Why bayesian “evidence for H1” in one condition and bayesian “evidence for H0” in another condition does not mean good-enough bayesian evidence for a difference between the conditions

Article (Accepted Version)

Why Bayesian "evidence for H1" in one condition and Bayesian "evidence for H0" in another does not mean good enough Bayesian evidence for a difference between conditions

Bence Palfi[1,2a], Zoltan Dienes[1,2]

[1]School of Psychology, University of Sussex, Brighton, UK

[2]Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK.

[a]To whom correspondence should be addressed: Bence Palfi

Address: Pevensey Building, University of Sussex, Falmer, BN1 9QH, UK

E-mail: b.palfi@sussex.ac.uk

Abstract

Psychologists are often interested whether an experimental manipulation has a different effect in condition A than in condition B. To test such a question, one needs to directly compare the conditions (i.e. test the interaction). Yet, many tend to stop when they find a significant test in one condition and a non-significant test in the other condition, and deem it as sufficient evidence for the difference between the two conditions. This tutorial aims to raise awareness of this inferential mistake when Bayes factors are used with conventional cut-offs to draw conclusions. For instance, some might falsely conclude that there must be good enough evidence for the interaction if they find good enough Bayesian evidence for H1 in condition A and good enough Bayesian evidence for H0 in condition B. The introduced case study highlights that ignoring the test of the interaction can lead to unjustified conclusions and demonstrates that the principle that any assertion about the existence of an interaction necessitates the comparison of the conditions is as true for Bayesian as it is for frequentist statistics. We provide an R script of the analyses of the case study and a Shiny App that can be used with a 2x2 design to develop intuitions on the current issue, and we introduce a rule of thumb with which one can estimate the sample size one might need to have a well-powered design.

“The manipulation in condition A was statistically significant and by contrast, we found no statistically significant effect in condition B”. Many believe that these findings are sufficient to support the claim that there is a difference between conditions A and B in the effect of the manipulation. However, such an inference does not follow from these results, as it requires the test of the difference between the conditions in the effect of the manipulation, or in other words, the test of the interaction of condition by manipulation (Abelson, 1995, p. 111; Gelman & Stern, 2006). This inferential mistake is common in neuroscience (Nieuwenhuis et al., 2011) and one can safely assume that psychologists are also not immune from committing it. While it is perhaps an old saw now when it comes to null hypothesis significance testing (NHST), how does this relate to the use of Bayes factors (see Box 1)? At this point, before reading further, the reader might like to consider the example in Box 2.

As soon as conventional cut-offs are used for Bayes factors (see Box 1 for a brief introduction on the interpretation of the Bayes factor via the conventional cut-offs), there may be conditions where the inferential mistake is even more likely than with frequentist statistics. When there is good enough Bayesian evidence for H1 in one condition and for H0 in another, surely one can conclude that the effect is bigger in the first condition than the second! Personal discussions with colleagues created the impression in us that it may not be obvious for many researchers that this latter conclusion is incorrect even if they are aware of the issue when it comes to frequentist statistics. The reader can explore the extent to which they are attracted towards this inappropriate conclusion, which asserts an interaction without directly testing it, by considering the hypothetical study in Box 2 (a scenario that is discussed in detail in Example 2). They can also test whether they would be more inclined to accept this inappropriate conclusion when it is based on frequentist or on Bayesian statistics.

The Bayes factor is a continuous measure of the strength of relative evidence for H1 over H0 based on the ability of these hypotheses to predict the data at hand (Dienes, 2016; Kruschke & Liddell, 2018; Rouder et al., 2009). A Bayes factor of 1 means that the two hypotheses under comparison predicted the data equally well. The convention we follow (but it is not universal) is that the larger the Bayes factor the better H1 fits the data compared to H0, and the smaller it is, H0 is more in line with the data compared to H1. To aid decision making about the hypotheses, Jeffreys (1961) suggested BF > 3 to be the cut-off of substantial evidence for H1 over H0. Note that this value was chosen by Jeffreys with the intention that the Bayes factor should lead to similar judgment as NHST, when one is about to reject H0 (i.e., a statistical test resulting in p = .05 will usually provide a Bayes factor around 3, so long as the obtained effect size is about that predicted). By symmetry, we interpret BF < 1/3 as substantial evidence for H0 over H1. However, it does not indicate that this cut-off should be automatically accepted as the level of good enough evidence. Indeed, it is a rough guideline and it remains a matter of scientific debate (e.g., currently, the level of good enough evidence for H1 is defined as BF > 6 at *Cortex* and as BF > 10 at *Nature Human Behavior* for Registered Reports). Nonetheless, in this tutorial, we apply the cut-off of substantial evidence as good enough

Box 1. The interpretation of the Bayes factor

The central goal of this tutorial is to substantiate the statistical intuition in the reader that to claim the existence of a difference between two conditions or groups, one always needs to test the interaction, and this principle is as true for Bayesian as frequentist statistics. In this tutorial, we present the scenarios that the reader can stumble upon when they calculate Bayes factors for the evidence of the presence of an effect in an experimental and a control condition (or group) that had a significant and a non-significant statistical test, respectively. By this approach, we aim to illustrate that there are cases when using the Bayes factor instead of frequentist statistics could make it more likely to commit the inferential mistake, and there are cases when it may be the other way around. We use a hypothetical study as a case study and by increasing the sample size or reversing the effect size, we cover all the scenarios.

At a Golf Club in Sussex, a coach stumbled upon a sport psychology paper concluding that mental training (e.g., imagining hitting the ball with a golf club) can help golfers improve their skills when it is used combined with real training. Before implementing the mental training in all of their groups, the coach decided to test whether players can benefit from it. Therefore, the coach asked the students in one of the groups to engage in mental training twice every week for the next 3 months (on top of the traditional, real training). The coach also had a control group in which the students underwent real training but they were not told to do the mental training, and the students had roughly identical skills to those in the mental training group. The coach assessed the performance of the students at baseline and after 3 months of training. The evaluation was performed on an interval scale from 0 to 10, and based on past studies with other sports the coach expected that after 3 months of training performance could improve by about 2 units.

To draw conclusions from the analyses, Null Hypothesis Significance Testing (NHST) was used, and the alpha level was set at the traditional .05. The coach reported the results of two statistical tests and a conclusion, which was based on these tests. Evaluate the appropriateness of this conclusion on a scale that ranges from 0 to 10 where 0 means that you feel that the conclusion is completely inappropriate and 10 means that you feel that the conclusion is completely appropriate based on the information at your disposal.

Comparing baseline and post-training performance in the control training group yielded a non-significant result ($t(19) = 0.29$, $M_{diff} = 0.11$, $p = .776$). However, when they analysed the data of the group of golfers who engaged in the mental training, they found a significant difference between baseline and post-training conditions, with better performance after 3 months of training ($t(19) = 2.61$, $Mdiff = 0.81$, $p = .017$). Based on these results, they concluded that traditional training is more efficient when it is combined with mental training than when it is not combined with it.

Let us suppose that the coach used Bayes factors to draw conclusions instead of NHST. A Bayes factor is a continuous measure of relative evidence, it can tell us the extent to which our data supports one model (H1) over another (H0), which is reported as BF (As we explain later, a model of H1 is needed. Given the researchers had reasons for expecting an effect of about 2 units, we used a half-normal with SD = 2.). By convention, Bayes factors larger than 3 indicate good enough evidence for H1 (i.e., we can conclude that an intervention works) and Bayes factors smaller than 1/3 indicate good enough evidence for H0. Assess the appropriateness of their conclusion that was, this time, based on the Bayes factors by choosing a value from the same scale of appropriateness. Zero indicates that you feel that their conclusion is completely inappropriate and 10 means that you feel that their conclusion is completely appropriate.

Comparing the baseline and the post-training conditions of the control group yielded good enough evidence for H0 ($t(19) = 0.29$, $M_{diff} = 0.11$, $p = .776$, $BF = 0.24$). However, when they analysed the data of the mental training group, they found good enough evidence supporting H1 (i.e., difference between the baseline and the post-training conditions), with better performance after 3 months of training than at baseline ($t(19) = 2.61$, $Mdiff = 0.81$, $p = .017$, $BF = 6.12$). Based on these results, they concluded that traditional training is more efficient when it is combined with mental training than when it is not combined with it.

Box 2. Test your intuitions

The case study

Consider the hypothetical study from Box 2, in which a golf coach is trying to test whether or not adding mental training to traditional training can improve golf performance. To investigate this question, they randomly assigned students into traditional training (henceforth control group) and traditional plus mental training (henceforth mental training group) groups. They assessed golf performance at baseline and after 3 month of training. Therefore, they had a 2x2 mixed design. Hence, the crucial test of the idea that one can benefit more from golf training if it is combined with mental boils down to a 2x2 interaction of time of assessment (baseline vs post-training) and type of the training (traditional vs traditional plus mental). For the sake of simplicity, imagine that golf performance was measured on a scale from 0 to 10, and the coach expected that the mental training should improve performance by about 2 units.

Justifying the model of H1 and the model of the data

To compute a Bayes factor, one needs to specify the parameters of the models representing the predictions of the hypotheses under comparison (see Box 3 for more information on the essential parts of the Bayes factor). The model of H0 assumes no difference in the population. To model the prediction of H1 we employed a half-normal distribution with a mode of zero. The properties of the normal distribution align with the scientific intuition that small effect sizes are more probable than large ones (Dienes & Mclatchie, 2018), and we opted for a half-normal as it is in line with the directional prediction of H1. To specify the standard deviation of the distribution, we applied the expectation of the coach from the hypothetical study who assumed that performance should improve by about 2 units. When we have an effect size estimate based on earlier studies for instance, we can use it as the standard deviation of the distribution modelling the predictions of H1 (Dienes, 2014). Nonetheless, one might think that the effect size of an earlier study is not a plausible representation of the alternative, or there are no relevant studies. In these cases, there are several heuristics (e.g., Dienes, 2019) that one can follow to specify the predictions of H1. In addition, we need a model of the data (also referred to as likelihood; see Box 3). We used the t-distribution as the likelihood function, which is recommended over the normal distribution when the variance of the data is estimated as it is unknown (Dienes & Mclatchie, 2018). Finally, we will notate all of the Bayes factors as $BF_{H(0,\ 2)}$ following the convention introduced by Dienes (2014). This notation includes all the necessary information about the model of H1: H indicates that it is a half-normal distribution; zero refers to the mode and 2 to the standard deviation of the distribution. All Bayes factors reported in this paper represent evidence for H1 over H0.

Specifying all of these parameters requires the researcher to make many decisions, which has the side effect of increasing analytic flexibility and so the opportunity to cherry pick the results supporting the researcher`s pet theory. The most crucial step when one can introduce bias is perhaps the model specification of H1 that, in some cases, can have drastic effect on our conclusions. One way to reduce bias is by constraining analytic flexibility through pre-registering the exact parameters of the model of H1 or the strategy with which one will acquire those parameters (Chambers, 2013; Munafò et al., 2017). One can also consider reporting a "Robustness Region" (RR) that indicates the range of parameters (e.g., SDs of the half-normal distribution modelling H1) that would lead to the same conclusion (i.e., good enough support for H1 over H0, insensitive evidence, good enough support for H0 over H1) as the chosen model specification (Dienes, 2019). RRs can diminish bias by increasing transparency over the analytic choices in a similar manner as multiverse analyses (Steegen, Tuerlinckx, Gelman, Vanpaemel, 2016). RRs have the additional benefit that in cases when model specifications can be motivated in different ways, we can ascertain the robustness of our conclusion to the model specification by simply checking whether all plausible parameters lie within the RR. In this tutorial, we report the RR for every BF in the format of "$RR_{conclusion}$[min, max]" where min indicates the smallest and max indicates the largest SD of the model of H1 that brings us to the same conclusion. We will indicate the original conclusion in the subscript of RR by reporting one of the following: "BF < 3", "1/3 < BF < 3" or "BF > 3".

In order to assess the predictive ability of the hypotheses, we need to create models that represent their predictions. Modelling the prediction of H0 as no difference is the straightforward part of the process. However, specifying the predictions of H1 requires scientifically informed decisions in every case and so it can be a subject of debate. For instance, one needs to define the shape and parameters of the distribution representing the predictions of H1 on the possible population effect sizes. This brings up questions, such as whether the distribution should be uniform, t or normal; one-tailed or two-tailed; centred on zero or on a non-zero population value; what should be the level of variance in the model. The discussion of these decisions is beyond the scope of the current tutorial and so we refer the reader to Dienes (2015, 2019), Dienes and Mclatchie (2018). Nonetheless, in the case study of the current paper, we justify all of the choices about the model specifications. Finally, one needs to define the likelihood function, which is modelling the probability of the data along different population effect sizes.

Box 3. The anatomy of the Bayes factor

Disclosure

All the materials of this tutorial are available on the Open Science Framework page of the project at: https://osf.io/xrctq/. The page includes the R script of the analyses introduced here and the script of the Bayes factor function. It also contains the R script of a Bayes factor Shiny app that is a simple and interactive web application that can calculate the Bayes factors of 2x2 between groups and within participants designs. Box 4 demonstrates an example of the usage of the Bayes factor R script (namely, the test of the interaction in Example 1) and Figure 1 portrays how the Shiny app can be applied to compute all three Bayes factors of Example 1. The Bayes factor Shiny app can be accessed at https://bencepalfi.shinyapps.io/Bayesian_Interaction_App/

To calculate the Bayes factor in R, one needs to obtain the summary statistics of the data (mean, standard error and degrees of freedom) and decide on the parameters of the model of H1.

The following R script reproduces the results of the test of the interaction of Example 1 (all the text preceded by the # symbol are comments helping the reader and will be ignored by R when the script is run):

```
#Loads the Bayes factor function
#Note that the current R file and the file containing the function should be placed in the same folder
source("BayesFactor_normalH1_tlikelihood.R")


#Calculates the Bayes factor
Bf(sd = 0.491, obtained = 0.7, dfdata = 38, meanoftheory = 0, sdtheory = 2,  tail = 1)
```

The first three arguments of the function specify the parameters of the likelihood: the standard error, the estimate (i.e., raw effect size) and the degrees of freedom of the distribution, respectively. The last three arguments define the parameters of the model of H1: the centre (or mode if it is a one-tailed distribution) and the SE of the distribution, and whereas it is one- or two-tailed. When all parameters are provided, the function returns a vector containing the value of the Bayes factor (evidence for H1 over H0).

Box 4. Calculating the Bayes factor in R

Example 1: When the Bayes factor helps us avoid committing the inferential mistake

Suppose that the researchers found that the test comparing baseline and post-training performance was significant in the mental training group (t (19) = 3.58, M_{diff} = 1.11, p = .002), and it was not-significant in the control group (t (19) = 1.08, M_{diff} = 0.41, p = .295). Based on this, they concluded that the effect of the training is only expressed (at least after 3 months of training) if it is combined with mental training. This conclusion is premature for two reasons. First, one cannot claim the absence of an effect based on a non-significant test (Cohen, 1994; Dienes, 2014, Rouder et al, 2007) and so it is false to imply that we have evidence against the effectiveness of the training in the beginner group. We don`t have evidence for the contrary either, simply, we need to refrain from decision-making. Second, a

more relevant point for the purpose of the current tutorial, the dissimilarity of two categorical statements (i.e. significant vs not-significant) does not grant a meaningful categorical statement about their difference (i.e., the difference between the two is not necessarily significant in itself; Abelson, 1995, p. 111). From the second point, it follows that one needs to test the difference directly to make any meaningful claim on the interaction of the groups. The test of the interaction, however, yields a non-significant result (t (38) = 1.43, M_{diff} = 0.70, p = .162) meaning that one needs to suspend judgment regarding the influence of the mental training on traditional golf training.

The question arises: how would we decide in this scenario if we were to rely on a Bayes factor to form conclusions about the hypotheses? These data translate into substantial evidence for the effect of the training in the mental training group ($BF_{H(0, 2)}$ = 46.36, $RR_{BF > 3}$[0.2, 36.3]) and leaves us with insensitive evidence for the effect of training in the control group ($BF_{H(0, 2)}$ = 0.57, $RR_{1/3 < BF < 3}$[0, 3.5]) as well as for the interaction directly comparing the effects of the groups ($BF_{H(0, 2)}$ = 1.14. $RR_{1/3 < BF < 3}$[0, 7.4]). Clearly, one cannot easily claim that good enough evidence for the effect in one group and insensitive evidence in the other group is good enough evidence in itself for the difference between the groups. Apparently, using the Bayes factor may help us avoid the inferential mistake regarding the interaction, even if we were to ignore the results of the direct test of the interaction. Hence, using Bayes factors may increase the chance that one would conclude that the available data are simply not enough to make a decision about the hypotheses.

The only way to come to a conclusion regarding whether or not mental training combined with traditional training is superior to traditional training is to collect more data until we obtain evidence in one direction or the other. Optional stopping is not a problem for Bayesian statistics, the Bayes factor will retain its meaning regardless of the stopping rule applied (Dienes, 2016; Rouder, 2014[1]). Thus, we can check the Bayes factor every time we recruit a new participant and stop once the Bayes factor reaches a good enough level of evidence. For example, in this scenario, assuming that the raw effect sizes and their variances remain constant, we would need to recruit 94 participants in total (47 per group) to have substantial evidence for the interaction ($BF_{H(0, 2)}$ = 3.09, $RR_{BF > 3}$[0.3, 2.0]). In this scenario,

[1] Recently, this claim has been called into question under some conditions by Heide and Grünwald (2017). For replies to the concerns see Rouder (2019), and Wagenmakers, Gronau and Vandekerchhove (2019): As soon as one specifies one's model of H1, BF indicates the evidence for that model of H1 over H0 regardless of stopping rule.

the evidence for the efficacy of the training in the control group would still be insensitive with a $BF_{H(0, 2)} = 0.89$, $RR_{1/3 < BF < 3}[0, 5.4]$. Thus, evidence for an effect in one group, coupled with no evidence one way or the other in the other group, could still be evidence for a difference in effects between the two groups.

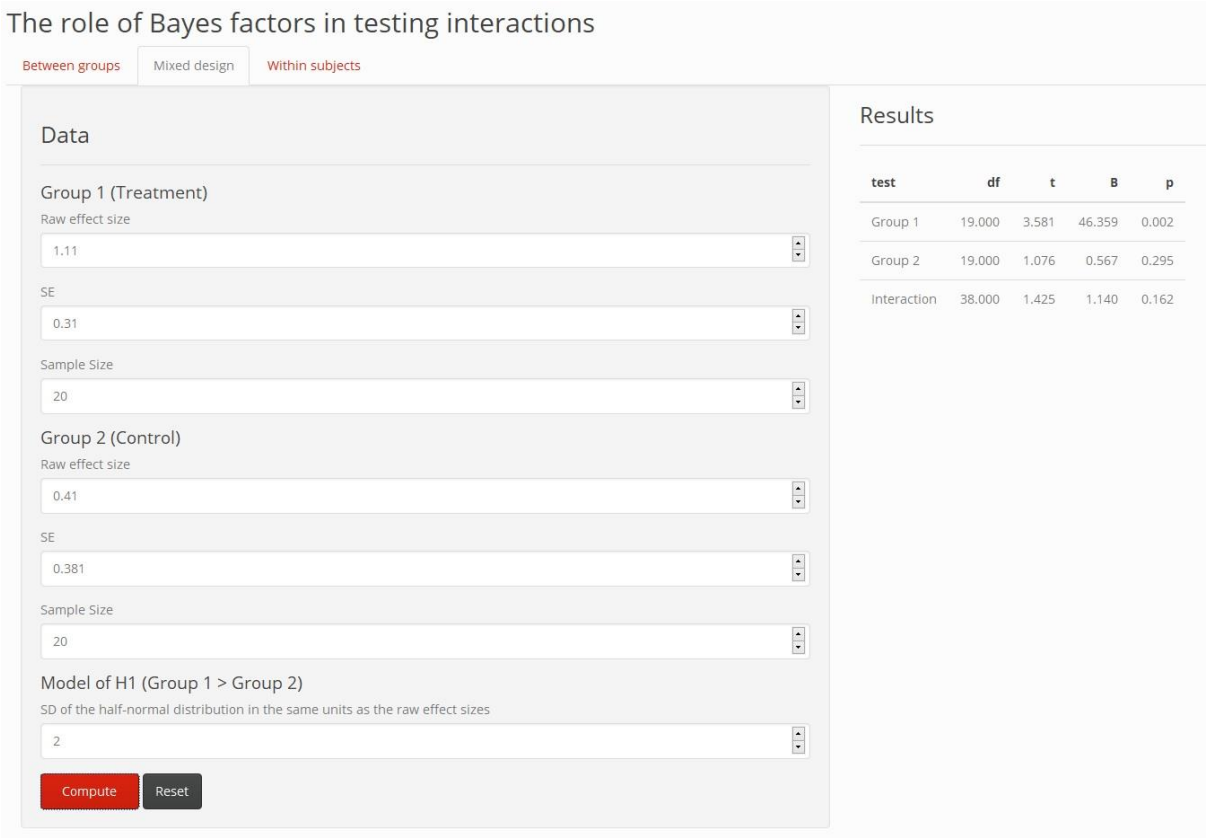

Figure 1. Print screen of the Shiny app that calculates the Bayes factor separately for the two groups and for the interaction based on the following statistical parameters: raw effect sizes and their SEs for the two groups and their interaction, the sample size, and the SD of the half-normal distribution that models the predictions of H1. In the top left corner, one can change between the "between groups", "mixed design" and "within subjects" options. The between groups and mixed design options are identical in that they run an independent t-test to test the interaction, and they request the same input parameters. For the within subjects design, one needs to provide the difference of the conditions and their standard deviation separately. The results appear on the right side of the screen, once the calculate button is pressed. The Shiny app reports the degrees of freedoms, the t-values, the Bayes factor and the p-values for the groups (or conditions) and for their interaction.

Example 2: When the Bayes factor might exacerbate the problem and seemingly creates an inferential paradox

Now let us consider the scenario from text Box 2 that only differs from Example 1 in that the raw effect sizes of differences between baseline and post-training conditions were reduced by 0.3 units in both of the groups. All other parameters (e.g., the standard deviations and the difference between the control and mental training groups) were kept constant. In this scenario, the results of significance tests probing the efficacy of the training, separately in the control and mental training groups, are identical to those of Example 1, being non-significant and significant, respectively. However, if we calculate the Bayes factors, it reveals that this scenario is different from Example 1 as we gain good enough evidence for the presence of the effect in the mental training group (t (19) = 2.61, M_{diff} = 0.81, p = .017, $BF_{H(0, 2)}$ = 6.12, $RR_{BF > 3}$[0.2, 4.3]) and good enough evidence for the absence of the effect in the control group (t (19) = 0.29, M_{diff} = 0.11, p = .776, $BF_{H(0, 2)}$ = 0.24, $RR_{BF < 1/3}$ [1.5, ∞]). It might seem intuitive to conclude that the evidence for the difference of the two must be substantial in itself as well (c.f., your feeling of appropriateness about the conclusion in Box 2). However, that is an unwarranted conclusion as the rule that "a meaningful categorical statement does not follow from the difference between the two categorical statements" (Abelson, 1995, p. 111) applies to Bayesian just as much as it applies to frequentist statistics. Hence, regardless of how tempting it feels to claim that the group with substantial evidence for H1 must be different from the group with substantial evidence for H0, we need to directly compare these two conditions to unravel whether there is an interaction.

Compared to Example 1, it appears that in this case relying on the Bayes factor rather than on the p-value would not help us avoid making the inferential mistake of ignoring the test of the interaction. On the contrary, using the Bayes factor may even amplify the problem as having good enough evidence for H1 in group A and for H0 in group B can easily create the false impression that there is no need for further statistical analyses and the two must be different. However, neglecting the test of the interaction is an inferential mistake, moreover, it would lead us to an incorrect conclusion, since the test of the interaction must yield the same result as in Example 1 as we kept the difference between the groups and their standard deviation fixed. It means that the test of the interaction is non-significant (t (38) = 1.43, M_{diff} = 0.70, p = .162) and the Bayes factor is insensitive ($BF_{H(0, 2)}$ = 1.14. $RR_{1/3 < BF < 3}$[0, 7.4]).

Seemingly, we got ourselves into a paradox in which we can claim that an effect exists in group A and it does not exist in group B, however, we cannot state that the effect is

stronger in group A than in group B. These conclusions are inconsistent with one another, but the Bayes factor should not take the blame for it. The cause of the existence of this paradox is that we introduced cut-offs to interpret the Bayes factor and so we reduced its continuous nature to a categorical one. That is the Bayes factors underlying the claims that there is an effect in group A and that the interaction is insensitive point in the same direction. Hence, we created the inconsistency by imposing a cut-off and labelling the first as good enough evidence for H1 and the second as insensitive evidence. Nevertheless, applying a cut-off to discern good enough from insensitive evidence is useful for scientific practice as it allows us to draw conclusions. And we often need to draw conclusions in order to move on with an experiment: have we established that manipulation does what it says, such as do we have evidence for lack of awareness in an unconscious condition; have we demonstrated that an effect replicates before we undertake the exploration of potential moderators; have we ruled out a nuisance alternative theory, and so on (Only a statistician and not a scientist would recommend not ever drawing any conclusions from statistics!). In other words, we gain a clear rule telling us when we have good enough evidence to make such a decision. On the other hand, if we rely on this decision rule, we need to accept that it can lead us to paradoxical situations.

Fortunately, there is a way to escape this paradox. There is no need to consider the evidence at our disposal as fixed. Therefore, the remedy to this problem is to collect more data until the Bayes factor of the crucial test exceeds one of the cut-off values (as mentioned earlier, optional stopping does not invalidate conclusions based on the Bayes factor). For instance, assuming that the raw effect sizes and their variances stay constant while we collect data for this study, we would need to recruit the same number of participants (47 per group) as we needed in Example 1 to obtain evidence for the difference between the groups. Note that optional stopping applies to multi-lab collaborations as well: if a lab runs out of participants before reaching good enough evidence, another lab can continue with the accumulation of evidence.

Discussion

In this tutorial, we aimed to illustrate how the application of Bayes factors with cut-offs relates to the old problem of the tendency to compare the statistical significance of the tests of two groups rather than the groups themselves. We introduced two scenarios in which group A had a significant effect whereas group B had a non-significant effect. In Example 1, employing Bayes factors instead of NHST may help us avoid the inferential mistake as the

test of the non-significant group turned out to be insensitive and it is unlikely that one would assume that the difference of good enough evidence for H1 in group A and insensitive evidence in group B indicates a clear difference between the two. In Example 2, however, the Bayes factor in group B provided good enough evidence for H0 and in such a scenario applying Bayes factors instead of NHST may increase the probability of committing the inferential mistake because the conclusions from the simple effects and the interaction contrast are literally inconsistent.

We observed that drawing a conclusion from the Bayes factor could sometimes lead to a paradox (i.e., good enough Bayesian evidence for H1 in group A, good enough Bayesian evidence for H0 in group B and the lack of good enough Bayesian evidence for their interaction). The reason for the paradox is that we use cut-offs that exchange the continuous measure of evidence to a categorical and ultimate claim about the state of the world in order to guide our decisions about the hypotheses. This situation bears a strong resemblance to Arrow`s theorem (1951) that demonstrates that there is no consistent way to explore the preference of a group ("will of the people"), and any ranked voting system (i.e. a system that turns strengths of opinion into a categorical outcome) will lead to paradoxes, perhaps undermining our faith in representative government and democracy itself. However, as Deutsch (2011) pointed out, Arrow`s theorem considers only a particular stage of decision making, as if preferences and options were fixed in social decision making. As long as preferences can be altered through open discussion and reasoning, and it is possible to modify or replace the options, democracy can be used consistently in selecting good policies. This conclusion is just as true for science, which seeks good explanations rather than policies or governments. When it comes to science, it would be a mistake to assume that evidence or the list of options (tested hypotheses) are fixed. Hence, even if we stumble upon a paradox, it remains a transient state and by accumulating more evidence, or by modifying the hypotheses (e.g., replacing a one-sided H1 with a one-sided H2 pointing into the other direction) we can dissolve the inconsistency. That is, the issue we raise about cut-offs leading to paradox is a very general one, not unique to science, let alone Bayes factors. The solution is just as general.

Continuing data collection until we obtain good enough evidence for or against the model predicting an interaction, which is critical to escape the inferential paradox illustrated in Example 2, can be challenging in some cases if there are not sufficient resources. Thus, estimating the sample size we might need to find good enough evidence for a hypothesis over

another one should play an essential role in the planning phase of an experiment. To this aim, we can compute the rough estimate of the sample size we need to probably obtain a Bayes factor that is equal to or larger than a specific value (i.e., the cut-off of good enough evidence defined by us). For instance, to have a long-term relative frequency of 50% to obtain a Bayes factor of 3 (or 1/3), we should simply replicate the steps of the sample size elevation of Example 1 and 2. That is, we can take the raw effect size and its standard deviation of a pilot study and assume that these parameters remain constant while we raise the sample size (see Dienes [2015] for a detailed tutorial). For an alternative view on how to plan the design of a future experiment to achieve good enough evidence, see Schönbrodt and Wagenmakers (2018), and for a tutorial see Stefan, Gronau, Schönbrodt, & Wagenmakers (2019). Finally, it is important to bear in mind that the sample size estimation is useful for planning, such as roughly estimating how long data collection will take, it has no influence on the inference made once the data are in. The final Bayes factor obtained is the measure of evidence for H1 over H0. The meaning of the Bayes factor is independent of the sample size estimation procedure (Dienes, 2016). That is, in Example 2, the sample size estimation suggests that we would need to recruit 47 additional participants to gain good enough evidence for the interaction. We may reach good enough evidence earlier or later than recruiting 47 more subjects, and once this happens, the conclusion regarding the presence or lack of the interaction should be solely based on the strength of the available evidence (i.e., the Bayes factor based on all the data collected to date).

In conclusion, it is evident that the Bayes factor is not a panacea for the inferential mistake discussed in this tutorial. In Example 1, we illustrated that the reliance on the Bayes factor may mitigate the issue, and in Example 2 we showed that it may exacerbate it. By depicting these two examples, we intended to raise awareness that any claim about the moderating effect of an independent variable should be supported by a sensitive test of the interaction regardless whether one uses frequentist or Bayesian statistics. Irrespective of how paradoxical it seems, good enough Bayesian evidence for H1 in group A and good enough Bayesian evidence for H0 in group B does not necessarily mean good enough Bayesian evidence for the difference of the two.

Author Contributions

BP performed the data analysis and wrote the script of the Shiny app with the supervision of ZD. BP drafted the manuscript and ZD provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

The authors declare no financial conflict of interest with the reported research. The project was not supported by any grant. Bence Palfi is grateful to the Dr Mortimer and Theresa Sackler Foundation which supports the Sackler Centre for Consciousness Science.

References

Abelson, R. A. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Arrow, K. J. (1951). *Social choice and individual values*. John Wiley & Sons, Inc.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*(3), 609-610.

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

de Heide, R., & Grünwald, P. D. (2018). Why optional stopping is a problem for Bayesians. *arXiv preprint arXiv:1708.08278*.

Deutsch, D. (2011). *The beginning of infinity: Explanations that transform the world*. UK: Allen Lane.

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

Dienes, Z. [Zoltan Dienes]. (2015, April 23). *How many participants might I need?* [Video file]. Retrieved from https://www.youtube.com/watch?v=10Lsm_o_GRg

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*. https://doi.org/10.1177/2515245919876960Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic bulletin & review*, *25*(1), 207-218.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331.

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

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*(1), 178-206.

Munafò, Marcus R., Brian A. Nosek, Dorothy VM Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J.

Ware, and John PA Ioannidis. "A manifesto for reproducible science." *Nature human behaviour* 1, no. 1 (2017): 0021.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*(9), 1105.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review*, *21*(2), 301-308.

Rouder, J. (2019). On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald. (Preprint: https://psyarxiv.com/m6dhw/)

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.

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, *25*(1), 128-142.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712.Wagenmakers, E. J., Gronau, Q. F., & Vandekerckhove, J. (2019). Five Bayesian Intuitions for the Stopping Rule Principle. (Preprint: https://psyarxiv.com/5ntkd)

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.