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

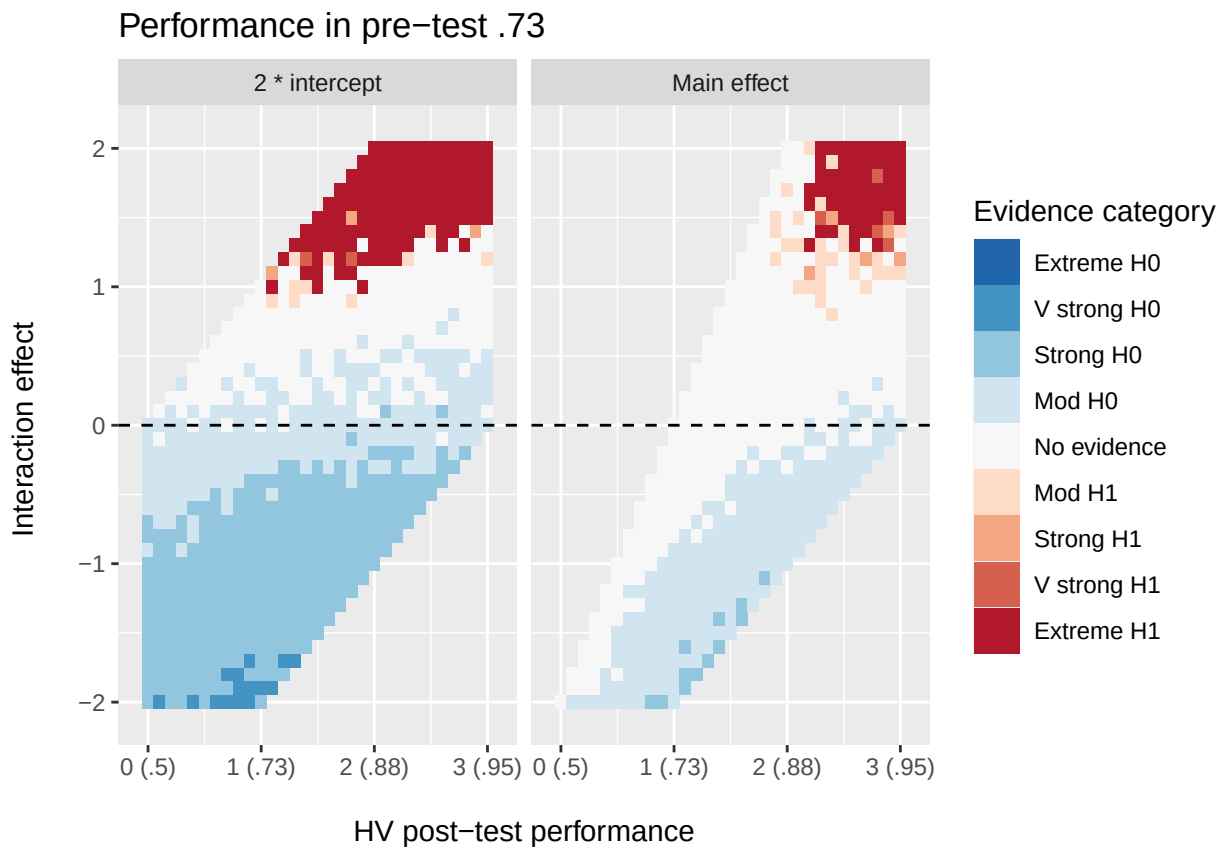

Figure 1: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using either 2* the intercept or the main effect of the within-subjects predictor from a mixed-effects model as the estimated effect. Performance in the two cells 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.

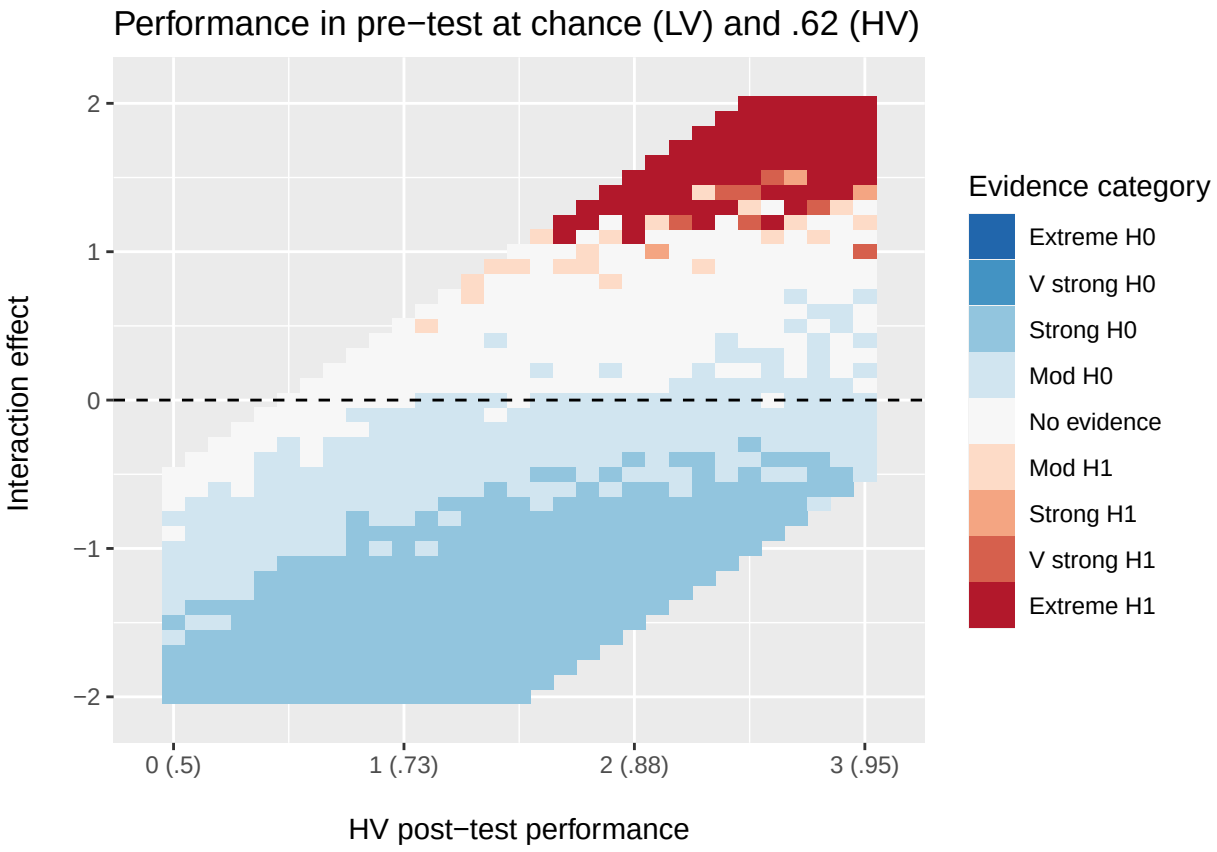

Figure 2: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using 2* the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .5 proportion correct (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.

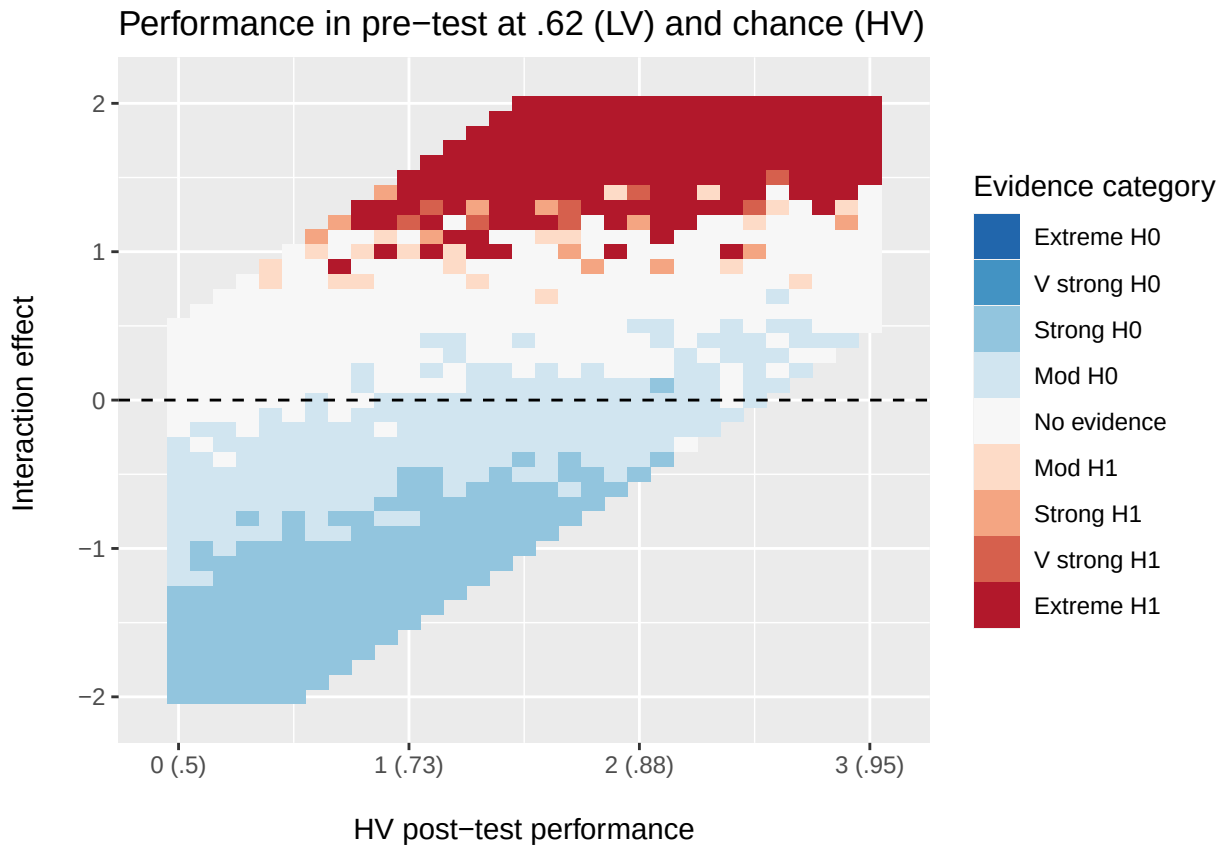

Figure 3: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 40 participants, using 2* the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .62 proportion correct (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.

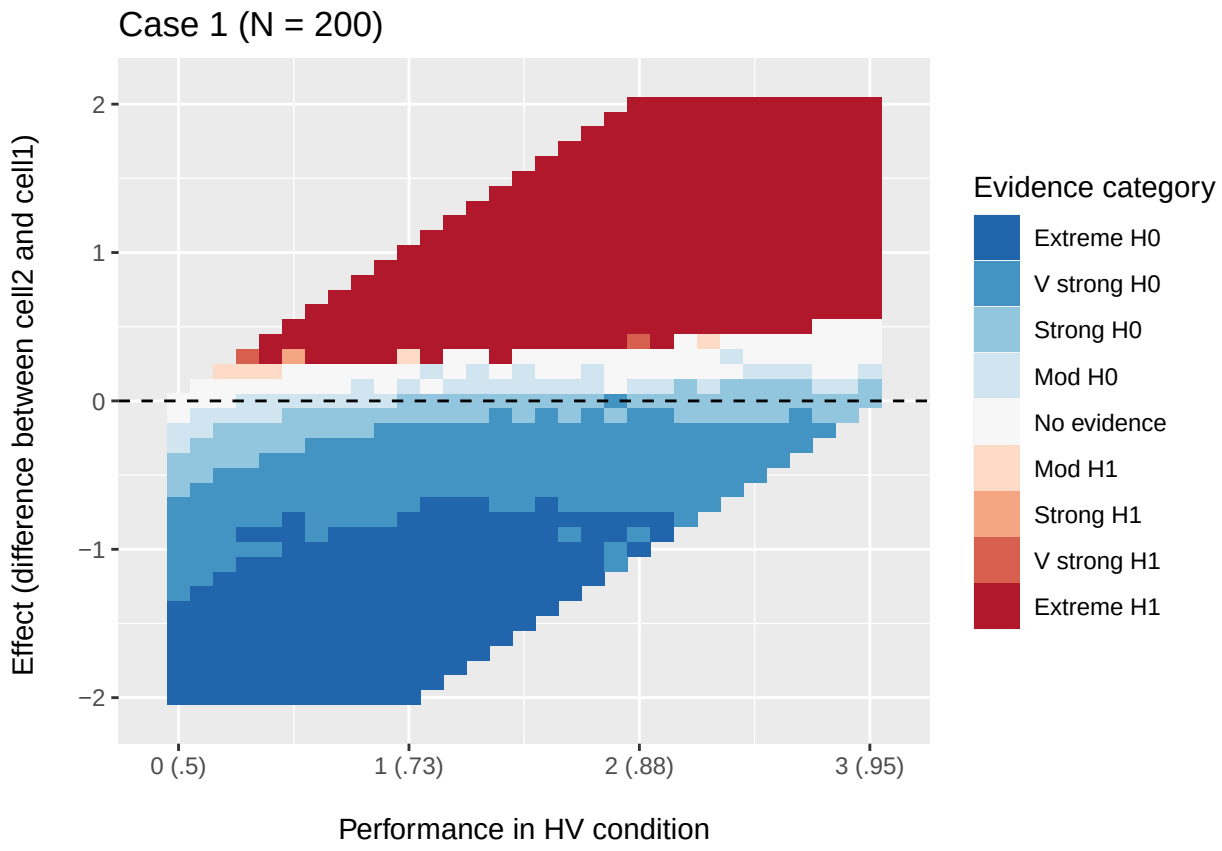

Figure 4: Results from a simulation calculating Bayes factors for the effect of a binary within-subjects predictor in a sample of 200 participants, using the intercept from a mixed-effects model as the estimated effect. The x-axis shows performance in the cell predicted to be highest (e.g., performance 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.

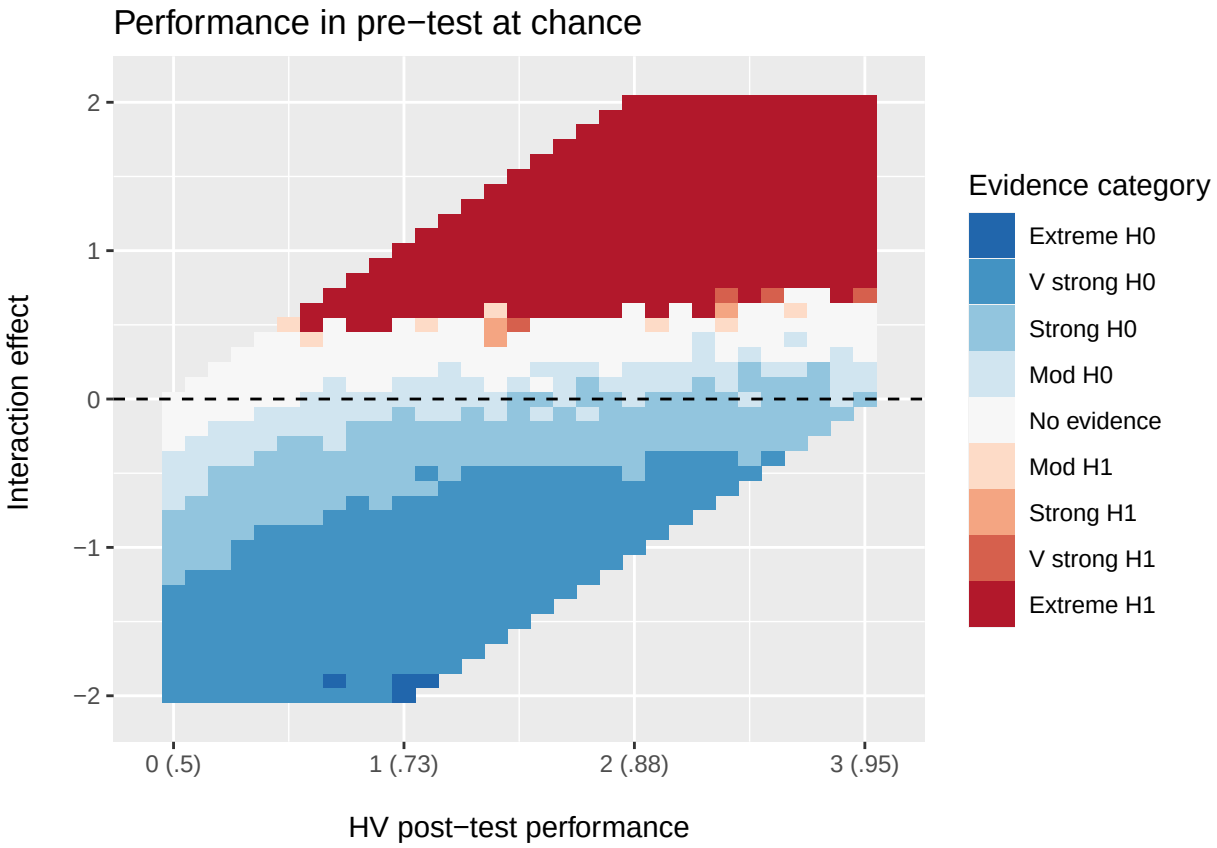

Figure 5: Results from a simulation calculating Bayes factors for the interaction of two binary predictors in a sample of 200 participants, using 2* the intercept from a mixed-effects model as the estimated effect. Performance in the two cells corresponding to the pre-test is fixed at .5 proportion correct (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:

$$
\begin{aligned}
i &= (i + i + 0 + h)/4 \\
i &= (2i + h)/4 \\
4i &= 2i + h \\
h &= 2i
\end{aligned}
$$

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:

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

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