

Bayes factors for logistic (mixed effect) models

Catriona Silvey, Zoltan Dienes, Elizabeth Wonnacott

21 August 2024

Author Note

A pre-print of this paper can be found at <https://psyarxiv.com/m4hju>. This link is also available on the authors' websites and has been shared via social media. This document was written as an R Markdown file and this file can be found on <https://github.com/silveycat/bayes-factor>. This repository also contains the case study data files and the code and output for the simulations. The case study is adapted from analyses of data reported in Singh et al. (2021).

The work is not under consideration at any other journal and has not been presented at any conferences.

Abstract

In psychology, we often want to know whether or not an effect exists. The traditional way of answering this question is to use frequentist statistics. However, a significance test against a null hypothesis of no effect cannot distinguish between two states of affairs: evidence of absence of an effect, and absence of evidence for or against an effect. Bayes factors can make this distinction; however, uptake of Bayes factors in psychology has so far been low for two reasons. Firstly, they require researchers to specify the range of effect sizes their theory predicts. Researchers are often unsure about how to do this, leading to the use of inappropriate default values which may give misleading results. Secondly, many implementations of Bayes factors have a substantial technical learning curve. We present a case study and simulations demonstrating a simple method for generating a range of plausible effect sizes, i.e. a model of H_1 , for treatment effects where there is a binary dependent variable. We illustrate this using mainly the estimates from frequentist logistic mixed-effects models (because of their widespread adoption), but also using Bayesian model comparison with Bayesian hierarchical models (which have increased flexibility). Bayes factors calculated using these estimates provide intuitively reasonable results across a range of real effect sizes.

Introduction

Do older people exert more prosocial effort than younger people (Lockwood et al., 2021)? Does high talker variability in training examples make it easier for non-native speakers to learn Mandarin tones (Dong et al., 2019)? Does informing people about the scientific consensus help correct false beliefs about climate change (Stekelenburg et al., 2021)? These research questions, drawn from across the psychological sciences, illustrate that we are often interested not only in the size of an effect, but in whether or not an effect exists at all.

The traditional way of determining the existence of an effect is to use frequentist statistics. In this approach, we begin by assuming that the null hypothesis (H_0) is true: older people and younger people exert the same amount of prosocial effort; high-variability training and low-variability training are equally beneficial for Mandarin tone learning; and people’s false beliefs about climate change remain unchanged after they are informed of the scientific consensus. Given the truth of H_0 , we then ask: what is the probability of observing data at least as extreme as the data we observed? Prior to collecting the experimental data, we set α , the probability below which we will reject H_0 . A standard frequentist test then returns one of two results. Either $p < \alpha$, H_0 is rejected, and we conclude that there is an effect, or $p \geq \alpha$ and we are unable to reject H_0 . However, crucially, the latter result does not discriminate between two possible states of affairs: a) we have evidence that H_0 is true, i.e. there is no effect, and b) we do not have sufficient evidence to tell whether or not H_0 is true.

We might want to distinguish these possibilities for several reasons. For the first research question, if older people and younger people show equivalent levels of prosocial behavior, this is interesting in itself, suggesting this behavior is stable across the lifespan. For the second research question, previous literature (Lively et al., 1993) suggests we should find an advantage for high-variability training; if we fail to replicate this advantage, we want to know whether our data actually provide evidence against it or are simply ambiguous. For the third research question, if there is evidence that informing people about the scientific consensus does not help correct false beliefs, science communicators should not waste time on this strategy, but if we simply do not yet have evidence one way or another, we should collect more data before dismissing the strategy out of hand.

An alternative approach that solves this problem is to use Bayes factors¹ (Dienes, 2019; Jeffreys, 1939; Kass & Raftery, 1995; Verhagen & Wagenmakers, 2014). Instead of assuming that the null hypothesis is true, we compare the probability of the data under H_0 to the probability of the data under a specific hypothesis of interest H_1 . Usually, H_0 is a point hypothesis, encoding the theory that the effect is exactly 0, whereas H_1 can be modeled as a distribution over the values an observed effect might be expected to take on, given a scientific context. For deriving the probability of the data given H_1 , the probability we calculate is therefore a weighted average which takes into account both (i) the probability of the data given a particular effect size (known as the likelihood) and (ii) how probable that effect size is based on a theory (i.e., the model of H_1). To obtain BF_{10} , the Bayes factor for H_1 against H_0 , we divide the probability of the data under H_1 by the probability of the data under H_0 . If the resulting value is larger than 1, the data are more probable under H_1 than under H_0 . If it is smaller than 1, the data are more probable under H_0 than under H_1 . The Bayes factor thus quantifies how much more confident we should be in H_1 compared to H_0 , or vice versa, now that we have observed the data. While a Bayes factor is a continuous measure, Jeffreys (1939) suggested 3 as the lower bound above which we should pay attention to evidence for H_1 ; thus by symmetry we can take $\frac{1}{3}$ as the upper bound below which we should pay attention to evidence for H_0 . While Jeffreys uses the term ‘substantial’ for this level of evidence, here we use the revised terminology ‘moderate’ proposed by Lee & Wagenmakers (2014), as this better captures the degree of evidence we are talking about. To situate this with reference to p -values, if the effect size we obtain is roughly the same as our H_1 predicted, a Bayes factor of 3 in favor of H_1 corresponds to a p -value of around .05 (Dienes, 2014). Indeed, our experience in generating Bayes factors with real data is that in cases where we find a Bayes factor roughly greater than 3, we generally also find a p -value less than .05 (though there is in fact no monotonic relation between Bayes

¹A further alternative is to test for equivalence (and thus the absence of the null) using two one-sided frequentist tests (TOST) or the Bayesian highest density interval region of practical equivalence (HDI-ROPE). See Linde et al. (2021) for a comparison of these approaches.

factors and p -values). However, where we find a p -value greater than .05, the Bayes factor has the advantage of disambiguating whether we have evidence for no effect ($BF < \frac{1}{3}$) or no evidence to speak of either way ($\frac{1}{3} < BF < 3$).

Bayes factors, then, can provide an intuitive answer to the question of whether or not an effect exists. In addition to this, Bayes factors offer a number of other advantages. Firstly, unlike p -values, they provide a continuous measure of the evidence: while threshold values can be useful as a heuristic, the change in evidence as the Bayes factor increases or decreases is quantitative rather than qualitative in both directions. Secondly, Bayes factors as a measure of evidence are not sensitive to optional stopping (Rouder 2014), making them ideal for use in fields where samples are hard to obtain, such as developmental research or research with specialized populations. Thirdly, inference using Bayes factors is not affected by whether the hypothesis was formulated before or after observing the data, nor in principle do Bayes factors need to be corrected for multiple testing as such. (Situations where multiple testing corrections seem appropriate arise for a number of reasons that do need to be addressed: one may have to consider, however, whether or not the hypotheses in a multiple testing situation had low prior probability; or consider the evidence that all tests combined give to a theory - see (Dienes, 2016, 2021) for further discussion; pre-registration and division of analyses into planned and exploratory are also advised to remove bias from the inferential chain from data to theory). In all these ways, Bayes factors tend to match better with researchers' intuitions than classical frequentist statistics (Dienes, 2011).

However, uptake of Bayes factors in the psychological sciences has not so far been particularly high. Beyond the fact that there are currently many competing recommendations regarding how Bayes factors should be used (Aczel et al., 2020), two obstacles may account for this. The first is that Bayes factors require the researcher to be explicit about the range of effect sizes their theory predicts. In cases where the previous literature does not provide a clear estimate, researchers are sometimes at a loss for how to do this in a principled yet accessible way. While some software packages provide "default" values for modelling H_1 (e.g. Morey & Rouder, 2018), the use of unprincipled estimates of predicted effect size is problematic. A Bayes factor is a ratio representing how strongly the data support one theory (i.e. the one being tested) over another (the null). It is therefore only sensible if we have a sensible model of the predictions of the theory being tested in the current experimental context (Vanpaemel, 2020). Default models make assumptions which may lead to distributions of predicted values which include highly implausible values. The resulting Bayes factor ratio is therefore not meaningful as a test our theory and if, as is generally the case, the default parameters are an overestimation of the scale of the effect predicted, this may lead to spurious evidence for H_0 (Dienes, 2023; Nicenboim & Vasishth, 2016). The second obstacle is that many implementations of Bayes factors require the researcher to get to grips with new software that has a steep learning curve. An approach that allows researchers to calculate Bayes factors on the basis of output from the statistical models they are already using would be more accessible and would potentially encourage greater uptake of this method in the field.²

The aim of this paper, then, is twofold. One is to outline how a researcher can use constraints inherent in their data to define a plausible maximum effect, and from there derive a distribution that represents their H_1 . We refer to this as the motivated-maximum approach. Another aim is to demonstrate how Bayes factors can be simply computed by researchers who are already using mixed-effects models implemented in lmer (Bates, Mächler, et al., 2015) to analyse their data. Methods that model the hierarchical structure of the data are the most appropriate way to analyse data that involve repeated or nested measurements, such as when a number of participants are tested on a number of items (a type of data that is ubiquitous across the psychological sciences) and have become widespread in the language sciences (the research area of the first and last author).³ Dienes (2008) calculator for computing Bayes factors has the benefit of requiring only two numbers to represent the data relevant to the hypothesis being tested: a mean difference and the associated standard error. We demonstrate how these may be obtained from the beta and standard error of the coefficient for a fixed effect in a mixed effect model. We also show how the motivated maximum approach may allow us to use other fixed effect coefficients within the same model to derive the scale factor parameter

²Johnson et al. (2023) for an alternative approach to hypothesis testing with Bayes factors using priors informed by theory, albeit using standardized effect sizes

³While this paper provides examples of the use of Bayes factors with estimates from mixed-effects models, this method can be applied in conjunction with other frequentist methods such as t -tests or ANOVA. See Dienes et al. (2012) and Ziori & Dienes (2015) for examples.

which informs the model of H_1 , which is the third and final value needed for computing the Bayes factor with the Dienes (2008) calculator. Note that while all of our examples use logistic mixed effects models, as recommended for use in analyzing data with a binary response (Jaeger, 2008), it is equally possible to compute a Bayes factor using the beta and standard errors from the fixed effects coefficients in a linear mixed effects as the data summary, and indeed the same is true for the coefficients of linear and logistic regression more broadly. The motivated maximum approach which we outline may also be applied more broadly; however the method does require that the researcher can define a meaningful baseline or lowest expected performance which is most straightforward in the case of binary data where there is a notion of chance performance. However it could also hold in other situations (for example with response time data there may be a minimum plausible response time which could be used as a baseline). Finally, in the current paper we only explore the case where we are considering difference relating to binary predictors, such as where there are two experimental conditions.

The advantage of this approach is that, while it still requires a researcher to think carefully about their theory and translate it into a relevant H_1 , it does not require them to master any additional technical skills beyond those they already possess. We present simulations showing that using the Dienes calculator with motivated-maximum estimates results in intuitively reasonable Bayes factors across a variety of cases. We also provide evidence that we get qualitatively similar results if we apply the motivated maximum approach to determine the priors for H_1 , but use an alternative common approach to computing Bayes factors with mixed models, namely, comparing two full Bayesian hierarchical models. Finally, we outline the implications of our results for researchers to consider when thinking about experimental design.

Methods

This paper will use the method outlined in Dienes (2014) for calculating Bayes factors that quantify the evidence for a specific alternative hypothesis H_1 compared to the null hypothesis H_0 . The model of H_0 is simple: it assumes the effect is 0. The remaining elements we need in order to calculate the Bayes factor are:

- a statistical summary of the data
- a model of H_1

Summary of the data

A reasonable summary of the evidence the data provides with respect to a specific effect of interest is the parameter estimate for this effect and its standard error (Dienes, 2014). For experiments which involve multiple trials per participant, mixed-effects models are the most appropriate analysis method (see Bates, Mächler, et al. (2015) for a frequentist solution; we discuss the Bayesian approach later). In the examples presented in this paper, we focus on logistic models, but the same principle applies for linear mixed-effects models. In both cases, we can use the parameter estimate and associated standard error for the relevant coefficient from the output of the mixed-effects model as our summary of the data. Using the Dienes calculator the calculation of the Bayes factor rests on the assumption that parameter estimates are drawn from a normal distribution across samples – the same principle that underlies the calculation of the standard error and associated t and z -statistics in a frequentist analysis. In the case of logistic regression, this assumption is met as the estimate and standard error are in log-odds space. (The model of H_1 must take this into account, as we describe below.)

Note that in doing this, we assume that the estimate and standard error are a good representation of the data. This will not necessarily hold if the data violate the assumptions of mixed-effects regression (Schielzeth et al., 2020). However, in this case the concern would apply equally to the frequentist results from the mixed-effects model itself as to the Bayes factor calculated on the basis of this output. Similarly, there is considerable debate about best practice in constructing mixed effect models, for example: Should a maximum random

slopes structure be used? What should be done in the case of non-convergence etc. (Barr et al., 2013; Bates, Kliegl, et al., 2015; Matuschek et al., 2017). These choices will affect the standard error terms for the fixed effects, and thus the computation of either a frequentist p value or a Bayes factor on the basis of this output. In the current paper, we do not address these questions about best practices in constructing mixed effects models. Instead, we assume a reader who is already using mixed effect models, and thus has already determined how to answer these questions when analysing their dataset, but who wishes to calculate Bayes factors rather than (or in addition to) p values on the basis of the output.⁴

Model of H_1

The final component we need in order to calculate the Bayes factor is the one that often gives researchers the most difficulty: the model of H_1 . The model of H_1 is the plausibility of different effect sizes under our theory. One misconception is that we have to be very precise in our predictions. While we do want to avoid an overly vague H_1 that includes implausible values, since this can result in spurious evidence for H_0 , we do not have to know exactly what effect size we expect to observe: all we have to do is translate the predictions of our theory into a distribution over probable effect sizes. This can be easier than it sounds. The starting point is to choose a distribution type, based on sensible assumptions. One common choice of distribution is the normal distribution. This can represent the expectation that smaller effects are more probable than larger effects, if it has a mean (and mode) of zero. The researcher must specify two parameters: the mean and the standard deviation. Note that the *mean* encodes what we expect to be the most probable effect size. At first glance, it might seem that this should be set to the rough predicted effect size. However, note that this would then make larger and smaller effects symmetrically less probable. In psychology, our effect sizes are generally small and we usually expect smaller effects to be more probable than larger effects. This is better captured by setting the mean to zero. The most probable effect is then 0, with larger positive and negative effects symmetrically less probable. A normal H_1 with a mean of 0 is harder to discriminate from H_0 , which makes it a more stringent test: if we find evidence for H_1 using this model, we may be correspondingly more persuaded in favor of our theory.

The standard deviation encodes the probable range over which the effect might vary. If specifying a mean of 0, the researcher may choose to set the SD to a rough predicted effect size. Because of the properties of the normal distribution, this encodes the prediction that the observed effect is unlikely to be more than twice this predicted value. In many cases, this is a reasonable prediction; however, the researcher should check that this makes sense with reference to their specific theory and measurement scale, rather than using it as a default. Note that a normal distribution centered on 0 encodes a non-directional hypothesis: positive and negative effects are equally probable. If our H_1 is directional - e.g., the theory predicts that learning in the experimental condition will be better than learning in the control condition, and the reverse would actually be counter to our theory - then we can model this by instead using a half-normal distribution, where negative values are given a probability of 0. This is true all of the examples we consider in the current paper.

For the remainder of this paper, we will model H_1 as a half-normal distribution with a mode of 0 and a standard deviation corresponding to our rough estimate of expected effect size. This encodes a directional theory where the theory predicts only zero or positive effects, smaller effects are more probable than larger effects, and the maximum effect we might observe is around twice the expected effect size. A useful property of this model is that, if we can derive a maximum effect, we can then work backwards and halve this to obtain an expected effect size to use as the SD of our half-normal distribution. It is this property that underlies the motivated-maximum approach. However note that in different circumstances, the assumptions of this distribution might not be appropriate. For example, we might have a non-directional prediction so that we should use a full rather than half normal. Other non-normal distributions might also be more appropriate, for example, if we might believe that all values within a range are equally likely we could use a uniform; if we believe that smaller values are more likely but have strong reason to believe that the sexpected effect

⁴NB The examples we use in the paper assume a maximum random effects structure and adhere to the principle of marginality, i.e. including both interaction and constituent main effects. However, the logic of the Bayes factor computation which we aim to exemplify is not affected by these choices

will be small relative to the maximum we might choose a Cauchy (where the maximum is around 7 times the scale factor, rather than 2 times as in the normal case) (see Dienes 2019 for further discussion). The researcher must determine which distribution is appropriate with reference to their specific theory. Note also that picking a distribution type - based on sensible assumptions - is not the same as using defaults since we still have to specify the scale factors for those distributions that establish a realistic range of effect sizes predicted by our particular H_1 (as tested in this experimental context). Specifically, we need a predicted effect size. How can this be obtained?

The most obvious way to obtain a predicted effect size is to consult the previous literature. In cases where a well-powered previous study exists investigating a similar enough question - or, better yet, a meta-analysis that takes into account the reliability of many previous studies to come up with a best estimate - this may be the most appropriate option. However, until recently, effect sizes were not routinely reported in much of the psychological literature (Funder & Ozer, 2019); in applied linguistics, for example, the rate of effect size reporting is around 50% (Wei et al., 2019). Even where effect sizes are reported, they tend to be standardized⁵ rather than raw. While standardized effect sizes may sometimes be useful for comparing effects across studies, they are often not well-suited for Bayes factor analyses because they are influenced by factors irrelevant to the theoretical predictions we want to formalize in our model of H_1 , such as measurement error. On the other hand, raw effect sizes will not be comparable unless the design of the previous and the current experiment are very similar. If no sufficiently similar previous experiment exists in the literature, a researcher could instead choose to run a pilot study in order to calibrate the expected effect size. However, pilot studies tend to have small samples and thus to be unreliable in their estimates of effect sizes; furthermore, researchers are more likely to follow up on pilot studies that produce large effects, resulting in systematic overestimation effects in pilot studies that are used to inform statistical analyses (Albers & Lakens, 2018; Dienes, 2017).

A third strategy is for a researcher to use information from independent sub-samples of their own data to inform their estimate of probable effect size. For example, imagine a researcher has previously observed the effect of an experimental manipulation in a study with adult participants. The researcher now wants to investigate the effect of the same manipulation in children. It might be reasonable to use the effect size observed in adults as the effect size our H_1 predicts we will observe in children, as we have done in previous work (Sinkeviciute et al., 2019; Wonnacott et al., 2017).

These strategies are ad hoc: they rely on the researcher having access to comparable data from other studies, or independent groups within the same study. This paper will focus on a more general approach that instead looks at the data from the study currently being analysed and asks: what is the maximum effect we could observe in these data? This *motivated-maximum approach* is similar in spirit to the ‘room-to-move heuristic’ described in Dienes (2019); however, rather than using one condition as a benchmark to define the maximum difference that could be observed between conditions, here we consider as our maximum the situation where one condition is at a baseline level (here chance) and the other condition has the maximum value possible based on the grand-mean. We first consider a simple case involving one binary predictor variable, where our H_1 is that there is a main effect. We demonstrate that in this case we may be able to estimate the effect as equal to the difference between the grand-mean and the baseline (i.e. chance). We then consider a more complex case involving two interacting binary predictor variables where our H_1 is that there is an interaction. We demonstrate that in this case, we may be able to base our estimate on twice the difference between the grandmean and the baseline (chance).

Note that while all the experiment designs described below are drawn from the language sciences, this is only for the purposes of illustration: the approach is equally applicable to similar experiment designs across other fields of psychology.

Case 1. Imagine a study investigating whether word learning benefits from high-variability materials - i.e. hearing words produced by multiple talkers. Adult participants are exposed to words and referents in a

⁵Though a log odds ratio can be considered a standardized effect size in some contexts, in the examples in the current paper it is best considered as a raw effect size. A log odds ratio can be used as a standardised effect when it is used to represent an underlying normal deviate which is itself a standardised effect (Chinn, 2000). For example, for a 2 by 2 table, the SE depends only on the relative Ns. In contrast, in the logistic mixed effect models we are using, the sample error for the log odds ratios of the fixed effects is not fixed by the number of observations: because of the different variances over participants and items, we could have two estimates of a fixed effect with the same log odds ratio but different SE’s, despite having the same number of participants and items. Thus, the log odds, as we use it, is not a standardised effect size.

novel language- all words are heard equally frequently, but for half of the words, they are always produced by the same talker (low-variability exposure) and for half of the words the tokens come from six different talkers. After exposure, a 2-alternative-forced-choice (2AFC) post-test is administered in which participants hear one of the words from training and must identify the correct referent from a choice of two. Our hypothesis, based on previous work (Barcroft & Sommers, 2017; Sinkeviciute et al., 2019), is that high-variability training will lead to better vocabulary learning. However, we may not know exactly how big this improvement will be. A reasonable assumption for motivating a maximum effect is if the lowest performing condition is at a meaningful baseline, such as chance. A further constraint comes from the average performance the researcher actually observes across conditions in the data.

In a design where predictors are centered, the intercept term i in a mixed-effects model represents the grand mean, or the average of performance in the two conditions. If performance in the lowest performing condition is l and in the high performing condition is h :

$$\begin{aligned} i &= (h + l)/2 \\ 2i &= (h + l) \\ h &= 2i - l \end{aligned}$$

If d is the difference between conditions (the effect of condition):

$$d = h - l$$

Substituting for h :

$$\begin{aligned} d &= (2i - l) - l \\ d &= 2i - 2l \end{aligned}$$

In the case where the test trials are 2AFC as described above, our baseline performance is chance which corresponds to .5 proportion correct, or 0 when transformed to log-odds. The difference between conditions d is at its maximum when the mean performance in the lower condition is at this baseline i.e. $l = 0$. Then the equation simplifies to:

$$d = 2i$$

The difference between the conditions in this maximal situation - *the motivated maximum* is thus twice the intercept. Using the heuristic described above, we set the expected effect s to be half this value:

$$s = i$$

We therefore set the standard deviation of the half-normal distribution that is our model of H_1 to equal the intercept from our mixed-effects model. An intuition for why this is reasonable is as follows. If there were no effect of condition, $s = (l + h)/2 = l = h$. So if performance in either condition is above a chance baseline of 0, the more scope there is for there to be a difference between them, and so the more the evidence should count against there being a difference should the sample difference be close to zero.

Note that in cases where baseline performance is not 0, we will instead set the standard deviation to equal the intercept minus the baseline value. For example, in a 3AFC task, chance is .33 proportion correct, or -0.69 log-odds; we would therefore set the standard deviation to equal $i + 0.69$.

Case 2. Now suppose that instead of investigating vocabulary learning, we are interested in whether multi-talker, high-variability (HV) phonetic training improves learners' ability to discriminate non-native phoneme contrasts more than single-talker, low-variability (LV) training (Brekelmans et al., 2022; Dong et al., 2019; Logan et al., 1991). The condition manipulation here is between participants, so that half of the participants are exposed to exemplars words produced by multi-talkers and half are exposed only to a single-talker. In

addition, we expect that our participants could have some ability to discriminate the contrast even before exposure, so we use a pre-/post- design where participants are administered a 2AFC discrimination test before and after training. We thus have one within-subjects predictor, test-session (pre-test vs. post-test) and one between-subjects predictor, condition (LV vs. HV) and we are interested in the interaction between test-session (improvement from pre-test to post-test) and condition (HV vs. LV): do participants in the HV condition improve more from pre-test to post-test than participants in the LV condition? Following previous literature (Lively et al., 1993), we may expect HV training to lead to more improvement from pre-test to post-test than LV training. Again, though, we may not have strong expectations for how much better we expect improvement in the HV condition to be.

As for Case 1, we can use the intercept and a measure of baseline lowest performance to constrain the maximum interaction effect that we might observe. We assume that the maximum interaction would occur where a) both conditions are at baseline in the pre-test and b) there is no improvement from pre-test to post-test in the LV condition. l is our baseline, or the lowest performance we expect; h , here, is performance in the one cell that deviates from baseline, post-test in the HV condition. Again, if both predictors are centered, the intercept i from a mixed-effects model represents the grand mean performance across cells in the design:

$$\begin{aligned} i &= (l + l + l + h)/4 \\ i &= (3l + h)/4 \end{aligned}$$

Expressing h in terms of our known quantities, i and l :

$$\begin{aligned} 4i &= 3l + h \\ h &= 4i - 3l \end{aligned}$$

The interaction effect (d) is the difference between the improvement from pre-test to post-test in the HV condition and the improvement from pre-test to post-test in the LV condition:

$$\begin{aligned} d &= (h - l) - (l - l) \\ d &= h - l \end{aligned}$$

Substituting for h :

$$\begin{aligned} d &= (4i - 3l) - l \\ d &= 4i - 4l \end{aligned}$$

Using the same heuristic as above, we set the expected effect s to be half this value:

$$s = 2i - 2l$$

Again, in the case where test trials are 2AFC, chance corresponds to .5 proportion correct and hence $l = 0$. Thus the expected effect s in this case is maximum difference d is:

$$s = 2i$$

For the interaction, we set the standard deviation of the half-normal distribution that is our model of H_1 to equal twice the intercept from our mixed-effects model.

An alternative method for Case 2 sets the standard deviation of the H_1 distribution to equal instead the main effect of test-session from the mixed-effects model. We outline the logic for generating this estimate and the results from simulations in the Supplemental Materials. Here, we focus on the intercept-based method outlined above, since our simulations find it performs better across a wide range of cases (though see caveats in the discussion below). It is also worth noting that this alternative method does not require that the researcher can define a meaningful baseline in terms of lowest expected performance.

There are important caveats to the motivated-maximum approach. The first is that, when using a distribution where the expected effect is set as the scale factor, the expected effect must be positive (since a scale factor such as the standard deviation cannot be negative). The second caveat is that the motivated-maximum approach as described above can only be used in the case of binary predictors; we are currently working on developing an alternative method for the case of a continuous predictor. The third, more general consideration is that the motivated-maximum approach should not be applied as a default. The difference between the intercept and a baseline can be a good estimate to use as a scale factor for H_1 , but the researcher should first check that the assumptions stated above hold for their data and their experimental design. There may be additional known constraints making the logic of the model of H_1 invalid. For example, in case 1, if it were known that LV would be zero, then the maximum difference estimated is directly determined by the sample difference obtained, and a false theory will never get evidence against it (see (Dienes, 2019), for the severeness of a test as a criterion of how adequately H_1 has been modeled; we discuss this later). A Bayes factor is only meaningful inasmuch as it tests an H_1 that formalizes the predictions of a theory. For case 1, the theory tested predicts that in the population the higher condition is higher than the lower; and that the lower could be any value from baseline (but not lower) up to (almost) the upper condition. For case 2, the theory tested predicts that the population change from pre-test to post-test is greater for the higher condition than the lower; and that all conditions apart from higher post-test could be baseline (but not lower) to (almost) as high as higher post-test.

In addition, the researcher should always conduct robustness checks, as we will demonstrate in the case study below, to examine the extent to which their conclusion depends on the particular model of H_1 they have chosen. In practice we formalize this requirement by specifying Robustness Regions (Dienes, 2019).

Having set up the logic of the motivated-maximum approach, we will now go on to present: 1) a case study of its use in practice and 2) results from simulations investigating how the approach fares across a range of possible data sets.

Note: code for the case study is embedded below and can also be found on github <https://github.com/silveycat/bayes-factor/tree/master> along with relevant data files. This study was not pre-registered.

Results

Case study

We now demonstrate how to apply the motivated-maximum approach to calculate a Bayes factor, using data and analyses drawn from Singh et al. (2021). The authors investigate the development of children's knowledge of graphotactic (spelling) patterns. In the specific analysis reproduced here, they contrast the knowledge of children who learned graphotactic patterns implicitly (by mere exposure) and children who learned the same patterns explicitly (by being told the rule that generates the patterns). In both experiments, children's knowledge is assessed via a 2AFC 'fill-in-the-blanks' task where they are given the beginning and end of a word and asked to drag the correct vowel into the gap. Figure 1 shows mean accuracy in the two experiments.

As a preliminary step, we run a mixed-effects model predicting accuracy in the test from experiment (implicit vs. explicit learning). Note that in the output below, `experiment.ct` is the predictor representing the contrast between experiments with a centered coding.

```
# this code uses the lme4 package (Bates D, Mächler M, Bolker B, Walker S (2015).
# "Fitting Linear Mixed-Effects Models Using lme4")

model_fill <- glmer(accuracy ~ experiment.ct + (1|participantCode),
                    control=glmerControl(optimizer = "bobyqa"), family = binomial,
                    data = lme_Children_implicit_vs_explicit_fill.df)
```

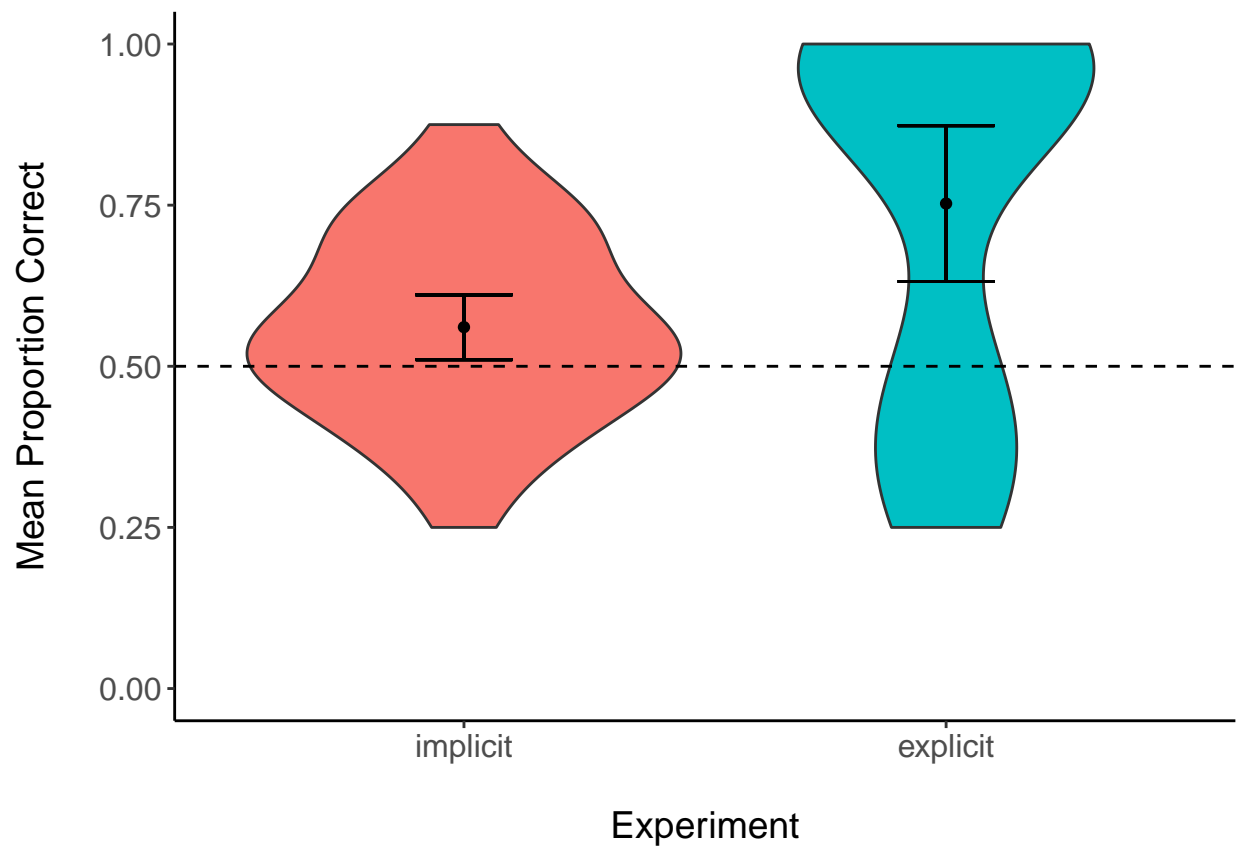


Figure 1: Children's accuracy on fill-in-the-blanks task after implicit or explicit learning of spelling rules

```
output <- summary(model_fill)
print(output)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
## Approximation) [glmerMod]
## Family: binomial ( logit )
## Formula: accuracy ~ experiment.ct + (1 | participantCode)
## Data: lme_Children_implicit_vs_explicit_fill.df
## Control: glmerControl(optimizer = "bobyqa")
##
##      AIC      BIC   logLik deviance df.resid
##  1132.8   1147.4   -563.4   1126.8     957
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.0797 -0.9227  0.2510  0.7910  1.5059
##
## Random effects:
## Groups           Name          Variance Std.Dev.
## participantCode (Intercept) 1.222     1.105
## Number of obs: 960, groups: participantCode, 60
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.8306    0.1686   4.928 8.31e-07 ***
## experiment.ct    1.3310    0.3481   3.823 0.000132 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr)
## expmmt.ct 0.131
```

The frequentist analysis shows us a significant effect of experiment, with a coefficient of 1.33 ($z = 3.82$, $p < .001$). Proceeding to our Bayes factor analysis, H_1 is that explicit learning will lead to better performance in the test – but how much better? Using the logic above, the maximal difference between the conditions in log-odds is equal to twice the difference between the grand mean in log-odds – i.e. the intercept from our mixed model – and chance – i.e. 0 (50% in log-odds). Thus the estimated effect we use as our scale factor for the H_1 distribution is therefore half this value, so the intercept itself, 0.83. The mixed-effects model also gives us the other element we need in order to calculate a Bayes factor: a summary of the data, in the form of the estimate for the effect of experiment (1.33) and the standard error of this estimate (0.35). First, we save the three parameters we need from the mixed-effects model: the estimate for the intercept, and the estimate and standard error for the effect of experiment:

```
intercept <- output$coefficients["(Intercept)", "Estimate"]
mean_effect <- output$coefficients["experiment.ct", "Estimate"]
se_effect <- output$coefficients["experiment.ct", "Std. Error"]
```

To calculate the Bayes factor, we then plug these values into an R code adaptation of the Dienes calculator (Dienes, 2008) ⁶ by Bence Palfi based on original code by Baguley & Kaye (2010) which is available here <https://github.com/bencepalfi/dienes-calculator>

⁶The calculator uses numerical integration to combine the likelihood function and the model of H_1 . Note the same results could be obtained by using the Savage Dickey method (Bartoš & Wagenmakers, 2023)

<https://users.sussex.ac.uk/~dienes/inference/Bence%20Bayes%20factor%20calculator.html>); see also the following shiny app implementation <https://harry-tattan-birch.shinyapps.io/bayes-factor-calculator/> by Harry Birch). To explain the parameters the function takes: `sd` and `obtained` are the standard error and the estimate of the effect of experiment taken from the mixed-effects model (i.e., the summary of the data). `likelihood` specifies our model of the likelihood function, which here is normal (a t distribution may be preferred for a small sample); `modeloftheory` specifies the distribution of our H_1 , which here is normal (uniform, Cauchy, and t are alternative possibilities); `tail` specifies whether we want to model H_1 as a normal or a half-normal distribution. The scientific theory we are interested in is that explicit learning will be stronger than implicit learning. This is one-tailed, we therefore we therefore model H_1 as a half-normal distribution by specifying `tail = 1`. H_1 as a half-normal distribution by specifying `tail = 1`.

```
# This code uses the BF function
# authored by Bence Palfi We enter our
# sample statistics (sd and obtained)
# the likelihood parameter gives us the
# assumed distribution of the data
# (normal) modeloftheory, modeoftheory,
# scaleoftheory and tail provide the
# model of h1 (i.e. a half normal with
# a mean/mode of 0 and an sd of 1)

bf <- Bf(sd = se_effect, obtained = mean_effect,
  likelihood = "normal", modeloftheory = "normal",
  modeoftheory = 0, scaleoftheory = intercept,
  tail = 1)
print(bf)
```

```
## [1] 386.2802
```

We obtain a Bayes factor of 386.28 in favor of H_1 . This represents extreme evidence for H_1 over H_0 . The standard way to report this is to include the distribution we used to represent H_1 in subscript: $BF_{HN(0,0.83)} = 386.28$. Here, HN denotes that we are using a half-normal distribution, the first number in parentheses indicates the mode of this distribution, and the second number indicates the scale factor (here, the standard deviation).

A good additional step to take in any Bayes factor analysis is to examine the extent to which the results are sensitive to the particular value chosen for the scale of H_1 . The standard way of doing this is to report robustness regions (Dienes, 2019), or the range of values for which the qualitative conclusion holds: here, that we have at least moderate evidence for H_1 . For this case study, we test a wide range of values that are reasonable given the scale of the dependent variable. Specifically, we test a range of predicted effects from 0 (no difference between experiments) to 4.595 (participants in one experiment being at chance and participants in the other experiment getting 99% correct). If the resulting Bayes factor is greater than 3 for a wide range of theoretically plausible values, we can consider our conclusion relatively robust.

The following code specifies a range of values, and uses bespoke function `Bf_range` to calculate Bayes factors using each of these values as scale factor for the H_1 distribution (code for `Bf_range` can be found within https://github.com/silveycat/bayes-factor/blob/master/BF_paper.Rmd)

```
# set up a vector with a range of values in logodds space
# from qlogis(50/100) (50% = chance) to qlogis(99/100) (99% is near ceiling)
# in incremental steps of 0.001

h1_range = seq(from=qlogis(50/100),to=qlogis(99/100),by=0.001)
```

```

range_test <- Bf_range(sd= se_effect,  obtained =mean_effect,
                      likelihood = "normal",
                      modeloftheory = "normal",
                      modeoftheory=0,
                      sdtheoryrange= h1_range,
                      tail=1)

# find values for which BF > 3

ev_for_h1 <- subset(data.frame(range_test), BF > 3)
low_threshold <- min(ev_for_h1$sdtypeory)
high_threshold <- max(ev_for_h1$sdtypeory)
print(low_threshold)

```

```
## [1] 0.104
```

```
print(high_threshold)
```

```
## [1] 4.595
```

We find a Bayes factor of greater than 3 starting from an estimated effect of 0.104 and this holds for the remainder of values that we tested up to our ceiling of an estimated effect of 4.595. We can report this as follows: $RR_{BF>3}[0.104, > 4.595]$.⁷

As with every step of this analysis, we must evaluate this robustness range in light of our theoretical knowledge to determine how it should affect our confidence in our conclusions. In this case, unless we had a strong *a priori* reason to expect an extremely small difference between experiments (equivalent to participants in one experiment performing at .50 proportion correct and participants in the other performing at .53), we can be confident that our data favor H_1 over H_0 to at least moderate standards of evidence.

This study demonstrates one situation where using the motivated-maximum approach to derive an estimate for H_1 from the output of a mixed-effects model works well. However, the difference between experiments here is large enough that a Bayes factor favoring H_1 would be likely regardless of how exactly we model the alternative hypothesis. How does the motivated-maximum approach perform across a range of possible datasets with different effect sizes? We set out to answer this question using simulations.

Simulations

General framework

We simulate two different experiments. Simulation 1 models a design with one binary within-subjects predictor, analogous to Case 1. Simulation 2 models a design with an interaction between two binary predictors, one within-subjects and one between-subjects, analogous to Case 2. Both simulations model a binary dependent variable, analogous to a 2AFC test, with 20 trials per participant. We vary performance in each condition from .5 proportion correct (chance) to .95 (near ceiling). This results in a range of possible values for baseline performance l (i.e. the condition(s) with the predicted lowest performance) and true effect size d . In the main paper, we report simulations with the number of participants N set to 40; in the Supplemental Materials, we report additional simulations with $N = 200$.

⁷There will always be a true upper limit to the range of value of H_1 yielding a Bayes factor above a threshold value (since as H_1 increases so does evidence for the null. In the current example, the true upper limit is 347.7 so we could alternatively write the range as $RR_{BF>3}[0.104, 347.7]$. Also note that when looking for the range of values that yield evidence for the null, the upper end of the range is always ∞ .

Table 1: Classification of Bayes factors into evidence categories, drawn from Jeffreys (1939) and Lee and Wagenmakers (2014).

Bayes factor	Category
$BF \leq \frac{1}{100}$	Extreme evidence for H_0
$\frac{1}{100} < BF \leq \frac{1}{30}$	Very strong evidence for H_0
$\frac{1}{30} < BF \leq \frac{1}{10}$	Strong evidence for H_0
$\frac{1}{10} < BF \leq \frac{1}{3}$	Moderate evidence for H_0
$\frac{1}{3} < BF \leq 3$	No evidence either way
$3 < BF \leq 10$	Moderate evidence for H_1
$10 < BF \leq 30$	Strong evidence for H_1
$30 < BF \leq 100$	Very strong evidence for H_1
$100 < BF$	Extreme evidence for H_1

For each combination of l and d , we simulate 20 datasets. We then analyse each dataset using a binomial generalized linear mixed-effects model implemented using the lme4 library in R (Bates, Mächler, et al., 2015), including a random intercept for participant and a random slope for the within-subjects predictor. We then calculate a Bayes factor using a half-normal model of H_1 with mode 0 and standard deviation s . We set s according to the motivated-maximum approach described above. For Case 1, we calculate the Bayes factor for the main effect of the within-subjects predictor, with s set to equal the intercept term from the mixed-effects model. For Case 2, we calculate the Bayes factor for the interaction between the two predictors, with s set to equal twice the intercept from the mixed-effects model. In the Supplemental Materials, we additionally report results from simulations using an alternative estimate for Case 2, where s is set to the main effect of test-session from the mixed-effects model.

Since we generate and analyse 20 datasets for each combination of parameters, we also generate 20 unique Bayes factors which will naturally vary. To characterize the most common qualitative outcome, in the figures below we plot the modal Bayes factor category across the 20 analyses. Table 1 shows how we categorized each Bayes factor (Jeffreys, 1939; Lee & Wagenmakers, 2014).

Full details of the simulations are given in the Supplemental Materials. All code and output is made publicly available on GitHub at <https://github.com/silveycat/bayes-factor>.

Case 1

In Simulation 1, we model an experiment corresponding to Case 1, with one binary within-subjects predictor. We model H_1 as a half-normal distribution with a mode of 0, with the standard deviation set to equal the intercept from the mixed-effects model. Do the Bayes factors we obtain using this method correspond to what we would expect, i.e., showing evidence for H_0 where effects are zero or near-zero, showing evidence for H_1 where effects are large, and showing no evidence where the small size of the effect and the imprecision of the estimate make the data ambiguous?

Figure 2 shows the modal Bayes factors that result from our simulations. As in Case 1, we can think of this simulated experiment as comparing participants’ performance following two different types of training (LV vs HV). The x -axis shows performance in the HV condition in log-odds. The y -axis shows the true effect of condition, or the log-odds difference between performance with the HV and LV items. We see that evidence for H_1 accumulates rapidly from an effect size of around 0.5 log-odds and higher. Evidence for the null accumulates more slowly (Weiss, 1997), as we can see from the wider bands of blue in Figure 2. The dotted line marks a true effect of 0. We observe that along this line and slightly above it, i.e., where effects are 0 or very close to 0, we get evidence for H_0 . We also consistently obtain evidence for H_0 where the effect is below 0, i.e., in the opposite direction from our prediction, with the degree of evidence increasing the larger the magnitude of the negative effect. While in this particular paradigm a negative effect is implausible in many cases (i.e. in all those cases where it means that the LV condition is below chance) it is important to

note that it is reasonable that the evidence for H_0 should increase in this way. To a researcher new to Bayes factors, this may seem counter-intuitive – doesn't a large negative evidence suggest evidence for an effect in the opposite direction, rather than evidence for H_0 ? But the key thing to remember about Bayes factors is that they test only the relative probability of the data under two competing hypotheses. Where H_1 is modeled as a half-normal with a mode of 0, instantiating our directional prediction, a negative effect is thus more consistent with H_0 than with H_1 .

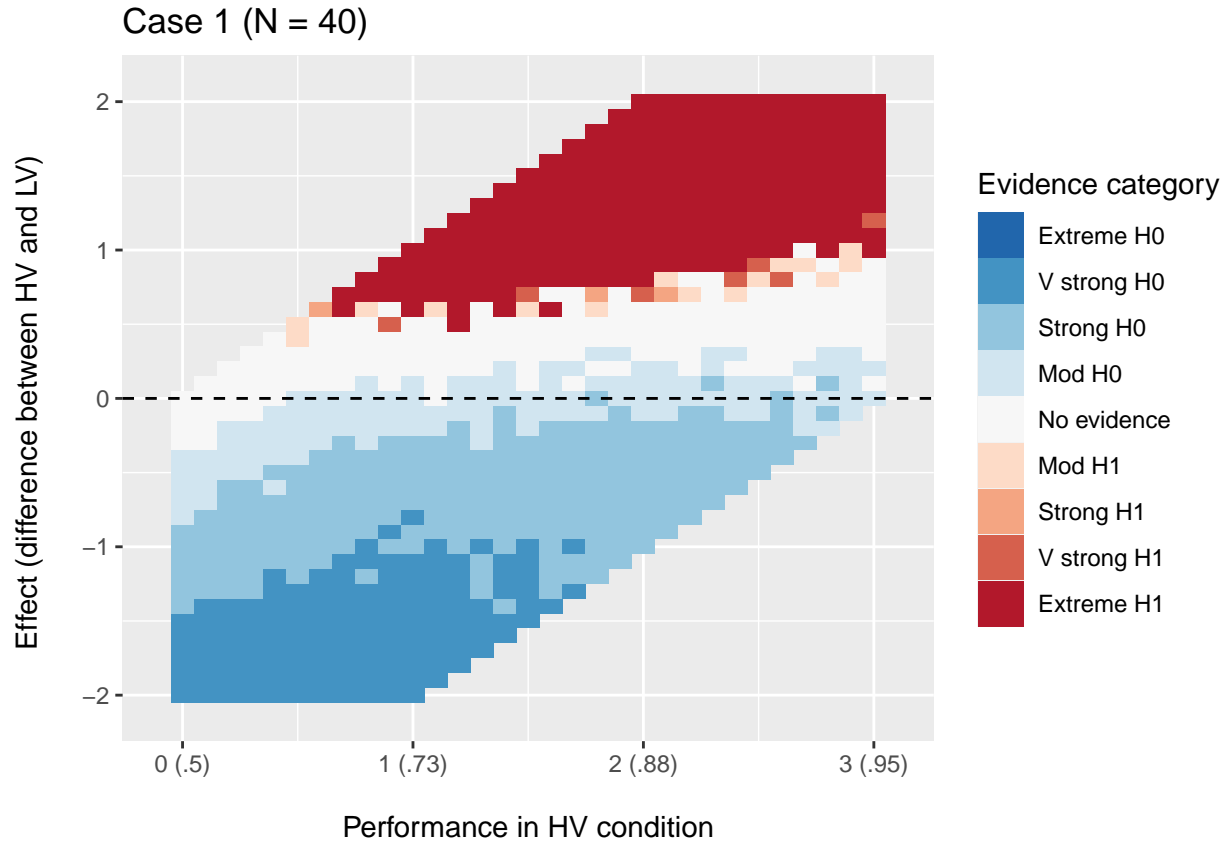


Figure 2: Results from a simulation calculating Bayes factors for the effect of a binary within-subjects predictor in a sample of 40 participants, using the intercept from a mixed-effects model as the estimated effect. The x -axis shows performance in the cell predicted to be highest (e.g., performance in HV), and the y -axis shows true effect size (e.g., performance in HV minus performance in LV). Units are log-odds, with corresponding proportion correct in parentheses. Colours show modal Bayes factor category across 20 generated datasets and analyses.

There are two further observations we can make from Figure 2. One is that we do not always see evidence for H_0 where the effect is truly 0. Towards the left of the plot, where the true effect is 0 but performance in the HV condition is lower than around 0.5 log-odds (corresponding to .62 proportion correct), we instead see no evidence either way. This happens because performance in both cells of the design is low, meaning the intercept of the mixed-effects model is correspondingly low: effectively, there is less ‘room to move’ and we are therefore looking for a smaller effect. All other things being equal, small effects are harder to find evidence both for and against. The outcome shows that if the condition meant to have the highest score is in fact very low, there is a de facto floor effect, there is no room to find a difference between conditions; the Bayes factor appropriately reflects this state of affairs by indicating no evidence.

A more general observation is that, unless the effect is in the opposite direction from predicted- which is implausible for all cases where this would mean performance in the LV condition is below chance- we rarely get more than moderate evidence ($BF < \frac{1}{3}$) for H_0 in a sample of 40 participants. The results in the

supplementary materials show that this can be improved by increasing N – with $N=200$, the modal Bayes factor category is strong evidence for H_0 ($BF < \frac{1}{10}$). (Also, with regard to floor effects, while we still see a region of no evidence where baseline performance is low, this is less pronounced than for the smaller sample). However, obtaining a large sample may be infeasible, for example in developmental research. Thus, we may instead wish to consider this in setting our Bayes factor thresholds. We return to consider these implications in the Discussion.

Case 2

Next, we simulate the more complex Case 2, involving an interaction between two binary predictors. As outlined above, this simulates an experiment where participants in two conditions (low-variability training, LV, and high-variability training, HV) take part in a pre-test and a post-test. Here, we are interested in the interaction effect, or how improvement from pre-test to post-test differs between conditions. For the cases illustrated below, we assume that performance in the pre-test is equivalent across the two conditions. In the Supplementary Materials (Section 2 Results for Interaction when pre-test varies across condition) we demonstrate that when this assumption is relaxed the overall pattern of evidence for H_0 and H_1 is reasonably similar).

We consider two plausible situations: a) performance in the pre-test in both conditions is at chance (.5 proportion correct = 0 log-odds); b) performance in the pre-test in both conditions is above chance (.73 proportion correct = 1 log-odds, similar to average pre-test performance in Logan et al. (1991)). In the following plots, the x-axis shows post-test performance in the condition we expect to improve more under H_1 , i.e., the HV condition. As for Case 1, the y-axis shows the veridical effect we are interested in: here, the interaction effect. Because we assume pre-test performance is equivalent across conditions, the interaction here equates to the difference in post-test performance between conditions.

Figure 3 shows the results in the situation where performance in the pre-test in both conditions is at chance. We can see that, compared to Case 1, the region of no evidence is wider: with the same sample size, it is harder to get evidence either for H_0 or H_1 for an interaction. Again, this is expected, and provides further reassurance that the motivated-maximum approach produces intuitively sensible results.

Figure 4 shows the results in the situation where performance in the pre-test in both conditions is above chance. We see that in this case, since the grand mean is higher and we are therefore looking for a larger effect, we obtain evidence for H_0 in a wider range of cases where the interaction effect is non-zero. We may then prefer to use an estimate based on the main effect (see Supplemental Materials) as a more conservative approach when seeking evidence for H_0 with this type of experimental design. However, as for every approach put forward in this paper, it should not be applied as a default in every situation. Where we have reason to expect that mean performance will be relatively low, either from previous results or the current study, the estimate based on twice the grand mean may be appropriate.

More generally, we emphasize the importance of taking into account the specific theoretical background and experimental design of the study in question when working out a motivated maximum. For example, in a situation where the two predictors were training variability (LV vs. HV) and item novelty (seen vs. unseen), reasonable assumptions could lead to using the intercept (rather than twice the intercept) as an estimate for the interaction effect (see Supplemental Materials for an outline of the logic). Indeed, when the intercept is low relative to chance, it can serve as an estimate for both main effects and interactions (Dienes et al., 2012; Ziori & Dienes, 2015).

Model comparison method

The simulations reported above and in the Supplemental Materials demonstrate that the motivated-maximum approach, where we use estimates from mixed-effects models to inform our model of H_1 , produces intuitively sensible results. To produce these results, we used an R code adaptation of the Dienes calculator (Dienes, 2008). This calculator takes as input a summary of the data and a distribution for H_1 . The

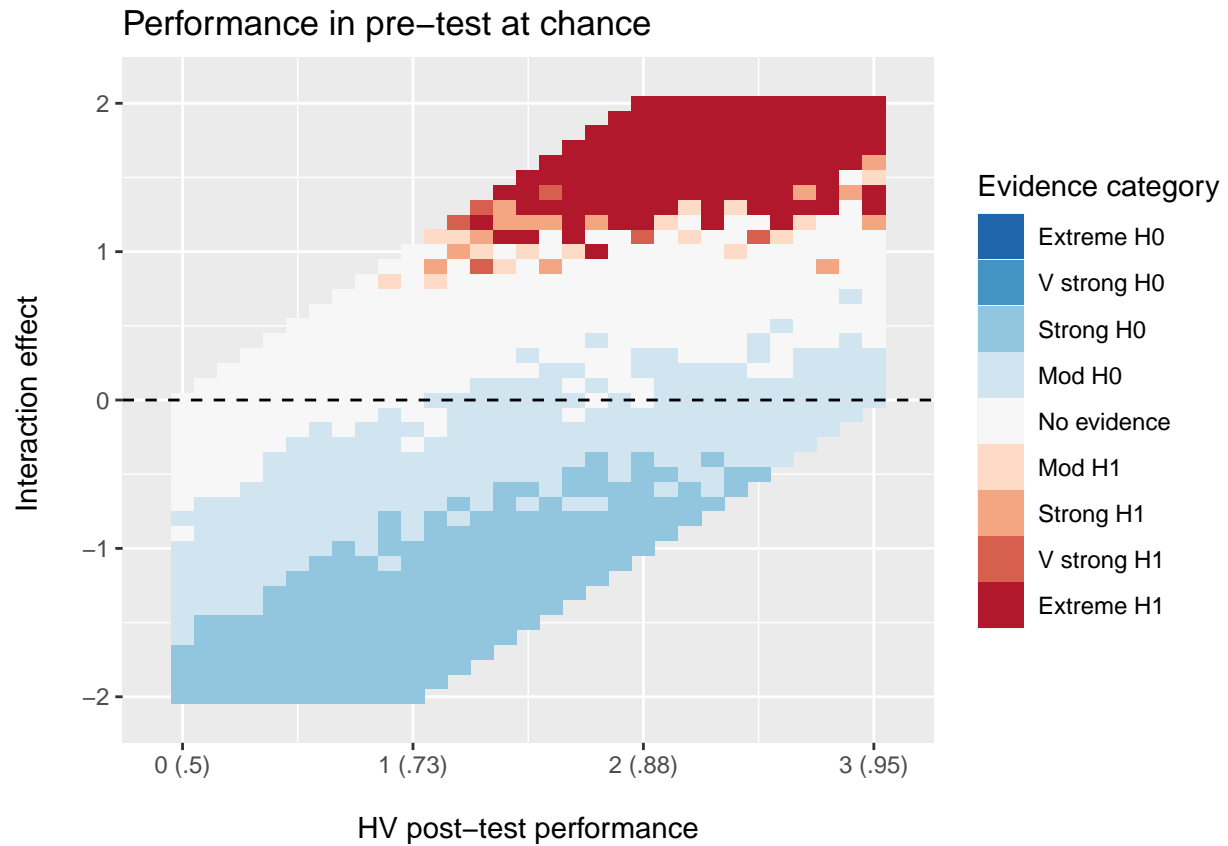


Figure 3: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using $2 \times$ the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .5 proportion correct (0 log-odds). The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

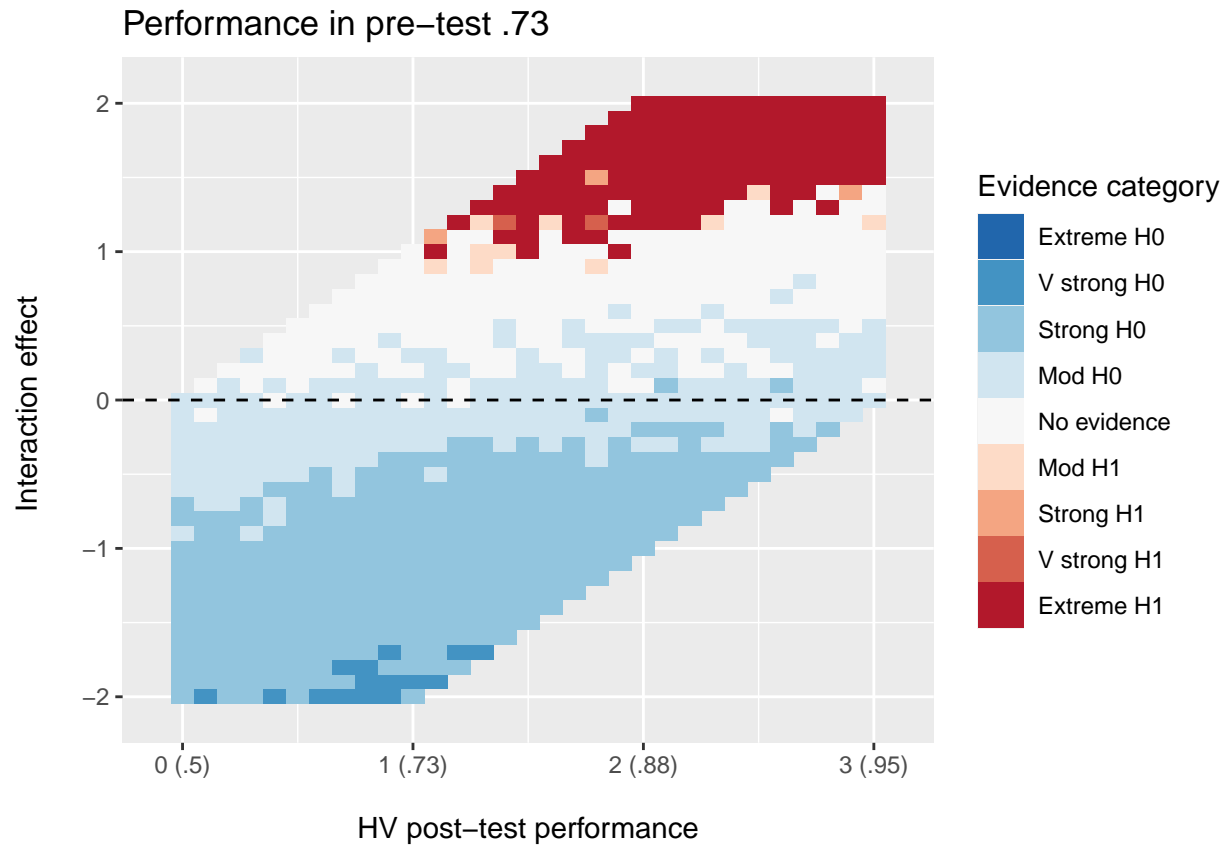


Figure 4: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using $2 \times$ the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .73 proportion correct (1 log-odds). The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

calculator then approximates the probabilities of the data under H_1 and a point null hypothesis H_0 and returns their ratio, i.e., the Bayes factor.

An alternative approach in the same spirit is to use Bayesian model comparison (Schad et al., 2021). The first step in this method is to build a full Bayesian mixed model of the data. Such models use Bayesian methods to estimate the parameters of mixed-effects models, returning an estimate and an associated credible interval within which we can be confident to a specific probability that the true value of the parameter lies. While credible intervals are sometimes used to argue for the presence of an effect (e.g., if the 95% credible interval does not cross 0), this is not an appropriate inference: the credible interval tells us nothing about the likelihood of the data under H_0 (Schad et al., 2021). Instead, the appropriate way to use these models to quantify evidence for H_1 compared to H_0 is to compute a Bayes factor based on comparing two models. It is important to realize that this method also requires the researcher to require define an informed prior for the hypothesis being tested- i.e. a model of H_1 . Specifically, the researcher must specify a full mixed model which includes the parameter of interest and, as in the Dienes calculator approach, the prior for this parameter represents H_1 . We can use the motivated maximum approach to determine this prior, just as we did for the Dienes calculator.

The researcher specifies a full mixed model which includes the parameter of interest; as in the Dienes calculator approach, the prior for this parameter represents H_1 . The researcher then specifies a null model where the parameter of interest is excluded, meaning it is set to zero. The probability of the data under each model is then approximated using bridge sampling (Gronau et al., 2020). Bridge sampling is more robust than our method, which assumes approximate normality of the likelihood, but our method is computationally easier and faster. Comparing entire models, rather than using a summary of the data is also likely more robust; however, it also requires the researcher to have expertise in running fully Bayesian analyses.

As a partial check on whether these conceptually similar approaches in fact produce similar results, we compared the Bayes factors resulting from each approach for a subset of the datasets we simulated for Case 1 (the simple design with one within-subjects predictor, with a sample size of $N = 40$). For the model comparison approach, we used brms (Bürkner, 2017) to construct the full model and the null model. The model had an identical structure in terms of fixed effects and random effects to the lmer model used to derive the data summary for the calculator.

Default priors (see supplementary materials for details) were used for all parameters except the main effect of interest. Note that this follows the principle that informed priors should be used for hypothesis testing, which applies to the effect of interest, while vague default priors are appropriate for estimation which is what we wish to do for all the effects which we are not of direct interest. The prior for the main effect of interest was the same as the H_1 distribution we used for the Dienes calculator approach: a half-normal distribution with a mode of 0 and a standard deviation set to equal the intercept from the model of the data. The null model was identical to the full model except that the fixed effect of interest was set to 0.⁸ Note that since we use the same priors for H_1 as with the calculator method, we still apply the motivated maximum approach. Our comparison between the methods here is thus not about how we obtain the prior, but rather on whether the results change when we compare entire models rather than extracting a data summary for the calculator. Full details of the model comparison process are provided in the Supplemental Materials.

Figure 5 shows scatter plots comparing the Bayes factors computed by the two approaches, separating cases where H_1 is true (left) and H_0 is true (right). While the points are not quite evenly scattered around the line of identity, they fairly closely follow the line and the departures do not involve any qualitative change in what conclusion would follow from the Bayes factors.

⁸Note that the random by participant slope for condition is thus maintained in the null model- only the fixed effect is removed. Doorn et al. (2021) and Linde & Ravenzwaaij (2021) refer to this as a “balanced null”, which is recommended for testing differences between means

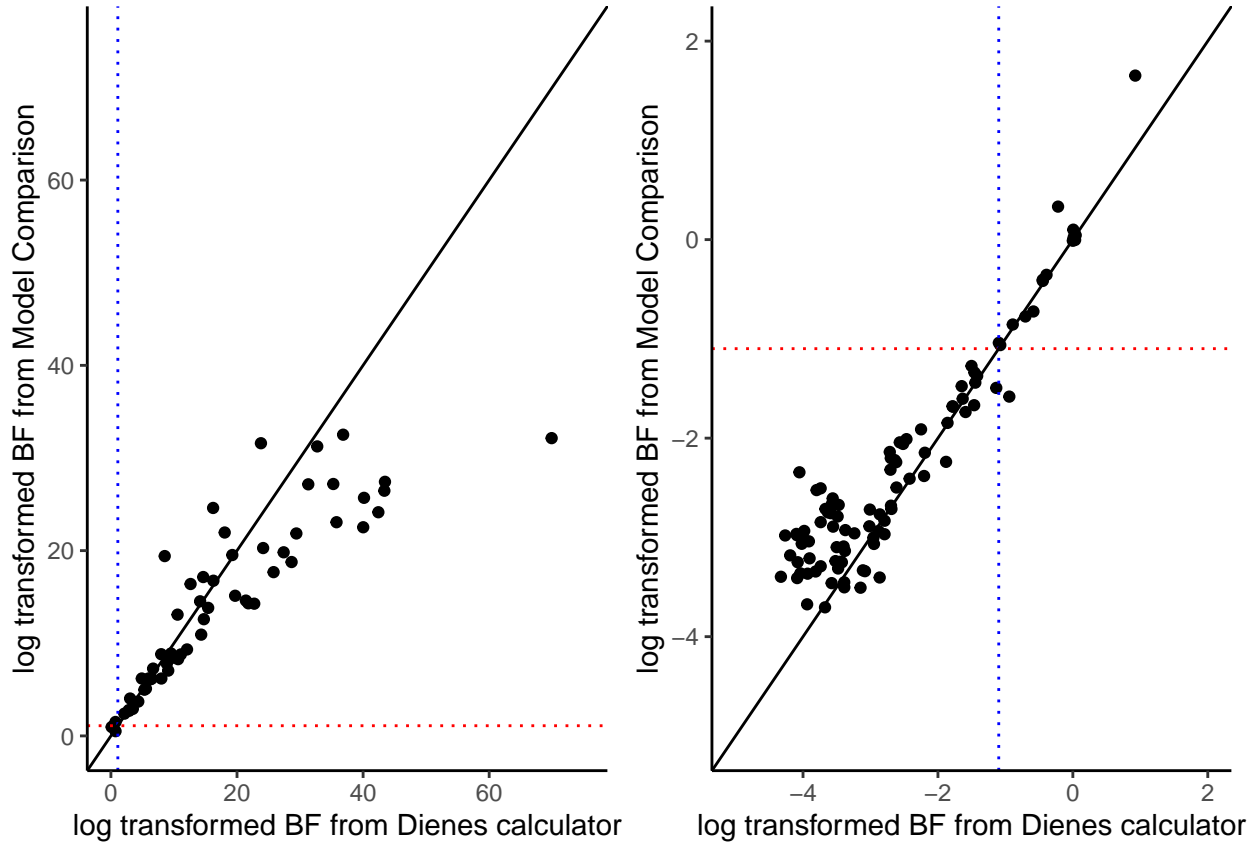


Figure 5: Scatter plot of the Bayes factors (log transformed) produced by the Dienes Calculator and the Model Comparison Methods. Different data points are from different runs with different effect sizes and baselines. We separate the case where H_1 is true (left) and H_0 is true (line). The solid black line is the line of identity which shows where the ratio of the two Bayes factors is 1. The dotted lines shows where the threshold of substantial evidence for H_1 ($BF > 3$) and for the null ($BF < \frac{1}{3}$) was found with each method (Dienes calculator blue, Model comparison red.)

Discussion

We demonstrate a method for calculating Bayes factors to quantify the evidence for H_0 and H_1 , using H_1 distributions constrained by estimates from mixed-effects models in what we call the motivated-maximum approach. We show using simulations that modelling H_1 as a half-normal distribution with the standard deviation set to the intercept (in the case of finding evidence for or against a main effect) or twice the intercept (in the case of finding evidence for or against an interaction) produces intuitively sensible results across a range of plausible cases. It is important to emphasize, however, that this approach should not be applied as a default in all situations. Researchers should base their model of H_1 on theoretical expectations informed by previous literature. However, our intent here is to provide an option researchers can consider using in cases where an estimate of the predicted effect is not directly available and it has the further advantage that it can be used in conjunction with frequentist mixed-effects models, which are already widely used across the psychological sciences. Note that, critically, the model we use embodies theoretical claims: i) the effect occurs in a certain direction; ii) there is a minimum value of performance we can expect to see in any cell (chance in our example) and iii) it is possible that the intervention postulated might not move performance above this minimum level.

The results also illuminate some general implications for Bayes factor analyses, both using this method and more generally. Firstly, in the case of low overall performance, we do not find evidence for H_0 even when the effect in question is veridically zero. This problem is exacerbated for small sample sizes (see Supplemental Material). Secondly, finding evidence for or against an interaction is much harder than finding evidence for or against a main effect. Thirdly, it is harder to get strong evidence for H_0 than strong evidence for H_1 . Indeed in case 1, when $N=40$, when the effect was veridically 0 the Bayes factors generally met the criteria for moderate ($BF > \frac{1}{3}$) but not for strong ($BF < \frac{1}{10}$) evidence. As suggested by Weiss (1997), this may be practical grounds for using asymmetric thresholds for accepting evidence for H_0 versus H_1 (see also Jeffreys, 1939). For example, a researcher might choose to accept only a Bayes factor above 10 as evidence in favor of H_1 , while accepting a Bayes factor below $\frac{1}{3}$ as evidence in favor of H_0 . Bear in mind this implicitly presumes that H_0 is more a priori likely than H_1 ⁹

Is it legitimate to base the model of H_1 (the “prior”, after all) on aspects of the data? Heide & Grünwald (2021) argue that priors where part of the model is fixed by the data should not be used with optional stopping because the concept of “calibration” (i.e. whether the Bayes factor reflects evidence appropriately) does not apply to such priors (because there is no fixed model of H_1 to test calibration for). While their argument concerned optional stopping (stopping when the Bayes factor reaches a desired level of evidence) one might also wonder if the point does not apply generally to any use of Bayes factors. However, one can still show whether the model of H_1 does reflect evidence appropriately by another means, namely by the concept of severe testing (see Vanpaemel, 2020). A severe test is one that can generate an extreme Bayes factor against a theory if the theory is false (Dienes, 2023). A Bayes factor can behave like this because it is a measure of evidence. Consider for example using the effect size in the sample to fix the predicted effect size in the model of H_1 , in order to test whether that effect exists using the very same sample. Clearly in this case one could not get evidence against a theory predicting an effect, even if the theory were false. Thus the concept of a severe test shows that it is inappropriate to fix e.g. the SD of a half-normal (or the scaling factor of a Cauchy) to the value of the sample effect size (Dienes, 2023). But in terms of the heuristic we advise, comparison of Figure 2 against figures 4 and 5 in the supplementary materials where the N is increased, shows a study can provide a severe test of a theory, whether the theory predicts an effect or no effect. That is, if either H_0 or H_1 is true, as N increases the Bayes factor tends to become more extreme in support of the true hypothesis and against the false one (cf. Dienes, 2019).

One might still ask, is it not better though to always fix the model of H_1 without reference to the data if one can? On the contrary it can be preferable to use the data just as we do. Consider the High Variability/Low Variability (HV vs LV) study we illustrated in case 1 and where the intercept is very small because HV produced little effect. The Bayes factor would be close to 1. Whether this is appropriate depends on the context and our assumptions. First take the situation where it was taken as given from past

⁹Note that a likely consequence of this decision is that evidence in favor of H_0 will be less robust than evidence in favor of H_1 . A sensitivity analysis using robustness regions, such as that reported for the case study above, would demonstrate this.

research that the HV condition should be substantially above chance (unless there was something odd about the implementation of learning in this study) and the interest was in whether LV would be lower. In this case, it is in fact appropriate for the Bayes factor to be close to 1, because there is no room to see if LV did worse. If we had instead used an estimate of HV effect from past studies, we might get evidence against the HV-LV difference; but that would be inappropriate. So the data dependent model of H1 automatically deals with what are effectively floor effects because the basic manipulation was not as effective as expected (Dienes, 2019). On the other hand, in the situation where the given was that the LV could be close to chance and the interest was in whether HV would still exceed LV, then use of an estimate of HV from past studies would be appropriate, precisely because one could then obtain evidence for H_0 if it were true. The general point is: Think carefully about whether the test severely tests one’s actual theory.

Finally, we show that using the the motivated-maximum method to inform the prior for H_1 while using estimates from frequentist mixed-effects models and the Dienes calculator, produces similar results to the method of using the same motivated-maximum method to set H_1 but comparing two fully Bayesian models, at least in the limited range of cases which we have tested. Our approach is simpler in that it can be used immediately by researchers already working with lmer without having to learn how to fit Bayesian hierarchical models. It is also notably faster: in a quick comparison, running a mixed-effects model and calculating the Bayes factor using our method took 0.3 seconds, whereas using bridge sampling to compare two fully Bayesian models took 11 minutes. However, there are some situations in which the model comparison approach may be preferred, despite its steeper learning curve for the user. For example, if the mixed-effects model fails to converge, this affects the trustworthiness of both the frequentist results and the Bayes factor based on these estimates. Although it is common in these cases to selectively exclude random effects until the model converges, this approach is not ideal (Eager & Roy, 2017). Indeed, researchers may turn to using Bayesian mixed models primarily because they solve the convergence problems associated with the lmer package. One alternative approach would be to use a Bayesian model implemented in brms (Bürkner, 2017) or equivalent software rather than an lmer model, to generate the parameter estimates and standard errors. Note that in this case, we are just using the model for estimation, and default uninformative priors should be used for all fixed and random effects, including the one of interest. The estimate of the fixed effect of interest and the error term –can then be extracted for the coefficient of interest exactly as with lmer (though note that in brms the equivalent of standard error in lmer is the standard deviation of the posterior distribution for the parameter, labelled as Est.Error). and fed into the Bayes factor calculation for hypothesis testing. While still more time-consuming than running an lmer model, this would provide more robust estimates for use in the Bayes factor calculation, without requiring the researcher to run and compare two Bayesian models.

The motivated-maximum method as presented here has limitations. In particular, the method cannot currently be applied to the case of a continuous predictor. We are currently working on developing an alternative method for this scenario. In the meantime, the ratio of scales or the ratio of means (Dienes, 2019) are appropriate heuristics that can be used in this case. It is also important to reiterate that the simulations reported above assume that performance in all cells of each experiment is at least at chance. If a researcher’s specific experimental design includes a reason why participants might perform below chance (e.g., the lure of a particular distractor), the simulation results may not offer a complete picture of the performance of the motivated-maximum approach when used on the resulting data. More generally, we reiterate that the specific theoretical background and experimental design of a given study should always inform the selection of an appropriate model of H_1 .

Conclusion

We describe a method for using Bayes factors to quantify the evidence for and against a given alternative hypothesis H_1 relative to a null hypothesis H_0 . In particular, we describe how to use the method pioneered by Dienes (2014) in conjunction with estimates obtained from frequentist mixed-effects models, both as a summary of the data and as a way of constraining the expected effect size under H_1 where estimates based on previous literature are unavailable. We call this the motivated-maximum approach. Using simulations, we show that the motivated-maximum approach performs well across a range of veridical effect sizes, and that in simple cases, it produces qualitatively similar results to the method of Bayesian model comparison

when using the same model of H_1 . We hope this paper will make Bayes factors more accessible to researchers who currently use frequentist mixed-effects models in their analyses and who are unsure exactly what effect size they should expect to observe. However, we emphasize that there is no substitute for properly thinking about what a theory predicts, and that as with all approaches, ours should be applied thoughtfully and not as a default.

Author Footnotes

Author affiliations

C. Silvey, UCL, Division of Psychology and Language Sciences

Z. Dienes, University of Sussex, Department of Psychology

E. Wonnacott, University of Oxford, Department of Education (corresponding author)

Author contributions

E. Wonnacott developed the motivated-maximum approach and worked out the estimates for each case. C. Silvey designed, coded and analysed the simulations. Z. Dienes originally developed the method for Bayes factor calculation and advised on its implementation. C. Silvey drafted the paper. All authors contributed to revising the paper and approved the final version.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The research was developed while C. Silvey was funded by a British Academy Postdoctoral Fellowship, Z. Dienes was funded by a grant from the Economic and Social Research Council (ESRC; grant number ES/P009522/1), and E. Wonnacott was funded by ESRC grant ES/K013637/2 and Leverhulme RPG-2019-160.

References

- Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E.-J., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., Vanpaemel, W., Dienes, Z., & van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour*, 4, 561–563. <https://doi.org/10.1038/s41562-019-0807-z>
- 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.
- Baguley, T., & Kaye, W. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by z. dienes. *British Journal of Mathematical and Statistical Psychology*, 63(3), 695–698.
- Barcroft, J., & Sommers, M. S. (2017). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, 27, 387–414. <https://doi.org/10.1017/S0272263105050175>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Bartoš, F., & Wagenmakers, E.-J. (2023). A general approximation to nested bayes factors with informed priors. *Stat*, 12(1), e600. <https://doi.org/10.1002/sta4.600>
- Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv Preprint arXiv:1506.04967*.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, 126, 104352. <https://doi.org/10.1016/j.jml.2022.104352>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, 19(22), 3127–3131.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290. <https://doi.org/10.1177/1745691611406920>
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5(July), 1–17. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z. (2016). How bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78–89.
- Dienes, Z. (2017). *Principles for Bayes, Talk at St Catherines College Oxford*. <https://www.youtube.com/watch?v=9hFN0csyeO4>
- Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, 1–18. <https://doi.org/10.31234/OSF.IO/YQAJ4>
- Dienes, Z. (2021). How to use and report bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice*, 8(1). <https://doi.org/10.1037/cns0000258>
- Dienes, Z. (2023). Testing theories with bayes factors. In J. E. Nichols Austin Lee & Edlund (Ed.), *Cambridge handbook of research methods and statistics for the social and behavioral sciences*. <https://psyarxiv.com/pxhd2>
- Dienes, Z., Baddeley, R. J., & Jansari, A. (2012). Rapidly measuring the speed of unconscious learning: Amnesics learn quickly and happy people slowly. *PLOS ONE*, 7(3), 1–9. <https://doi.org/10.1371/journal.pone.0033400>
- 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.
- Doorn, J. van, Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *Computational Brain & Behavior*. <https://doi.org/10.1007/s42113-021-00113-2>
- Eager, C., & Roy, J. (2017). *Mixed effects models are sometimes terrible*. arXiv. <https://arxiv.org/abs/1701.04858>

- 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(2), 156–168. <https://doi.org/10.1177/2515245919847202>
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10), 1–29. <https://doi.org/10.18637/jss.v092.i10>
- Heide, R. de, & Grünwald, P. D. (2021). Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin & Review*, 28, 795–812. <https://doi.org/10.3758/s13423-020-01803-x>
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4), 434–446. <https://doi.org/10.1016/j.jml.2007.11.007>
- Jeffreys, H. (1939). *The theory of probability*. Clarendon Press.
- Johnson, V. E., Pramanik, S., & Shudde, R. (2023). Bayes factor functions for reporting outcomes of hypothesis tests. *Proceedings of the National Academy of Sciences*, 120(8), e2217331120.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
- Linde, M., & Ravenzwaaij, D. van. (2021). Bayes factor model comparisons across parameter values for mixed models. *Computational Brain and Behavior*. <https://doi.org/10.1007/s42113-021-00117-y>
- Linde, M., Tendeiro, J. N., Selker, R., Wagenmakers, E.-J., & Ravenzwaaij, D. van. (2021). Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the bayes factor. *Psychological Methods*. <https://doi.org/10.1037/met0000402>
- Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, 94(3), 1242–1255.
- Lockwood, P. L., Abdurahman, A., Gabay, A. S., Drew, D., Tamm, M., Husain, M., & Apps, M. A. J. (2021). Aging increases prosocial motivation for effort. *Psychological Science*, 32(5), 668–681. <https://doi.org/10.1177/0956797620975781>
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and/l: A first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type i error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.
- Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs*. <https://CRAN.R-project.org/package=BayesFactor>
- Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part II. *Linguistics and Language Compass*, 10, 591–613. <https://doi.org/10.1111/lnc3.12207>
- Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). *Workflow techniques for the robust use of Bayes factors*. <https://arxiv.org/abs/2103.08744>
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegate, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, 11(9), 1141–1152.
- Singh, D., Wonnacott, E., & Samara, A. (2021). Statistical and explicit learning of graphotactic patterns with no phonological counterpart: Evidence from an artificial lexicon study with 6–7-year-olds and adults. *Journal of Memory and Language*, 121, 104265. <https://doi.org/https://doi.org/10.1016/j.jml.2021.104265>
- Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, 41(4), 795–820. <https://doi.org/10.1017/S0272263119000263>
- Stekelenburg, A. van, Schaap, G., Veling, H., & Buijzen, M. (2021). Boosting understanding and identification of scientific consensus can help to correct false beliefs. *Psychological Science*, 32(10), 1549–1565. <https://doi.org/10.1177/09567976211007788>
- Vanpaemel, W. (2020). Strong theory testing using the prior predictive and the data prior. *Psychological Review*, 127, 136–145.

- Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4), 1457–1475. <https://doi.org/10.1037/a0036731>
- Wei, R., Hu, Y., & Xiong, J. (2019). Effect size reporting practices in applied linguistics research: A study of one major journal. *SAGE Open*, 9(2), 2158244019850035. <https://doi.org/10.1177/2158244019850035>
- Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, 46(2), 185–191.
- Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, 95, 36–48. <https://doi.org/https://doi.org/10.1016/j.jml.2017.01.005>
- Ziori, E., & Dienes, Z. (2015). Facial beauty affects implicit and explicit learning of men and women differently. *Frontiers in Psychology*, 6, 1124. <https://doi.org/10.3389/fpsyg.2015.01124>

Supplemental Materials for ‘Bayes factors for mixed-effects models’

1 Alternative approach to Case 2: using the main effect

As noted in the paper, when deciding on an H_1 distribution for an interaction effect, there is more than one approach we could take. The approach in the paper uses the intercept from the mixed-effects model as a basis for generating a motivated maximum. An alternative approach is to base our estimate on the size of one of the main effects. This is related to the approach suggested by Gallistel (2009), described in Dienes (2019) as a special case of the ‘room-to-move’ heuristic. However, where that approach uses the simple effect (e.g., the difference between pre-test and post-test in the LV condition), the current approach uses the main effect t (the difference between pre-test and post-test, averaged across conditions). The logic here is the same as that outlined for the intercept-based approach in the paper: for the maximum interaction d , we assume that all improvement from pre-test to post-test happens in the HV condition. If this is the case, then improvement from pre-test to post-test in the LV condition is 0, and the difference that represents the interaction effect d is equal to the improvement from pre-test to post-test in the HV condition. In a centered design, the main effect of test-session t is the average of these two values, or $d/2$. The main effect t is therefore half the maximum effect we might observe. We set the standard deviation of the half-normal distribution that is our model of H_1 to equal t , the main effect of test-session from our mixed effects model.

It is an open question which of these two versions of the motivated-maximum approach (using twice the intercept or using the main effect) performs better on average for returning appropriate Bayes factors. In the plot below, we contrast these two approaches. In the situation where performance in the pre-test in both conditions is at chance, the estimate and hence the results are the same for the two approaches. Figure 1 shows the results in the situation where performance in the pre-test in both conditions is above chance (.73 proportion correct, similar to average pre-test performance in Logan et al. (1991)). Here, the estimates from the two approaches diverge, and we can observe which estimate enables us to disentangle H_0 and H_1 most effectively.

From Figure 1, two things are apparent: 1) the estimate based on the intercept can be used in a wider range of situations, because the grand mean remains positive even where the main effect of session is not; 2) the

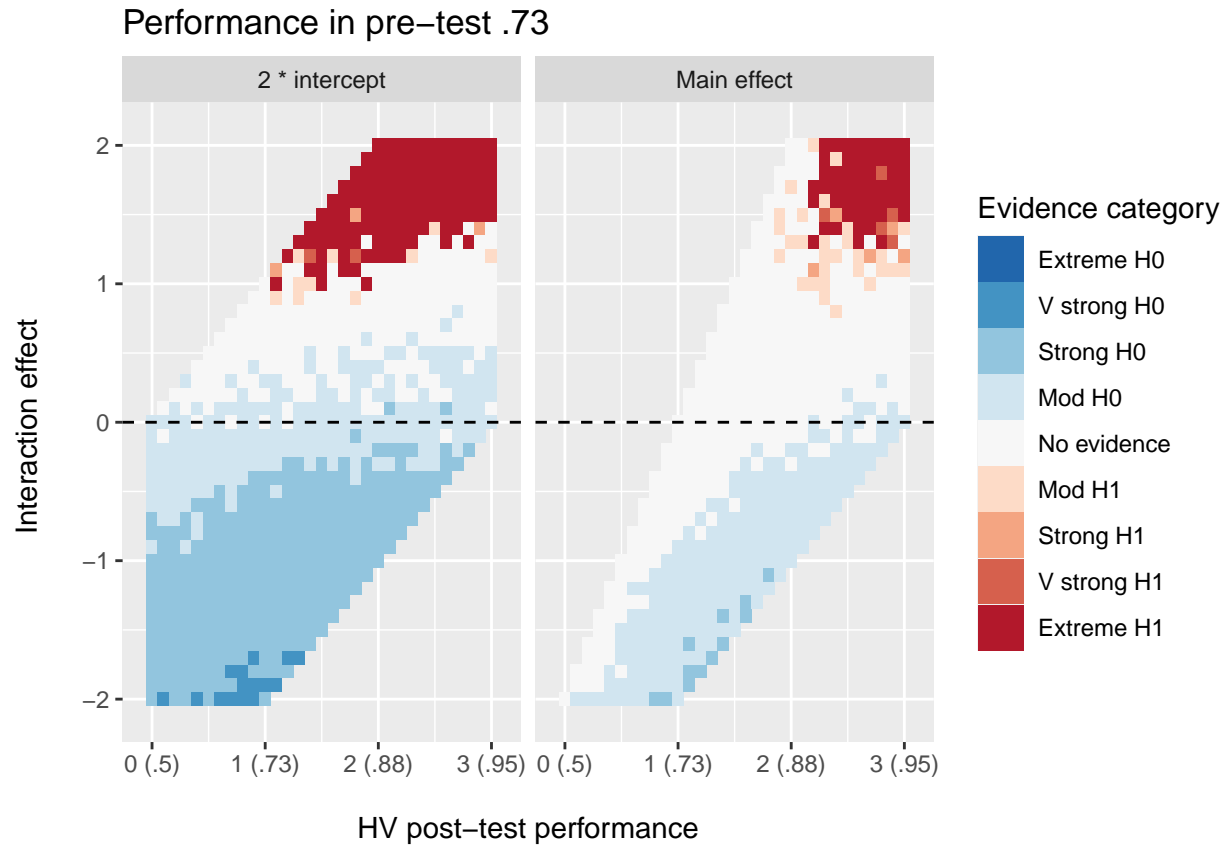


Figure 1: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using either 2* the intercept or the main effect of the within-subjects predictor from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .73 proportion correct (1 log-odds). The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

estimate based on the intercept both suffers less from floor effects: we can obtain evidence for H_0 even where post-test performance in both conditions is low) and even when performance is in the strongest condition is low, it obtains evidence for H_1 more robustly when the veridical interaction is at least 1 log-odds. The estimate based on the intercept tests the theories more severely, that is, is more likely to find them wrong when they are wrong. However, small real effect sizes can result in evidence for H_1 , which is a normative consequence of a measure of evidence (Morey, 2010); as shown in Figure 4 below, when N is increased evidence is obtained for H_1 for increasingly smaller effect sizes.

2 Results for interaction when pre-test varies across conditions

The paper reports results for the interaction case only where pre-test performance is equivalent across conditions. Below we present examples of what happens when this assumption is relaxed and pre-test performance varies across conditions (with a sample size of $N = 40$; as noted above, results where $N = 200$ are similar but with a narrower band of no evidence). For comparison, Figure 3 in the main paper shows the case where performance in the pre-test in both conditions is at chance.

Figure 2 shows the results where performance in the pre-test is at chance in the LV condition and at .62 proportion correct in the HV condition. Figure 3 shows the results in the opposite case, where performance in the pre-test is at .62 proportion correct in the LV condition and at chance in the HV condition. While the range of results we can observe varies (since the range of possible interaction effects is constrained by the difference in pre-test performance), the overall pattern of evidence for H_0 and H_1 is similar across the different cases, suggesting that small differences in pre-test performance across conditions should not affect the applicability of the approach.

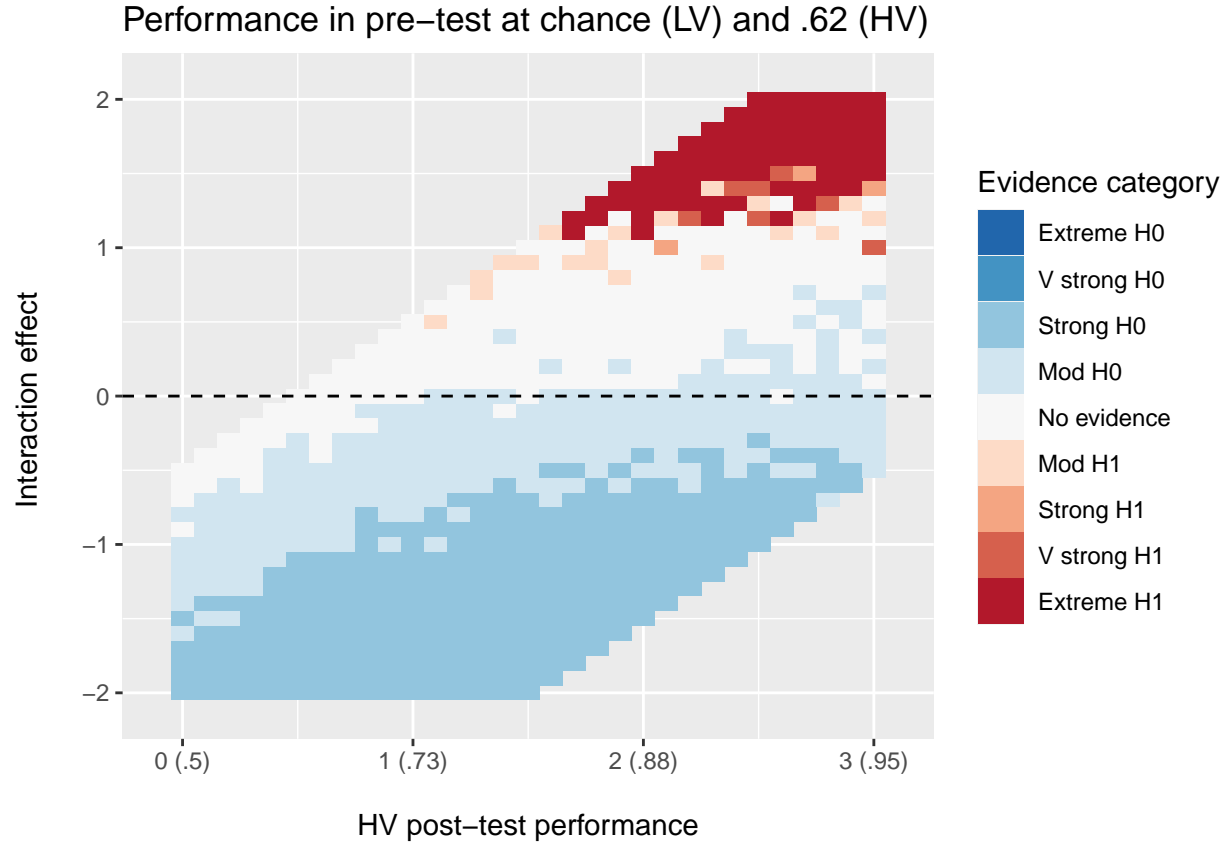


Figure 2: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using 2^* the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .5 proportion correct (0 log-odds) in the LV condition and .62 proportion correct (0.5 log-odds) in the HV condition. The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

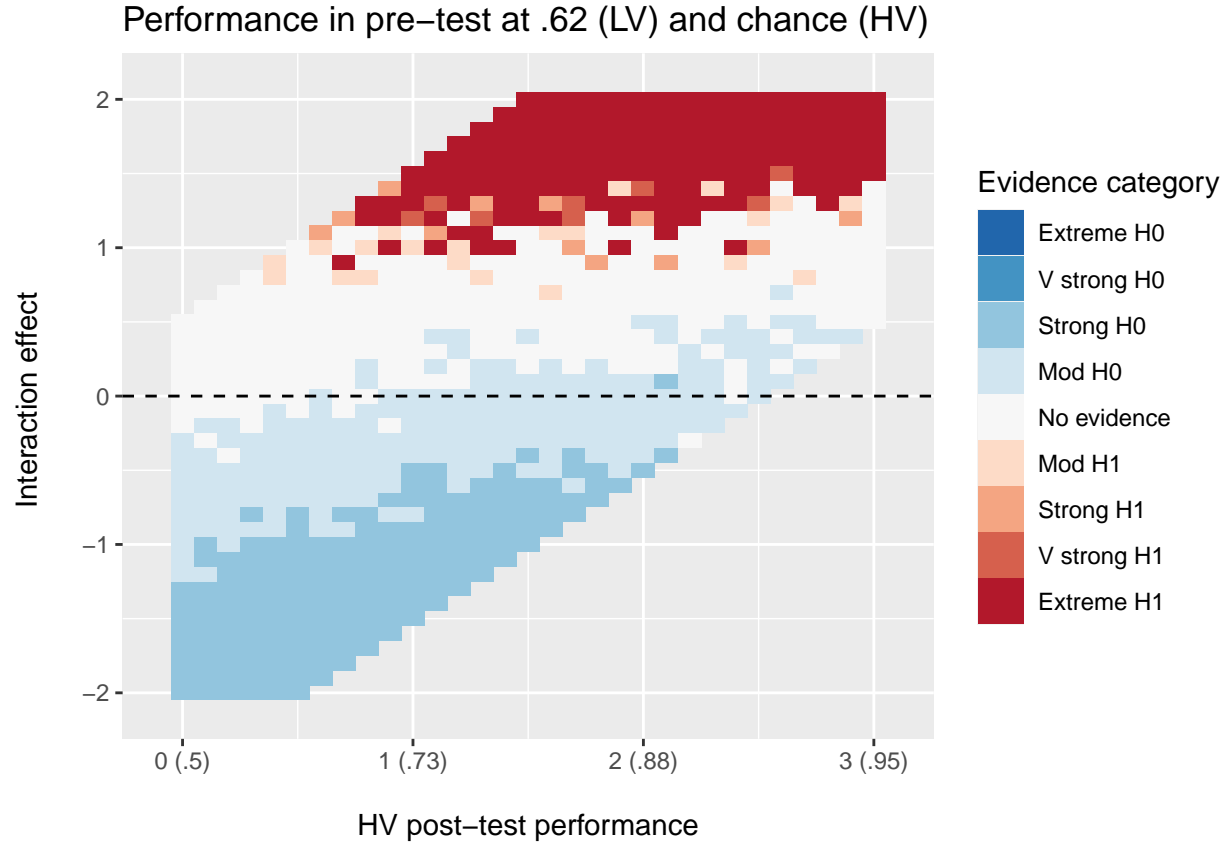


Figure 3: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using $2 \times$ the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .62 proportion correct (0.5 log-odds) in the LV condition and at chance (0 log-odds) in the HV condition. The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

3 Results in $N = 200$ case

The paper reports simulations of a dataset where $N = 40$, a relatively small sample size. When the sample size is larger and we therefore have more information, do we see a narrowing of the band of no evidence, as we should expect?

Figure 4 shows the Bayes factors that result from simulations in Case 1 where $N = 200$. With this larger sample, the band of no evidence has narrowed: we are able to get evidence for H_1 in the case of smaller veridical effects (from an effect size of around 0.25 log-odds and higher). We also get better evidence for the null where the effect is 0 or close to 0, with the modal Bayes factor category in most cases being strong evidence for H_0 . While we still see a region of no evidence where baseline performance is low, this is less pronounced than for the smaller sample: we are able to obtain evidence for H_0 from a lower level of baseline performance. In short, when we have more data, our Bayes factors tend to be more informative, as we would expect.

Figure 5 shows the results in Case 2 with 200 participants, where performance in the pre-test in both conditions is at chance. As for Case 1 above, a higher number of participants gives a narrower region of no evidence, making it easier to get evidence for H_0 or H_1 .

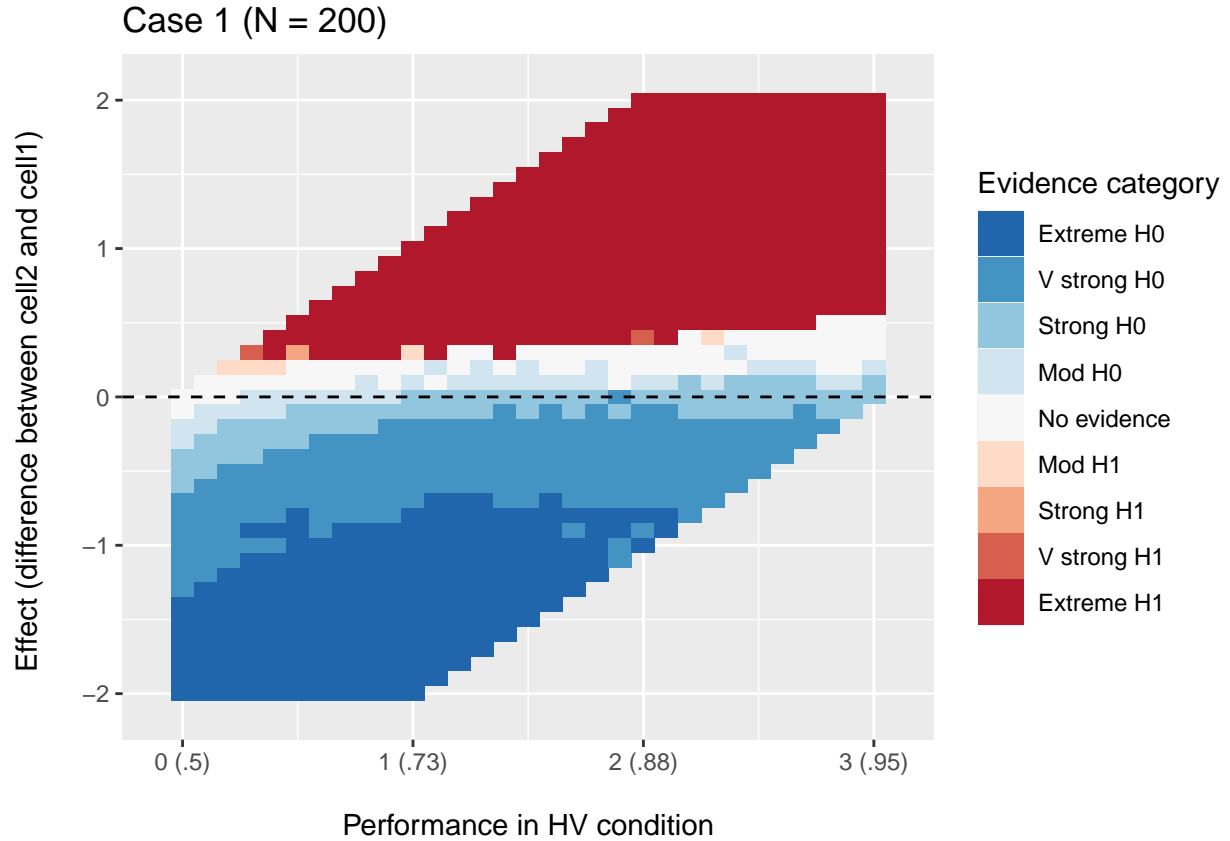


Figure 4: Results from a simulation calculating Bayes factors for the effect of a binary within-subjects predictor in a sample of 200 participants, using the intercept from a mixed-effects model as the estimated effect. The x -axis shows performance in the cell predicted to be highest (e.g., performance in HV condition), and the y -axis shows true effect size (e.g., performance in HV condition minus performance in LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

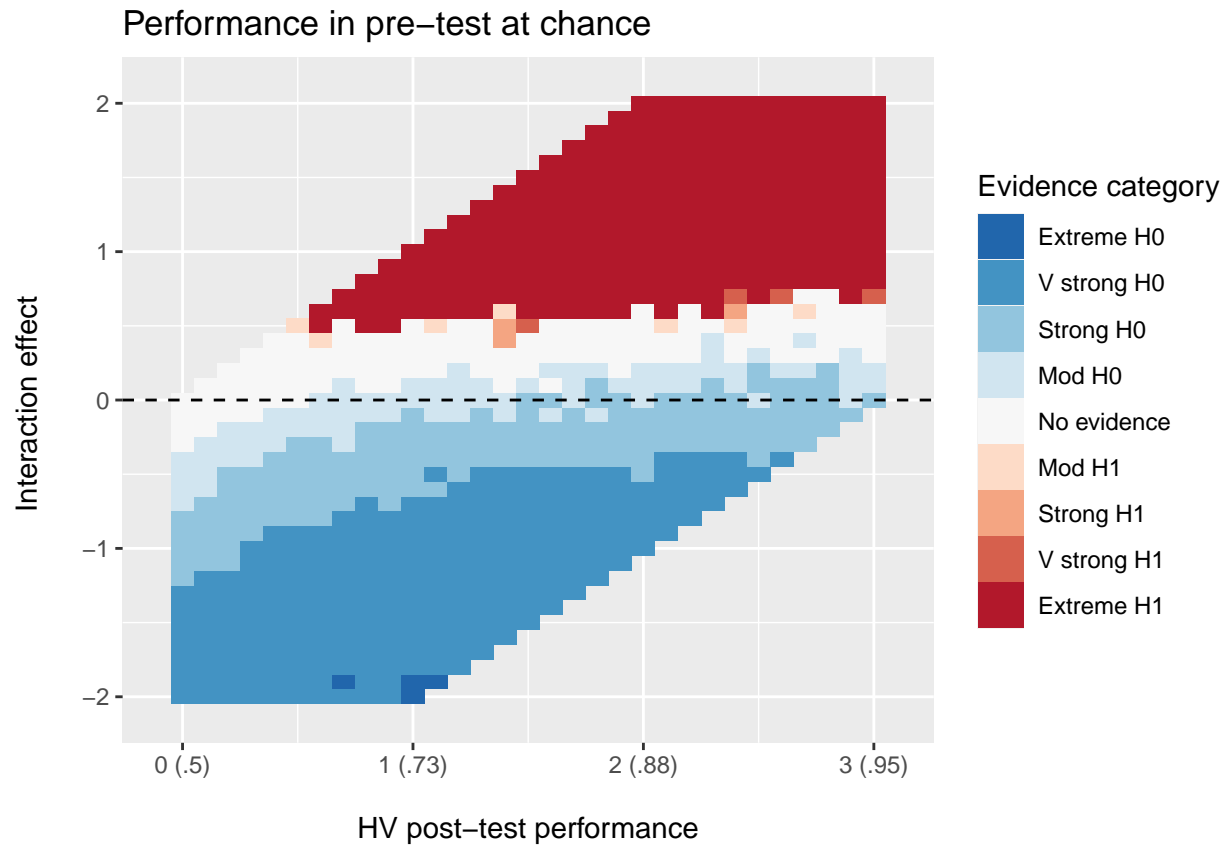


Figure 5: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 200 participants, using 2* the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .5 proportion correct (0 log-odds). The x -axis shows performance in the post-test cell predicted to be highest (e.g., performance in post-test in the HV condition), and the y -axis shows true interaction effect size (e.g., improvement from pre-test to post-test in the HV condition minus the equivalent in the LV condition). Colours show modal Bayes factor category across 20 generated datasets and analyses.

4 Other ways of generating estimates

As mentioned in the paper, exactly how a researcher generates estimates using the motivated-maximum approach should be informed by a study's specific theoretical background and experimental design. Here, we outline the logic of the situation mentioned in the paper, where the between-subjects predictor is training variability (LV vs. HV) and the within-subjects predictor is item novelty (seen vs. unseen). We are interested in the interaction: do participants in the HV condition perform better on unseen items relative to seen items (i.e., do they show better generalization) than participants in the LV condition? To constrain the maximum interaction effect d , we assume the following: a) participants in the LV condition perform at chance on unseen items; b) participants in both conditions perform equivalently on seen items; c) performance on seen items is equivalent to the grand mean; and d) performance in all cells of the design is at least at chance. Let l denote baseline or chance performance, and h denote participants' performance on unseen items in the HV condition (all values in log-odds). If both predictors are centered, the intercept i represents the grand mean performance across cells in the design:

$$i = (i + i + 0 + h)/4$$

$$i = (2i + h)/4$$

$$4i = 2i + h$$

$$h = 2i$$

Performance on unseen items in the HV condition is therefore twice the grand mean. The interaction (d) is the difference between unseen and seen items in the HV condition, minus the difference between unseen and seen items in the LV condition:

$$d = (h - i) - (l - i)$$

$$d = h - i - l + i$$

$$d = h - l$$

Again, in the case where test trials are 2AFC, chance corresponds to .5 proportion correct and hence $l = 0$. In this case, the equation simplifies to:

$$d = l$$

Substituting d for h in the previous equation:

$$d = 2i$$

The maximum interaction is twice the intercept. Using the same heuristic as before, we set the expected effect s to be half this value:

$$s = i$$

We set the standard deviation of the half-normal distribution that is our model of H_1 to equal the intercept from our mixed-effects model.

This is simply an example; many different methods of generating estimates are possible within the motivated-maximum approach. The most important steps are for the researcher to 1) justify the logic and assumptions behind their estimate, and 2) report robustness regions to demonstrate how sensitive the conclusions are to different assumptions.

5 Full details of the simulations

The code and output of the simulations is available on GitHub at <https://github.com/silveycat/bayes-factor>. Below is a description of the parameters used in each simulation.

5.1 Case 1

The Case 1 simulation first generates a number of datasets from a simulated experiment with one within-subjects predictor (analogous to condition, HV vs. LV) and one binary outcome (analogous to correct/incorrect on a series of 2AFC trials). The simulation generates each dataset according to the following parameters:

n_subj: number of participants, set to either 40 (small sample) or 200 (large sample)

n_obs: number of observations per participant, set to 20

subj_tau: standard deviations of the within-subject random effects. SD of the participant random intercepts is set to 0.4; SD of the participant random slopes by test-session is set to 0.9. These values were representative of datasets from similar studies run in the Language Learning Lab.

subj_corr: correlation between participant random intercepts and slopes. This was set to 0.2, again since this was representative of datasets from similar studies run in the Language Learning Lab.

b: true performance in log-odds in the pre-test, set to range from 0 (= .5 proportion correct) to 3 (= .95 proportion correct), in steps of 0.1

v: true performance in log-odds in the post-test, set to range from 0 (= .5 proportion correct) to 3 (= .95 proportion correct), in steps of 0.1

We ran separate simulations for the small sample of 40 participants and the large sample of 200 participants. Within each simulation, for each combination of **b** and **v**, we generated 20 datasets. For each dataset, we analysed it using a mixed-effects model with a main effect of test-session and a by-participant random intercept and slope. We then calculated a Bayes factor using an updated version of the Bf function by Bence Palfi, based on original code by Baguley & Kaye (2010). Parameters used in calculating the Bayes factor were:

sd: standard error of the estimate for the main effect of test-session from the mixed-effects model

obtained: estimate for the main effect of test-session from the mixed-effects model

likelihood: likelihood function of the data, set to normal

modeloftheory: distribution of H_1 , set to normal

modeoftheory: mode of the H_1 distribution, set to 0

scaleoftheory: scale parameter for the H_1 distribution (here, standard deviation, since the H_1 distribution is normal), set to the intercept from the mixed-effects model

tail: parameter to indicate whether H_1 encodes a directional (1) or non-directional (2) prediction. We set it to 1, meaning a half-normal distribution is used for H_1 .

The resulting Bayes factor was then categorized according to the scheme set out in Table 1 of the paper.

5.2 Case 2

The Case 2 simulation first generates a number of datasets from a simulated experiment with one within-subjects predictor (analogous to test-session, pre-test vs. post-test), one between-subjects predictor (analogous to training condition, low-variability vs. high-variability) and one binary outcome (analogous to correct/incorrect on a series of 2AFC trials). The simulation generates each dataset according to the following parameters:

n_subj: number of participants, set to either 40 (small sample) or 200 (large sample)

n_obs: number of observations per participant, set to 20

subj_tau: standard deviations of the within-subject random effects. SD of the participant random intercepts is set to 0.4; SD of the participant random slopes by test-session is set to 0.9. These values were representative of datasets from similar studies run in the Language Learning Lab.

subj_corr: correlation between participant random intercepts and slopes. This was set to 0.2, again since this was representative of datasets from similar studies run in the Language Learning Lab.

a: true performance in log-odds in cell 1 of the design (pre-test in the LV condition). Values tested for the paper were 0 (= .5 proportion correct) and 1 (= .73 proportion correct).

b: true performance in log-odds in cell 2 of the design (pre-test in the HV condition). Values tested for the paper were 0 (= .5 proportion correct) and 1 (= .73 proportion correct).

c: true performance in log-odds in cell 3 of the design (post-test in the LV condition), set to range from 0 (= .5 proportion correct) to 3 (= .95 proportion correct), in steps of 0.1

d: true performance in log-odds in cell 4 of the design (post-test in the HV condition), set to range from 0 (= .5 proportion correct) to 3 (= .95 proportion correct), in steps of 0.1

We ran separate simulations for the small sample of 40 participants and the large sample of 200 participants, and for each pair of values of **a** and **b**. Within each simulation, for each combination of **c** and **d**, we generated 20 datasets. For each dataset, we analysed it using a mixed-effects model with a main effect of test-session and a by-participant random intercept and slope. We then calculated two Bayes factors using an updated version of the Bf function by Bence Palfi, based on original code by Baguley & Kaye (2010). Parameters used in calculating the Bayes factors were:

sd: standard error of the estimate for the interaction of test-session and condition from the mixed-effects model

obtained: estimate for the interaction of test-session and condition from the mixed-effects model

likelihood: likelihood function of the data, set to normal

modeloftheory: distribution of H_1 , set to normal

modeoftheory: mode of the H_1 distribution, set to 0

scaleoftheory: scale parameter for the H_1 distribution (here, standard deviation, since the H_1 distribution is normal). This was set to 1) twice the intercept from the mixed-effects model and 2) the main effect of

test-session from the mixed-effects model. In cases where the main effect of test-session was negative, we did not calculate a Bayes factor using this estimate.

tail: parameter to indicate whether H_1 encodes a directional (1) or non-directional (2) prediction. We set this to 1, meaning a half-normal distribution is used for H_1 .

The resulting Bayes factors were then categorized according to the scheme set out in Table 1 of the paper.

5.3 Model comparison simulation

To compare the Bayes factors from our approach to the Bayes factors produced by **brms** model comparison with bridge sampling, we focused on a limited number of situations drawn from Case 1. Specifically, we varied the following parameters:

b: true performance in log-odds in the pre-test, set to values of 0 (= .5 proportion correct), 1 (= .62 proportion correct), 2 (= .88 proportion correct), and 3 (= .95 proportion correct)

v: true performance in log-odds in the post-test, set to values of 0 (= .5 proportion correct), 1 (= .62 proportion correct), 2 (= .88 proportion correct), and 3 (= .95 proportion correct)

For each combination of **b** and **v**, we first calculated a Bayes factor using the Dienes method as described in the paper, modelling H_1 as a half-normal distribution with a mode of 0 and an SD corresponding to the intercept from a mixed-effects model of the data. We then calculated a Bayes factor using the **brms** model comparison method. To do this, we first defined priors for the **brms** models. Since our hypothesis testing only applied to the prior for the parameter of interest, we used default priors - which are vague and uninformative and thus appropriate for estimation - for all parameters except the main effect. (Specifically, Intercept: a student-t prior with 3 degrees of freedom, a location of 0, and a scale parameter of 2.5; SDs of the random effects: a half student-t prior with 3 degrees of freedom, a location of 0, and scale parameter of 2.5; Correlation matrix of correlations between random effects: an LKJ prior with $\eta = 1$). The prior for the main effect matched the H_1 distribution used for the Dienes method, i.e. it was a half-normal distribution with a mode of 0 and an SD corresponding to the intercept from a mixed-effects model of the data. We then used **brms** to run 1) a full model which corresponded exactly to the mixed-effects model used to generate the estimate - i.e., including a main effect of test-session, and a by-participant random intercept and slope - and 2) a null model which did not include the main effect of test-session, but was otherwise identical to the full model (i.e. a balanced null approach (Aust et al., 2021; Linde & Ravenzwaaij, 2021) since the variance of random by participant slopes for test-session is *not* removed). We then ran the **brms** `bayes_factor` function

to perform bridge sampling on the two models and generate a Bayes factor by comparing their marginal likelihoods. Finally, we categorised each Bayes factor according to the scheme set out in Table 1 of the paper.

Supplementary References

- Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *Computational Brain & Behavior*. <https://doi.org/10.1007/s42113-021-00113-2>
- Baguley, T., & Kaye, W. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by z. dienes. *British Journal of Mathematical and Statistical Psychology*, 63(3), 695–698.
- Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science*, 1–18. <https://doi.org/10.31234/OSF.IO/YQAJ4>
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116(2), 439.
- Linde, M., & Ravenzwaaij, D. van. (2021). Bayes factor model comparisons across parameter values for mixed models. *Computational Brain and Behavior*. <https://doi.org/10.1007/s42113-021-00117-y>
- Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/and/l: A first report. *The Journal of the Acoustical Society of America*, 89(2), 874–886.
- Morey, R. (2010). *All about that "bias, bias, bias" (it's no trouble)*. <http://bayesfactor.blogspot.com/2015/04/all-about-that-bias-bias-bias-its-no.html>