is also a value for which there is often good information).

If there is no basis for justifying an effect size, consider if one only need estimate (e.g. derive means and confidence or credibility intervals) to draw the inferences needed (Calin-Jageman, & Cumming, 2019; Cumming & Calin-Jageman, 2017; Gelman, Carlin, Stern, Dunson, et al., 2013; McElreath, 2016; Rothman & Greenland, 2018; Wagenmakers, Marsman, Jamil, Ly et al., 2018). One recommendation might be to find evidence for an effect before estimating it, to make sure there is something to estimate. On this approach, if there were evidence for H0, one would not estimate. In fact, one can only obtain evidence for nothing being there given grounds for claiming what size it would be were it to be there. So a theory, however minimal, is always needed for hypothesis testing. One should always estimate parameters if they are reported at all; in addition, one may also hypothesis test given a theory to test.

Both estimation and hypothesis testing may use priors. The prior used in Bayes factors (i.e. the model of H1) serves a different purpose from that used in estimation and therefore will rarely be the same. The prior distribution used in estimation has the purpose of allowing the most accurate estimate of parameters. The purpose of the model of H1 is to represent the predictions of a theory (for this contrast in purposes see Dienes, 2021a). One may find it hard to specify the prediction of a relevant theory, but still use a vague prior to estimate parameters from data. If there is only a sense of credibility that a parameter is relevant, one can still estimate. Remember that if estimation is used, it gives no grounds for asserting H0; one can only say the effect is (or is probably) between certain bounds. (Estimation also typically gives no grounds for concluding there is an effect (of any size), and that is because estimation typically assumes that there is an effect.)

The philosophy underlying the approach in this paper is critical rationalism (e.g. Notturno, 1999: Popper, 1972): Claims should be rationally criticized by considering the objective relations between them. Specifically, for inferential statistics, there should be reasons intrinsic to the scientific context for defining models of H1 and H0 such that tests of the models also test the corresponding theories whose predictions they represent. This may sound to be a simple truism. But note how often inferential approaches do not respect this truism. Significance testing against a no-effect H0 often fails to model H1 at all. If power or equivalence testing uses conventional effect sizes, there is no reason why the statistical model should be relevant to any particular theory: The theory has not

been rationally criticized. Objective Bayesians may use a default Bayes factor: but this presumes that every theory always makes the same prediction in every context. Subjective Bayesians postulate that probabilities are purely personal, and so may advise one to ask experts (not functioning as end users) to generate models of H1: But this presupposes that the predictions of a scientific theory can be different for different people depending on how they feel. The problem of obtaining evidence for something not being there is not solved by not using a model of H1, by using someone's default model of H1, nor by going by feelings without searching for the objective reasons that motivate them. That is because the evidence for something not being there is only as good as the grounds for claiming the effect, should it be there, is of the size modelled in the model of H1.

This article has only considered one aspect of the process of statistical inference. Other key aspects of inference, glossed over here, are modelling the data generation process (simple linear model? Hierarchical model?), choosing other distributions (in Bayesian statistics, for example, the likelihood function), data exclusion criteria and so on. This article is about how those other important choices interface with theory.

In sum, this article argues a change in approaching statistical testing is needed, compared to typical practice (though for exceptions see e.g. Vanpaemel, 2016), in order to test our theories severely: To obtain data counting against a theory predicting a difference, a range of predicted effect sizes must be specified that follow from the theory itself in its scientific context. Without there being good reasons for specifying such a range, there won't be good reasons for supporting the null hypothesis. But we need to be able to obtain support for the null hypothesis if we want to severely test our theories. We are only a small step away from achieving this, should we wish to take it.

References

Abelson, P. (2003). The Value of Life and Health for Public Policy. Economic Record, 79(SpecialIssue), S2–S13. https://doi.org/10.1111/1475-4932.00087

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ, 311*(7003), 485. doi: https://doi.org/10.1136/bmj.311.7003.485

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science, 28,* 1547-1562. https://doi.org/10.1177/0956797617723724

Anvari, F., & Lakens, D. (2019). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. *PsyArXiv. 10.31234/osf.io/syp5a*

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., Morris, J. N., Rebok, G. W., Smith, D. M., & Tennstedt, S. L. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. JAMA, 288(18), 2271–2281. DOI:10.1001/jama.288.18.2271

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence Trials, Intersection–Union Tests and Equivalence Confidence Sets. *Statistical Science, 11*, 283–319. doi:10.1214/ss/1032280304

Blume, J. D., McGowan, L. D., Greevy, R. A., & Dupont, W. D. (2018). Second-Generation p-Values: Improved Rigor, Reproducibility, & Transparency in Statistical Analyses. *PLoS One, 13*, e0188299. DOI: 10.1371/journal.pone.0188299

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*, 261–304. DOI: https://doi.org/10.1177/0049124104268644

Calin-Jageman, R. J., & Cumming, G. (2019). Estimation for Better Inference in Neuroscience. *eNeuro* 1 August 2019, 6 (4) ENEURO.0205-19.2019; DOI: 10.1523/ENEURO.0205-19.2019

Cumming, G., & Calin-Jageman, R.(2017). *Introduction to the new statistics: estimation, open science, and beyond*. New York: Routledge.

Díaz-García, A., González-Robles, A., Mor, S. et al. (2020). Positive and Negative Affect Schedule (PANAS): psychometric properties of the online Spanish version in a clinical sample with emotional disorders. *BMC Psychiatry 20,* 56. https://doi.org/10.1186/s12888-020-2472-1

Dickey, J. (1973). Scientific reporting and personal probabilities: Student's hypothesis. *Journal of the Royal Statistical Society. Series B (Methodological), 35*, 285-305. DOI: 10.1111/j.2517-6161.1973.tb00959.x

Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psycholology, 5*: 781. DOI: 10.3389/fpsyg.2014.00781

Dienes, Z (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press, pp 199-220.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 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. DOI: 10.1177/2515245919876960

Dienes, Z. (2020). http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm Downland 1 Feb 2020.

Dienes, Z. (2021a). Testing theories with Bayes factors. In Austin Lee Nichols & John E. Edlund (Eds), *Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences*, https://psyarxiv.com/pxhd2

Dienes, Z. (2021b). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*, 9–26. https://doi.org/10.1037/cns0000258

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian over significance testing. *Psychonomic Bulletin & Review, 25*, 207-218. DOI: 10.3758/s13423-017-1266-z

Dong, H., Clayards, M., Brown, H., Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ* 7:e7191 https://doi.org/10.7717/peerj.7191

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. DOI: 10.1177/2515245918773087

Flores, M., & Orcana, J. (2018). Heteroscedasticity irrelevance when testing means difference. *SORT: Statistics and Operations Research Transactions, 42*, 59-72. DOI: 10.2436/20.8080.02.69

Freedman, L. S. , & Spiegelhalter, D. J. (1983). The Assessment of Subjective Opinion and its Use in Relation to Stopping Rules for Clinical Trials. *Journal of the Royal Statistical Society: Series D (The Statistician), 32 (1-2)*, 53-160. DOI: 10.2307/2987606

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*, 156–168. DOI: 10.1177/2515245919847202

Gallistel, C.R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439–453. DOI: 10.1037/a0015251

Gelman, A., & Carlin, A. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science, 9,* 641-651.DOI: 10.1177/1745691614551642

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis*, 3rd Edition. Boca Raton: Chapman & Hall.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20. DOI: 10.1037/h0076157

JASP (2020). https://jasp-stats.org/ Downloaded 1 Feb 2020.

Jeffreys, H. (1939). *The theory of probability*. Oxford, England: Oxford University Press.

Jeffreys, H. (1948). *The theory of probability (2nd edition)*. Oxford, England: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 377–395. DOI: 10.1080/01621459.1995.10476572

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91*, 1343–1370. DOI: 10.1080/01621459.1996.10477003

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*(2), 171–184. DOI: 10.1586/erp.11.9

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General 142* (2), 573-603 . https://doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: a tutorial with R and BUGS*, 2nd Edition. London: Academic Press.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science, 1*, 270–280. DOI: 10.1177/2515245918771304

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. DOI: 10.3758/s13423-016-1221-4

Lakatos, I. (1978). *The methodology of scientific research programmes: Volume 1: Philosophical papers*. Cambridge University Press.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses: Sequential analyses. *European Journal of Social Psychology, 44*, 701–710. DOI: 10.1002/ejsp.2023

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology, Series B: Psychological Sciences*, *75*, 45–57. DOI: 10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*, 259–269. DOI: 10.1177/2515245918770963

Lee, M. D., & Vanpaemel, W, (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review, 25*, 114–127. DOI 10.3758/s13423-017-1238-3

Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E., & van Ravenzwaaij, D. (2020, November 10). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. https://doi.org/10.31234/osf.io/bh8vu

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192. DOI: 10.2307/2333251

Lovell, M., & Dienes, Z. (2021). Minimal mindfulness of the world as an active control for a full mindfulness of mental states intervention: A Registered Report and Pilot study. https://doi.org/10.31234/osf.io/3umz7

Lush, P. (2020). Demand characteristics confound the rubber hand illusion. Collabra: Psychology, 6(1), 22. DOI: 10.1525/collabra.325

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods, 51*, 2498-2508. DOI: 10.3758/s13428-018-1092-x

Ly, A., Verhagen, J., & Wagenmakers, E-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology, 72*, 19-32. DOI: 10.1016/j.jmp.2015.06.004

Malejka, S., Vadillo, M. A., Dienes, Z., & Shanks, D. R. (in press). Correlation analysis to investigate unconscious mental processes: A critical appraisal and mini-tutorial. *Cognition*,

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology, 3* (1), 25. doi:https://doi.org/10.1525/collabra.78

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond he statistics wars.* Cambridge: Cambridge University press.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman & Hall.

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference, 26*, 231–245. DOI: 10.1016/j.foodqual.2012.05.003

Morey, R. (2018). https://medium.com/@richarddmorey/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91

Morey, R. (2021). https://richarddmorey.github.io/BayesFactor/ Viewed 15 Feb 2021

Morey, R., Homer, S. and Proulx, T. (2018). Beyond statistics: accepting the null hypothesis in mature sciences. *Advances in Methods and Practices in Psychological Science, 1*, 245-258. DOI: 10.1177/2515245918776023

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. J*ournal of Mathematical Psychology, 72*, 6–18. DOI: 10.1016/j.jmp.2015.11.001

Morey, R. D., and Rouder J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*, 406–419. DOI: 10.1037/a0024377

Notturno, M. A. (1999). *Science and the Open Society*. Budapest: Central European University Press.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley: Chichester.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783. DOI: 10.1037/h0043424

Palfi, B., & Dienes, Z. (2019). When and how to calculate the Bayes factor with an interval null hypothesis. PsyArXiv https://psyarxiv.com/9chmw

Perugini, M., Gallucci, M., Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*, 319–332. DOI: 10.1177/1745691614528519

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

Price, D. D., Milling, L. S., Kirsch, I., Duff, A., Montgomery, G. H., & Nicholls, S. S. (1999). An analysis of factors that contribute to the magnitude of placebo analgesia in an experimental paradigm. *Pain, 83*, 147-156. https://doi.org/10.1016/S0304-3959(99)00081-0

Raftery, A. E. (1999). Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection." *Sociological Methods & Research, 27*, 411–427. DOI: 10.1177/0049124199027003005

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553–565. DOI: 10.1037/0033-2909.113.3.553

Rothman, K. J., & Greenland, S. (2018). Planning study size based on precision rather than power. *Epidemiology, 29*, 599–603. DOI: 10.1097/EDE.0000000000000876

Rouder, J. N. (2020). http://pcl.missouri.edu/bayesfactor Downloaded 1 Feb 2020.

Rouder, J. N., Morey R. D., Verhagen J., Province J. M., & Wagenmakers E. - J. (2016). Is There a Free Lunch In Inference? *Topics in Cognitive Science, 8*, 520-547. DOI: 10.1111/tops.12214

Rouder, J. N., Morey R. D., & Wagenmakers E. - J. (2016). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra. 2*, 1-12. DOI: 10.1525/collabra.28

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*, 225–237. DOI: 10.3758/PBR.16.2.225

Sandham, A.L., Dando, C.J., Bull, R, & Ormerod, T. C. (2020). Improving Professional Observers' Veracity Judgements by Tactical Interviewing. *Journal of Police and Criminal Psychology*, https://doi.org/10.1007/s11896-020-09391-1

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*, 322–339. DOI: 10.1037/met0000061

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*, 128–142. DOI: 10.3758/s13423-017-1230-y

Skora, L., Livermore, J. J. A., Dienes, Z., Seth, A., & Scott, R. B. (2020, May 4). Feasibility of Unconscious Instrumental Conditioning: A Registered Replication. *Cortex*, Stage 1 RR. https://doi.org/10.31234/osf.io/p9dgn

van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., ... Wagenmakers, E. (2019). The JASP Guidelines for Conducting and Reporting a Bayesian Analysis. *PsyArXiv*, https://doi.org/10.31234/osf.io/yqxfr

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology, 55*, 106 – 117. DOI: 10.1016/j.jmp.2010.08.005

Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology 72*, 183-190. DOI: 10.1016/j.jmp.2015.10.006

Vanpaemel, W. (2020). Strong Theory Testing Using the Prior Predictive and the Data Prior. *Psychological Review, 127*, 136–145, http://dx.doi.org/10.1037/rev0000167

Vanpaemel, W., & Lee, M. D. (2012). The Bayesian evaluation of categorization models: Comment on Wills and Pothos (2012). *Psychological Bulletin, 138*, 1253 – 1258. DOI: 10.1037/a0028551

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. DOI: 10.1037/a0036731

Wagenmakers, E-J. (2020). https://www.bayesianspectacles.org/rationale-and-origin-of-the-one-sided-bayes-factor-hypothesis-test/

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J.,... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57. DOI: 10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., ... Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. Lilienfeld & I. Waldman, (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123-138). New York: John Wiley and Sons.

Watson, D., Clark, L. A., Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063-1070. DOI:10.1037/0022-3514.54.6.1063

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences, 61*, 1340–1341. DOI: 10.1002/jps.2600610845

Zeidan, F., Gordon, N. S., Merchant, J., & Goolkasian, P. (2010). The Effects of Brief Mindfulness Meditation Training on Experimentally Induced Pain. *The Journal of Pain, 11*, 199-209. https://doi.org/10.1016/j.jpain.2009.07.015.

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

Funding information. This work was partially funded by a grant from the Economic and Social Research Council (ESRC; grant number ES/P009522/1).

Acknowledgements. Thanks to Julia Haaf, Wolf Vanpaemel and Daniel Lakens for valuable comments.