Standard Operating Procedures For Using Mixed-Effects Models

A Principled Workflow from the Decision, Development, and Psychopathology (D2P2) Lab
document version 2.0 – June 24, 2024

[This document will be continuously updated and expanded; it may contain typos and other errors--both unintentional errors and errors based on incorrect or outdated knowledge--we will try to improve these things in future versions. Feel free to let us know if you spotted such things, how to further improve this document!]

Authors (in alphabetical order except that the youngsters were so kind to put the oldest guy in the lab first; BF)

The people involved in creating the new version 2: **Bernd Figner, Floor Burghoorn, Zhang Chen, Jesse Fenneman, Mingqian Guo, Leslie Held, Felix Klaassen, Joppe Klein Breteler, Meghana Vadakkedath Dharmapalan, Fritz Wienicke**

The original (version 1) was created by: **Bernd Figner, Johannes Algermissen, Floor Burghoorn, Leslie Held, Afreen Khalid, Felix Klaassen, Farnaz Mosannenzadeh, Julian Quandt**

When citing our SOP, the following authors should be cited:

Figner, B., Algermissen, J., Burghoorn, F., Chen, Z., Fenneman, J., Guo, M., Held, Khalid, A., L., Klaassen, F., Klein Breteler, J., Mosannenzadeh, F., Quandt, J., Vadakkedath Dharmapalan, M., Wienicke, F.

Content/Analysis Steps

Benefits/advantages of mixed models:
Mixed models offer several advantages (compared to, for example, linear regression or ANOVA-type approaches), notably accounting for the nested structure of data, for example in cases such as repeated measures within individuals, participants grouped in classes, repeated measures within stimuli, or combinations thereof. This is crucial as neglecting such nesting can result in correlated errors, violating standard regression model assumptions and therefore likely to untrustworthy results, for example due to inflated Type 1 errors. Furthermore, mixed models handle missing data more gracefully, not automatically excluding whole sets of data points (e.g., participants) from analyses as long as the missing-at-random assumption holds. Additionally, by capturing variability between groups (e.g., subjects), mixed-effects models can enhance statistical power. In summary, they provide a robust framework for analyzing the intricate data structures prevalent in diverse fields like psychology, biology, neuroscience, education, and economics.

How to read this document:
The structure of this SOP parallels the steps that we usually go through when doing research. Thus, the first section covers study planning, including statistical power analyses. Next, we consecutively cover data preparation, running a mixed-effects model, model diagnostics (which we inspect before inspecting our model results), inferring significance, post-hoc/follow-up tests, and reporting the results. Finally, the Appendix includes more detailed information or *R* code for some of the steps described in the main document.

The SOP provides a description of the approaches that we, in our lab (https://decision-lab.org), usually take when using mixed-effects models for our research. We believe these approaches to be well-suited for our situations. Nevertheless, this document is not meant to be an exhaustive overview of *all* possible approaches. Thus, other approaches may also be feasible and plausible. In some of these cases, we refer to other resources for further reading. Moreover, sometimes, we do not all use the same approach within the lab. In these cases, we try to describe the various possible approaches that we use in the lab.

Please keep in mind that first and foremost, we created this document for ourselves in our own lab, for the type of studies and analyses that we typically do. However, we were happy to hear from other people outside of our lab that they found our SOP helpful. That means, however, that of course your views, thoughts, best-practices, etc might be different from ours, and this is of course completely fine! In case you want to get in touch with us to give us input or feedback or found inaccuracies or mistakes, please feel free to do so.

1. Before data collection: Power/ design/ planning/ sample size

Before starting data collection, whenever possible, we make use of power analysis, sensitivity analysis, and/or employ sequential sampling or other stopping rules to achieve an adequate sample size. A set of different approaches and tools for sample size calculation and planning the design is introduced below. Whatever approach from the list below is chosen, we strongly recommend taking this part seriously and pre-registering the sampling plan. In our lab, we do not all agree on one specific method. That's why we don't have a single recommendation for one method (and sometimes sample sizes are simply determined by logistical constraints such as time or money; but even in these cases, power considerations are often helpful, for example to help decide whether or not to conduct that study at all, given the logistical constraints).

1.1. Power analysis

There are several approaches and tools for power analysis in mixed-effects models (some tools are similar to software like G*Power). Here, we group them into three general approaches: Simulation-based power analyses from scratch, simulation-based power analyses based on estimates from previously collected data, and power analysis tools / apps. We explain each of these approaches below (see section 1.1.3). First, however, we have two notes in advance (if you are looking for information on standardized effect sizes, please go to section 7.3).

1.1.1. What are you powering for?

Power can be defined as the probability of achieving the goal of a planned study, if a suspected underlying state of the world is true. Traditionally, in NHST research, the goal of a study is to reject the null hypothesis. Consequently, the traditional definition of power is the probability of rejecting the null hypothesis when the null hypothesis is false. However, alternative study goals exist, such as the goal to accept the null hypothesis (e.g., Kruschke, 2018; Lakens, 2017), or the goal to achieve a desired level of precision in the estimation of an effect (Kruschke, 2015, Chapter 13). We encourage researchers to think carefully about their study goal, and to power accordingly. Although the targets of statistical power differ per goal, the basic principles of power analyses are the same, and the approaches outlined below therefore generally apply to all of these. In section 1.3, we briefly elaborate on the possibilities of using a sequential sampling approach when powering for precision. In section 1.4, we briefly elaborate on powering for a null effect.

1.1.2. Obtaining informed estimates as input for power analyses

What makes informative power analyses for mixed-effects models more complicated compared to power analyses for simpler statistical models is that they require input for many more parameters. That is, one does not have just one fixed effect (as would be the case in, e.g., a *t*-test), but one has multiple fixed *and* random effects. Even if one is only powering for one effect of interest, using well-informed estimates for all other effects is typically also important to be able to conduct an informative power analysis. Given that coming up with

only one well-informed fixed effect size or precision estimate can already be challenging, coming up with many different fixed and random effect estimates is even more complicated. Here are a few tips that might be helpful to obtain informed estimates:

- Use estimates from previous research, for instance studies that used a similar design, or a meta-analysis. A limitation of this approach is that even if papers and meta-analyses provide detailed information on their fixed effects, they usually do not provide much, if any, information regarding random effects. Therefore, if possible, try to obtain the raw data from previous studies, allowing you to run your mixed-effects model on these data to get an idea about random effects estimates.
- Collect pilot data, run your mixed-effects models on the pilot data, and use the parameter estimates from the pilot data as basis for a power analysis. If you have the resources to collect pilot data, we highly recommend doing so, because, in addition to other benefits of running pilot studies, it provides you with estimates for fixed and random effects. Note that the uncertainty around your estimates increases as the sample size of the pilot study decreases, and a small pilot study will have lots and lots of uncertainty around these estimates. Nevertheless, the estimates may give you an idea at least of the ballpark that one is in regarding the effects (which may be especially helpful for random effects). Try to incorporate this uncertainty into your power analysis (see, e.g., the section on power analyses using **brms**) or rerun your power analysis for different effect sizes.
- Determine the *smallest effect size of interest* (SESOI). This can be based on theoretical considerations and/or practical considerations. In our experience, determining a SESOI is easier when thinking about the raw (unstandardized) effect instead of a standardized effect. We try not to use so-called “T-shirt effect sizes”, i.e., universally accepted “small” (e.g., Cohen’s *d* of 0.2), “medium” (Cohen’s *d* of 0.5), and “large” (Cohen’s *d* of 0.8) effects. These effect sizes are often not informed, and e.g., a “medium” effect size is often unrealistically large for most psychology studies. More realistic effect sizes in psychology often seem to be in ballpark of *d* = .15 to .4 (see, e.g., Brysbaert, 2019 for references of replication studies finding quite different average effect sizes; effect sizes can vary of course also substantially depending on the research field and the study design). If, as a last resort, we do end up using a T-shirt effect size as input for a power analysis (because we do not have a more informed effect-size estimate), we tend to be conservative and use a “small” effect size.

1.1.3. Three general approaches to power analyses

Below, we describe general approaches to power analyses that we use in the lab. All of these approaches are used in our lab; which approach we use depends on the complexity of the mixed-effects model, time constraints, and the experience the specific researcher has with simulating data.

The first approach is creating your own **simulation-based power analysis from scratch**. The basic idea behind a simulation-based power analysis is that one simulates many datasets, runs the mixed-effects model on all these datasets, and checks how often the effect of interest is statistically significant. If, for instance, it is significant 90% of the time, one has a power of 90%. Doing a simulation-based power analysis from scratch is flexible and recommended for situations where more control over the parameter space (e.g., about the

sampling errors) is wanted or when the other tools/apps (explained below) are not sufficient. However, since the simulations are created from scratch, this approach can become very complex as the model complexity increase.
There are several noteworthy resources regarding this approach, including:

- Julian Quandt wrote a comprehensive four-part blog post including *R* code on how to simulate a power analysis from scratch, starting with *t*-tests and ending with linear mixed-effects models.
- A very short step-by-step guide by Ben Bolker, best suited for people already familiar with data simulation.
- Another brief tutorial for custom simulations by Tood Jobe
- A tutorial paper by Tom Snijders
- SimDesign is an *R* package that provides flexible Monte Carlo based simulation framework. Organizing simulations can be a challenge, particularly to those new to the topic, where all too often coders resort to the inefficient and error prone strategies (e.g., the dreaded "for-loop" strategy, forever resulting in confusing, error prone, and simulation-specific code). The package **SimDesign** is one attempt to fix these and other issues that often arise when designing Monte Carlo simulation experiments, while also providing a templated setup that is designed to support many useful features when evaluating simulation research.
- **simstudy**, an *R* package for simulation-based power analysis (or, more generally, for simulating data) which can handle also more complex and clustered data (e.g., patients nested in therapists, in clinics, etc; it has the possibility to introduce different types of missingness, etc); see also helpful example blog post here on how to best do pre/post comparisons with treatment and control group: https://www.rdatagen.net/post/thinking-about-the-run-of-the-mill-pre-post-analysis/ (but note that some authors showed using simulations that using pre-treatment measures as predictors for post-treatment outcomes can lead to inflated Type 1 errors if the pre-treatment measure correlates with another continuous covariate in the model; thus, proceed with caution and read up on the latest insights on this issue!)
- **longpower**, an *R* package for power-simulation of longitudinal data.
- Optimal design, a software to find the optimal research design.
- MLPowSim, an extensively annotated software for power simulation of mixed-effects models (https://sites.google.com/site/optimaldesignsoftware/home)

The second approach is largely similar to the first approach, but it makes use of previously collected data to inform you about the parameters to use in the power simulations. That is, we run our mixed-effects model on previously collected data and use the estimates from this model. The advantage of this approach is that you do not have to come up with the effect estimates yourself, which becomes more helpful once the complexity of the model (and with that, the number of model parameters) increases.

You could still, of course, use the estimates from previous research as input for a simulation-based power analysis that you created from scratch. Alternatively, however, some packages exist that help you do a simulation-based power analysis after you run your model on previously collected data such as **simr,** an *R* package for calculating power for generalized linear mixed models, using simulation.

Brms, a Bayesian *R* package (which can also be used for frequentist inference, such as Null Hypothesis Significance Tests). This package allows you to enter estimates from previously collected data, and/or your own expectations, as prior information, simulate those priors, and predict values for the outcome variable using these priors. Uncertainty about the estimates can be incorporated through, e.g., the standard deviation of the prior distribution. This power analysis approach is explained in more detail in Appendix C of blog post IV by Julian Quandt. It should be noted that power is not really a Bayesian concept. Nevertheless, we have experienced this approach as an accessible form of simulation-based power analysis, reducing the number of things one has to simulate or code manually. Note that the structure of the data set still has to be simulated manually. This is facilitated by the function `generate_design()` that Julian created, also explained in blog post IV.

Another approach which is less flexible, but perhaps easier to use is applying **power analysis tools/apps** such as:

- Power Analysis with crossed random effects by Jake Westfall for a design where, e.g., subjects and items are both random factors.
- Power Analysis with two random factors (crossed or nested) by Jake Westfall for similar purposes as the previous one, but more flexible.
- PANGEA, a comprehensive App by Jake Westfall for mixed ANOVA designs, in which within and/or between subject factors are present and factors can be nested in multiple levels.
- Simulating for LMEM, an app by Lisa DeBruine for power quasi-simulations in which each parameter of the model can be adjusted using a slider (allows for random factors for subjects and items).
    - *R* code by Lisa DeBruine for flexible data simulation
    - Tutorial paper by Lisa DeBruine and Dale Barr for flexible data simulation (including logistic mixed-effects regression)
    - Tutorial paper by Brysbaert (2019; see also Brysbaert and Stevens, 2018): It's not focused on mixed-effects models, but simpler analysis approaches, but still informative and also touching upon mixed-effects models.

If none of the approaches described above is suited and/or feasible, but we do want to obtain an estimate of the rough sample size we need for our study, we may do an extremely rough power analysis for a repeated measures ANOVA, for example using G*Power or something similar.

Lastly, for direct replications, Murayama, Usami, and Sakaki have argued to just use the *t*-test of the respective effect of a previous study and compute power as for a one-sample *t*-test. Another approach for direct replications is the "small telescopes" approach by Simonsohn (2015) which means using 2.5 times the sample size of the original study.

1.2. Sensitivity analysis

In some contexts, it might be useful to use sensitivity analysis rather than power analysis. Sensitivity analysis takes a given sample size (and other relevant information such as number of trials, number of stimuli in the case of random effects for stimuli, etc.) as input and

computes which effect sizes could be detected with a certain power given the sample size (in contrast, conventional power analysis takes expected effect sizes as input and computes the required sample size). Recently, some journals, e.g., the Journal of Experimental Social Psychology, have adopted the approach to ask for sensitivity analyses as a more objective alternative to the rather subjective choice of expected effect size estimates. Furthermore, sensitivity analysis appears to be more realistic in projects with limited budgets and/or time constraints, e.g., student projects. If a sensitivity analysis with the maximum possible sample size (given the resources) yields a minimally statistically detectable effect that seems reasonable, the study can be conducted; otherwise, this might be an indication that the research design and/or research question(s) need to be changed. In cases where the budget or time constraints are less strict, power analysis may be conducted with the tools described above.

1.3. Sequential sampling with stopping rules

Other approaches include flexible sampling plans that allow for sequential sampling until a stopping criterion is met. Kruschke (2015; Chapter 13) showed that this approach is unbiased when the precision of the parameter estimate of interest is taken as stopping criterion (compared to when taking e.g., a critical *p*-value or Bayes Factor as stopping criterion, which may result in serious biases – but see the resources below for more detailed discussions on the use of Bayes Factors as stopping criterion, including work from proponents of this criterion). When sampling for precision, one determines (and pre-registers) a precision criterion a priori, and continues sampling until the precision has been achieved. This precision criterion is often reflected by the width of the confidence or credible interval, and applies to both frequentist and Bayesian analyses. The stopping criterion can also be combined with pragmatic stopping criteria, such as time or budget constraints, which may be particularly relevant for, e.g., student projects. An advantage of sampling for precision is it does not require an a priori estimate of the effect size. However, one does need to determine the precision criterion a priori. The tips in section 1.1.2 can also be applied here, although we encourage researchers to not simply use the precision obtained in previous studies as criterion, but to also consider the *desired* precision criterion. Finally, we wish to note that power analyses for precision can also be done a priori, instead of using a sequential sampling approach.

Relevant resources on sequential sampling are:

- This blog post by Geoff Cumming
- This blog post by John Kruschke about optional stopping in a Bayesian context
- de Heide, R., & 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
- Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226. https://psycnet.apa.org/record/2017-15257-001
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710. https://onlinelibrary.wiley.com/doi/full/10.1002/ejsp.2023

- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-308. https://link.springer.com/article/10.3758/s13423-014-0595-4
- Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods*, *22*(2), 322. https://psycnet.apa.org/record/2015-56330-001
- Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128-142. https://link.springer.com/article/10.3758/s13423-017-1230-y
- For a more critical view, see this blogpost by Richard Morey: https://medium.com/@richarddmorey/power-and-precision-47f644ddea5e

1.4. Powering for a null effect

Sometimes, one might want to show evidence that there is *no* effect, or that two treatments or groups are the same in their scores on some measure. In these cases, one seeks to accept, rather than reject the null hypothesis. However, showing that there is no statistically significant effect, or no statistically significant difference between groups, is not sufficient evidence for the null hypothesis (in other words, absence of evidence is not evidence of absence). Evidence for the null hypothesis can be acquired using equivalence tests, which also need to be sufficiently powered:

- In a frequentist equivalence test (Lakens 2017), one first establishes an equivalence range, consisting of values that one considers equivalent to the null, or that one considers negligible (bounded by the SESOI). See also 1.1.2 for more information about SESOIs. One subsequently performs two one-sided significance tests: one to show that the effect is significantly smaller than the upper bound of the equivalence range, and one to show that the effect is significantly larger than the lower bound of the equivalence range. If both are significant, this can be interpreted as equivalent to a negligible effect or null-effect, given the specified SESOI / equivalence range. If the *p*-value criterion of the two one-sided tests is < .05, this is equivalent to testing whether the 90% CI falls within the equivalence range.
- In a Bayesian ROPE test, the equivalence range is termed a Region of Practical Equivalence (ROPE). One examines how much of the posterior distribution, or the Highest Density Interval (HDI) of the effect of interest falls inside the ROPE, taking this as a measure of evidence for the null effect (see Kruschke 2018, and Makowski et al. 2019, for several criteria). Several of us prefer ROPE tests as these allow us to more quantitatively express the support for/against the null.

Some tools are available for conducting power analyses for equivalence tests, such as the **toster** package in *R* (there also used to be the **BEST** package, but this has been removed from CRAN and might be deprecated). For determining equivalence in mixed-effects models and power of such a design, see also this blog. We think, however, that for mixed-effects models, simulation-based power analyses are a better choice here, as they provide the amount of flexibility required for mixed-effects models.

1.4.1 ROPE test versus Bayes Factors

When doing Bayesian analyses, people sometimes also use Bayes Factors to quantify evidence for a null effect (we have even seen examples where authors first conduct a frequentist significance test and in the cases where it is non-significant, they then report a BF in support of the null; we think this is a very bad idea). Some of us advise against using Bayes Factors approach, unless one uses very well-informed priors and is familiar and OK with the properties of BFs (which at least to some of us are rather unappealing, as nicely described in a series of data colada blog posts: https://datacolada.org/78). Instead, we prefer a ROPE test. Our reason for this preference is twofold. First, the ROPE test uses the posterior distribution to evaluate the evidence for the null value, which is generally robust to the choice of prior distribution (Kruschke, 2013; Wagenmakers et al., 2010). Bayes Factors, in contrast, do not rely on the posterior distribution, but reflect the ratio of the marginal likelihoods of the null and alternative model. These marginal likelihoods are extremely sensitive to the choice of prior distribution, and using a weakly informative prior to convey a state of minimal prior knowledge will result in Bayes Factors, but not posterior distributions, that are biased towards the null model (Kruschke & Liddell, 2018; Rouder et al., 2012; Schad et al., 2022; Tendeiro & Kiers, 2019; Wagenmakers et al., 2010). Thus, if we do not have a strong theory regarding our prior distributions, we favor the approach that is more robust to the choice of prior distribution, and that allows for the use of a weakly informative prior (as recommended by e.g., Kruschke & Liddell, 2018; Liao et al., 2021; Makowski et al., 2019). Second, when visualizing the posterior distribution in relation to the ROPE, one can communicate information about the uncertainty of the parameter estimate (reflected by the width of the posterior distribution) in a more explicit and transparent manner compared to Bayes Factors (Kruschke, 2013). A more detailed discussion of this matter can be found in the Appendix.

1.5. More reading suggestions

More literature on power analysis:

- Brysbaert, M. (2019). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, 2(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6640316/
- Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). http://www.journalofcognition.org/articles/10.5334/joc.10/
- Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*. https://www.sciencedirect.com/science/article/abs/pii/S1364661319302979
- Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7-31. http://journals.sagepub.com/doi/10.1177/0265407517710342
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology* (2022), *8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

For more information about boosting power by increasing the number of trials or number of stimuli (under certain conditions), see:

- Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(3), 295–314. https://doi.org/10.1037/met0000337
    - Associated App: https://shiny.york.ac.uk/powercontours/
- Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. http://doi.wiley.com/10.1111/psyp.13049
- Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19-26. https://journals.sagepub.com/doi/full/10.1177/2515245917745058
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020. https://psycnet.apa.org/doi/10.1037/xge0000014

For an extensive paper on Bayesian design planning:

- Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765*. http://arxiv.org/abs/1904.12765

2. Preparing data

2.1. Categorical variables

We most commonly use sum-to-zero coding for categorical predictors (via the `options(contrasts = c("contr.sum", "contr.poly"))`) for factors. We use this coding scheme because we are typically interested in *main effects* and *main interactions* rather than *simple effects* or *simple interactions* (see also this blog post by Dale Barr; the link seems to be dead at the time of finishing this SOP version? But it's still available via this URL here: https://web.archive.org/web/20230103224304/https://talklab.psy.gla.ac.uk/tvw/catpred/). One option is also to use the command `mixed()` from the package **afex**, as it will automatically set all contrasts to sum-to-zero. Some of us prefer a -0.5/+0.5 coding scheme instead of the default -1/+1 coding, for ease of interpretation of the regression coefficients.

Other reasons to deviate might include the use of *custom contrasts* to test specific hypotheses (see section 6.4.2); an approach that we think is underused.

We usually will *follow-up* on significant effects involving factors with more than two levels by either restricting analyses to only two levels in the form of follow-up models (i.e., analyzing a subset of the data comprising only two levels of the given factor) or, alternatively, we use some post-hoc procedures, e.g., using the package **emmeans** (for more details on both, see the section on post-hocs and follow-ups below). Note that in **brms**, it is also possible to model monotonic effects of ordinal predictors (Bürkner, P. C., & Charpentier, E. (2020),

Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420-451), which some of us like to use.

2.2. Continuous variables

As a default, we typically use mean-centering or z-standardization *or divide mean by 2 standard deviations* for the continuous predictors (to help with model estimation). Centering has several advantages, for example, it can help us interpret the intercept of the model. For instance, when using the previous trial's reaction time (RT) to predict the current trial's RT, an unstandardized previous trial RT would mean that the intercept represents the current trial's RT when the previous one is zero. However, if the predictor is standardized, the intercept is the current trial's RT at the mean of the previous trial's RT. Also, centering (of which standardizing is one variant) is essential to make interactions interpretable and avoid so-called nonessential multicollinearity (Astivia & Kroc, 2021; Dunlap and Kemery, 1987; Marquardt, 1980; also see this blog post by Philipp Masur). Note that different types of centering exist, such as grand-mean centering and group-mean centering, which will change the interpretation of the coefficients (see Enders, & Tofighi, 2007).

3. Running the model

3.1. Model specification and random effects

As a general guideline, we strive to follow the approach of fitting maximal models (in the sense of Barr et al., 2013), i.e., including all random intercepts, slopes, and correlations justified by the experimental design/the data structure.

We are aware that there is a possible trade-off between Type 1 errors (fitting maximal models should avoid inflated Type 1 errors; Barr et al., 2013) and Type 2 errors (maximal models can reduce power if they are too complex given the data, see Matuschek et al., 2017). In our studies, we are often more concerned about not inflating Type 1 error than about inflating Type 2 error and thus maximal models appear to be an appropriate default strategy. Furthermore, the increase in inflating Type 1 error seems to be much much bigger than the possible loss in power: Based on our own simulations, as well as the work by others like Barr et al. and Matuschek et al., inflating Type 1 errors by 1000 % [sic!] due to omission of random effects is quite likely while (e.g., the *p* value is 10 times too small) the maximum drop in power due to an overly complex random effect structures never seems to be larger than about 10% (e.g., the power drops from .8 to .72)

However, if in a specific study, we prefer a different trade-off, we will make this explicit in the pre-registration of that study. A possible scenario might be the following: We test for a certain effect and it is not significant. However, the model is potentially too complex. In this context, the burden is to try and show as convincingly as possible that the effect is indeed not significant. Therefore, one could remove the random slope for the fixed effect of interest and see whether one still obtains a non-significant result. If it is still non-significant, we would

be quite convinced that this non-significance is not due to a loss in power caused by an overly complex random-effects structure of the maximal model.

For clustering variables (e.g., subjects, items), a minimum of 5 levels (e.g., 5 subjects) should be available (better more, e.g., > 30); otherwise, we add these clustering factors just as fixed effects, as commonly recommended.

For control/nuisance variables, we try to add random slopes where appropriate. However, if we have convergence issues, we may opt to not add them as random slopes in order to reduce model complexity (but in such a case we will refrain from reporting and interpreting the associated *p*-value as it likely suffers from inflated Type 1 error; see Barr et al., 2013).

3.2. Addressing convergence and singularity warnings

3.2.1. Convergence warnings in R's lme4

In case of convergence warnings, we attempt the approaches listed below, typically in the order in which they are listed (these steps are based mainly on recommendations by Ben Bolker/the **lme4** team and Barr et al., 2013). For each step, we check whether it resolves the convergence issues.
Before going through the listed steps, some of us would always set the optimizer to `bobyqa` as default via `optimizer = c("bobyqa")` (since it has been suggested that it might work better for the kind of data that we typically have in psychology) and/or switch off the calculation of the gradient and Hessian via `control = [g]lmerControl(calc.derivs = FALSE)`; these settings might already resolve the convergence issues.

1. We increase the number of iterations to the maximum.
2. We use the estimates from the previous (non-converged) fit as our new starting values.
3. We compare the estimates of different optimizers (e.g., using `allFit()`); if different optimizers give highly similar estimates (even if they give convergence warnings), the convergence warnings can be considered false positives. If many of the different optimizers give a singularity warning, it might be an indication the convergence warning is the result of a singularity warning.
4. We follow the steps suggested in Ben Bolker's blog post:
    a. Center independent (and dependent) variables instead of scaling (or vice versa); rescale the independent variables.
    b. Robustness check: Check whether certain random correlations are close to +/-1 and/or certain random slope variances are close to 0. If yes, remove those; afterwards check whether the estimates are still the same. Note that removing random slopes might severely inflate Type 1 error rates!
    c. Double check gradient calculations: Check the (parallel) minimum of the absolute and relative gradients. If those gradients are > 0.001, gradient calculation is likely not a problem.

When speaking of singularity warnings, in simple terms this indicates that the model has estimated the variance of a random effect as 0 (in the case of random slopes and random intercepts) or a random correlation as 1/-1. For the more complete and technical term and explanation, please see Bolkers FAQ. In simpler models, the cause of singularity warnings can be checked by simply inspecting the random variances and correlations after running the model. However, in more complex models, inspecting will sometimes not show the singularity warning. Therefore, in principle, it is recommended to use the `rePCA()` function from **lme4** (see Bates et al., 2015).

After identifying a singularity warning and the potential cause, there are different approaches and much discussion on how to deal with singularity warnings (or not do much anything about them at all), also outlined by Bolker. Please also check out the more unbiased list of options provided there.

Here, we present an approach that prioritizes avoiding underfitting / inflated Type 1 error rates by simplifying the model in steps: Underfitting a linear mixed-effects model can severely inflated Type 1 error rates (e.g., Barr et al., 2013), whereas overfitting in extreme cases can reduce power up to 12% (Matuschek et al., 2017). This approach is what works best for some of us and could be considered:

1. We drop random correlations for one random grouping factor at a time (in the case there are multiple random grouping factors in the model). If the singularity warning is no longer present after dropping one or all random correlations, we use/interpret that model. Else, we continue to the next step.
2. We check the model (without random correlations) using **`lme4::`**`rePCA()` to determine the effect with the smallest variance and remove this from the model. If two variances are equally small, we start with the higher-order variance. Then, we check if the simplified model still gives a singularity warning ***and*** use a likelihood ratio test (LRT) to test whether the model fit becomes significantly worse, using an alpha level of .2 as recommended by Matuscheck et al (2017).
    a. If the simplified model gives no singularity warning and the LRT is not significantly worse, we use/interpret that model.
    b. If the simplified model gives no singularity warning and the LRT is significantly worse, indicating a worse fit for the simplified model, we do not simplify the model and instead use the more complex model and accept the singularity warning. This is also recommended by Singmann, H., Kellen, D., Spieler, D. H., & Schumacher, E. (2019, *New methods in cognitive psychology*).
    c. If the simplified model still gives a singularity warning (and the LRT is not significantly worse), we repeat this step 2 by again identifying the smallest variance using `rePCA()`.
3. Lastly, always compare the results of the simplified model and the maximal model and report both of them (with the maximal model in the supplementary materials for example). If the results diverge, please interpret the results with caution, keeping in mind that the simplified model might suffer from inflated Type 1 errors and/or the full model might suffer from inflated Type 2 errors (e.g., similar as you would interpret results that differ in their interpretation when including versus excluding outliers).

If the above steps don't work, we try further model simplifications:

1. We drop additional random effects in the following order: random correlations, random slopes of covariates (where significance is of no interest), random intercepts ("`0+`" instead "`1+`") (following Barr et al., 2013). Whenever possible, we never remove the random slopes of the variables of interest (i.e., the ones for which we want to conduct significance tests).
   Please note that removing random correlation terms can be tricky if random slopes are estimated for factors with 3 or more levels. In that case, it is probably easiest to use `afex::mixed()` with `expand_re = TRUE` (an alternative option is to create manually the relevant contrasts yourself and add them as predictors to your model, which allows you to suppress the random corrections using the double pipe symbol `||`).
2. We try to run separate analyses: For example, one model to only test the fixed and random effect of A (with fixed effect of B present); then one model to only test the effect of B.
   If we really have to drop random slopes, we follow the next step:
3. We follow the PCA approach suggested by **rePsychLing** (see Bates et al., 2015) that is performing a PCA on the random effects and following the guidelines described in the paper (note that this is similar to what we described above but in principle, here we might simplify to a more extreme extent than above).
   a. We use a likelihood ratio test to test whether the model fit becomes significantly worse. As we prefer a more conservative approach here (i.e., rather err on the side of keeping too many random effects; we prioritize avoiding inflated Type 2 errors for this kind of decision), we use larger alpha-level of .2 (Matuschek et al., 2017).
   b. Alternatively, we suggest an Information criterion approach to avoid using a *p*-value for our inclusion/exclusion decision, but choose the best model based on the Bayesian Information Criterion (*BIC)* or Akaike Information Criterion (*AIC*).

As a last resort, we might use:

- Two-stage regression
  (also called summary statistics approach, e.g., Gelman, 2005):
  Estimate a separate linear/logistic regression per participant, extract the regression coefficients, perform a one-sample *t*-test (or a two-sample *t*-test if testing for group differences) to test whether a certain regression coefficient is significantly different from zero on a group level.
  - This approach constitutes a special case of mixed models with stronger assumptions, i.e., all participants are assumed to provide equally reliable estimates and none of them is an outlier. Also, no shrinkage to the group-level mean is applied in such a case. See, e.g., this comparison of both approaches by Eshin Jolly.
  - This approach is very common in fMRI analyses.
  - As a slightly more sophisticated variant of the same idea, a meta-analysis approach can be used to conduct the test across the per-participant regression coefficients, for example using the **metafor** package. The

advantage is that this approach also carries forward the uncertainty (i.e., standard errors) from the first level (akin to meta analysis).

- Sandwich estimator
  See e.g. the Huber-White sandwich estimator provided by the **merDeriv** package using the **sandwich** package.

3.2.2. Or we choose a Bayesian approach

As an alternative to addressing convergence and/or singularity issues within **lme4**, we instead fit the same model with **brms** and compare its estimates to the **lme4** estimates of the maximal model (i.e., the one with the singularity and/or convergence warning). Often, our main interest is in the fixed effects and in our experience, the conclusions of the **brms** model (e.g., in terms of which fixed effects are "significant" or not) are identical to the conclusions of the **lme4** model, which we interpret that we can trust the **lme4** results. But similarity checks can also be done regarding all the different parameter estimates. Thus, **brms** can be used to check and verify an **lme4** model with convergence and or singularity issues. As a matter of fact, though, many if not most people in our lab have moved to **brms** as their main framework for conducting mixed-effects analysis, so the convergence/singularity warnings of **lme4** have become less of an issue for many of us.

In **brms,** in our experience, there are far fewer issues with estimation (i.e., far fewer warnings that in **lme4**); nevertheless to make sure the **brms** model/results can be trusted, we investigate the convergence of chains by *at least* checking the following (please note that a section on more general model diagnostics follows further below):

1. Trace-plots of the **brms** chains: plot(model)
   a. Did the chains converge (no change of variance across time & chains look like fat caterpillars)?
   b. Are the posterior distributions of the parameters of interest approximately normal and unimodal?
2. Are the Rhats (also sometimes spelled R-hat) reported in `summary(model)` between 0.99 and 1.01? (More recently, there seem to be no values below 1, so the criterion becomes just that Rhat should be smaller than 1.01)
3. Are tail and bulk n-eff ("ESS" in summary output) big enough (bulk n-eff should be bigger than 100 times the number of chains (warnings will be provided if they are very low)?
4. Are no other convergence warnings issued (e.g. exceeding maximum tree-depth, divergent transitions)? If there are, check the Stan Manual convergence guide.

A more extensive tutorial on model-checking by Rens van der Schoot, provides additional information on how these things can be checked and what else might be worth investigating. In case of influential observations, instead of removing them, changing the model family (blog post by Solomon Kurz) provides a more robust alternative (see above).

3.2.3. Mixed Models in Python

While *R* is the predominant tool for conducting mixed-effects models in our lab, thanks to its user-friendly packages **lme4** and **brms**, there are also viable alternatives available in *Python*, which some of us use.

The first option is **statsmodels**, a comprehensive library that encompasses a wide array of commonly used statistical tools beyond mixed-effects models. However, **statsmodels** offers support only for the binomial and Poisson distributions for generalized mixed-effects models. The second option is **Bambi**, which, much like the popular **brms** package in *R*, adopts a Bayesian approach for parameter estimation. **Bambi** stands out by supporting a diverse range of distributions, including Wald, beta, gamma, and more. However, it lacks the capability to estimate the covariance matrix of random effects, which means that correlations between different random effects cannot be assessed. In terms of model-fitting speed, **Bambi** offers slightly faster performance compared to **brms**.

For more information, you can refer to the following links:

- https://www.statsmodels.org/stable/mixed_linear.html
- https://bambinos.github.io/bambi/

3.2.3. Mixed Models in Julia

Because *Julia's* most significant advantage is its high performance, we might consider using `MixedModels` in *Julia* (currently, we do not have much experience with it in our lab). For more information, see also these links:

- https://github.com/RePsychLing/MixedModels-lme4-bridge/blob/master/using_jellyme4.ipynb
- https://github.com/JuliaStats/MixedModels.jl/
- https://github.com/palday/JellyMe4.jl

3.3. Important notes/considerations

3.3.1. Families/ distributions and link functions

When residual scores are non-linear, heteroskedastic or non-normally distributed, a linear model fit is likely to be poor. For example, approximating count data with a ‘standard’ regression model likely yields poor model fit. In such situations, we typically do **not** use (non-linear) transformations of the ‘raw’ dependent variables. Instead, we recommend relying on generalized linear models that better approximate the observed distribution. It's important to note that if our model includes continuous predictors, some authors recommended to apply a log transformation to both the predictors and the dependent variable. This transformation is argued to facilitate the interpretation of the regression coefficients, indicating the percentage change in the predictors and how this relates to the percentage change in the DV (see this blog post). While this sounds relevant and appealing, at the moment, not many of us have adopted this approach (yet?).

3.3.1.1. Deciding on a family

Ideally, the choice of which distribution to use should be based on theoretical ideas, not statistical measures of fit – or even worse, based on the statistical significance of predictors. That is, some outcome variables can be reasonably expected to not follow a normal distribution. For example, dichotomous outcomes are usually modeled with a binomial distribution.

If there are multiple candidate distributions that might be appropriate, but we are not sure which one to use, we normally fit the same model with the different distributions separately and select the one that shows the best fit to the data. When using **brms**, we use the `pp_check()` function to run posterior predictive checks, allowing us to examine the fit between what the model predicts and the observed data. Alternatively, one could use *AIC*, *BIC*, or some other deviance measure (looic, waic or DIC for Bayesian estimation) for model comparison. To confirm that the model with the best fit is healthy, we check the *model diagnostics* (see section 4).

- See also this shiny-app by Jonas Lindeløv for a demonstration of the various distributions in brms that can be used to model reaction times.
- Real-world data do not always adhere to a normal distribution, often exhibiting right or left skewness. In psychological research, it's commonly observed that reaction times (RT) are right-skewed, while normal distributions are symmetric and ill-suited for modeling skewed data. In this context, two options present themselves.
  The first option involves comparing different distributions, such as the Wald or log-normal distributions, as discussed in Jonas Lindeløv's blog, using methods like `pp_check()`, *AIC*, or *BIC* for evaluation.
  The alternative approach is to adhere to the normal distribution model. As highlighted in a blog post by Andrew Gelman, the assumption of normality may not be critically important. If the data exhibit slight skewness, employing a normal distribution model may not significantly impact the validity of statistical inferences drawn from regression coefficients. The primary effect of deviating from normality pertains to the generation or simulation of new data based on the model. Both approaches mentioned have their merits and are not inherently correct or incorrect. However, it's important to note a specific consideration related to psychological theories, such as Donders' Subtractive Method, which often posits a linear relationship between predictors/task manipulations and reaction time. When employing distributions like gamma or Wald, which typically require a non-linear link function, this assumed linear relationship may be compromised. This deviation from linearity is a crucial factor to consider, as it could impact the interpretation and applicability of such models in light of psychological theories that predict linear dynamics.

3.3.1.2. Some commonly used families per DV type along with their respective following link functions

- Continuous
  - Gaussian (default).
    - Link functions: The Gaussian distribution link function is ‘identity’.
    - Examples: amount of money offered/returned, some psychophysiological measures, quasi-continuous rating-scales (i.e.

with many > 10 levels), speeded reaction times (without long tail) Robust alternative: Student.
    - (Shifted) lognormal / ex-gaussian / skewed normal.
        - Link functions: Exponential function or squared function since these distribution require all positive input.
        - Examples: for skewed data such as reaction times, skin conductance responses, quasi-continuous rating-scales
- Categorical / Ordinal/ Counts with defined maximum
    - Bi- or multinomial/ Bernoulli.
        - Link function: inverse logit function and probit function. Inverse logit function has lower chance of having an numerical error.
        - Examples: binary choice (approach/avoid; LL/SS; risky/sure, ambiguous/unambiguous); multinomial choice (healthy, neutral, unhealthy foods).
    - Cumulative. Examples: for ordinal data such as height (low/medium/high), size (small/medium/large), attractiveness (unattractive/neutral/attractive), rating-scales with few levels.
- Counts without maximum
    - Poisson.
        - Link function: Exponential function, squared function, and softplus function.
        - Examples: number of books sold within a week
    - Negative binomial
        - Link function: Exponential function, squared function and softplus function.
        - Examples: Negative binomial distribution is similar to the Poisson distribution. The major difference is that the negative binomial allows a separate standard deviation parameter whereas for a true Poisson distribution, the mean is identical to the standard deviation.

See also the documentation of the `family()` and `brmsfamily()` functions. Based on more anecdotal evidence from our lab, beta-binomial distributions seem to work well for us for data bound between 0 and a maximum (e.g., rating data). Particularly for rating data, see also this helpful post here (on using zero-inflated beta models): https://mvuorre.github.io/posts/2019-02-18-analyze-analog-scale-ratings-with-zero-one-inflated-beta-models/

Ordered beta regression, a more recently developed model, may also be suitable for rating data, and may in some cases be superior to zero-one inflated beta regression because the ordered beta regression has fewer parameters and can be more easily interpreted (Kubinec, 2022).

The distributions mentioned previously apply to cases where data are not censored, but often our data may be censored. For example, in risky decision tasks like the BART (Balloon Analogue Risk Task) and the CCT (Columbia Card Task), a trial may be terminated due to a balloon explosion or turning over a loss card, which are examples of censored data. This is because the subject might have attempted to inflate the balloon more times or turn over

more cards, but had to stop due to the exploding balloon/turned over loss card. Bayesian approaches, such as **brms**, may be more suitable for modeling censored data. Indeed, in our lab, we typically model data with the “hot” CCT as censored using **brms** (see, e.g., Schaefer et al., 2023; note that this paper also links to an OSF repository with a detailed *R* script).
In addition, data might also sometimes be truncated, e.g., follow a normal distribution that is bounded at a lower and/or higher level (e.g., the CCT data are truncated at 0 at the lower level and at 32 at the upper level); **brms** allows to model data as truncated if one wishes to do so (model estimation can become difficult, though, in our--admittedly rather limited—experience).

3.3.2. Estimation method: ML versus REML versus Bayesian

This is how we decide which estimation method to use:

- Bayesian versus (RE)ML:
  We have differing preferences in our lab and thus the individual pre-registrations will describe which approach each project will use. Some pros and cons involve that Bayesian methods are more flexible (e.g., in terms of available families or multivariate models, censoring, truncation, monotonic predictors, etc) but can be more time-consuming. Also, **brms** can usually fit models that are difficult to fit without convergence and/or singularity issues in **lme4**.
  An important feature of **brms** is its capability to estimate complex multivariate models, which offer more than just insights into the correlation between different dependent variables. For instance, in a scenario where a model incorporates both an individual's risky preference and intertemporal choice as DVs, **brms'** multivariate modeling can elucidate the correlation between these two aspects of decision-making. Additionally, it goes further by providing the correlation between the effects of shared predictors of the risky preference model and those of the intertemporal choice model on the respective DVs. This functionality can be used to enhance our understanding of how different factors and preferences may be interrelated, allowing for a more holistic analysis of decision-making processes. We have found that this feature can be particularly useful for studying individual differences.
- ML versus REML:
  The default in **lme4** is REML and we use it unless we have good reasons to use ML instead (e.g., if we intend to use likelihood ratio tests). Since there is a debate about whether ML or REML is more advantageous, in the future, we might change our position.

3.3.3. Priors when using a Bayesian approach

In a Bayesian approach, it is necessary to specify priors. In general (and simplifying things), there are two groups of people with different views on priors. Subjective Bayesian statisticians argue that each study should have its own custom prior which is selected based on personal beliefs or previous studies. This type of prior is often informative which contains a lot of information about parameter values. In contrast, the second group proposes priors should be chosen using objective criteria (e.g., the prior should have minimal impact on the parameter estimation). The latter group has introduced the concept of "default priors," which

are commonly employed as the default priors in software packages like **brms**. This type of prior is uninformative or weakly informative which means it provides zero or little information to the parameter values. There is no definitive superiority between these two perspectives; users should choose the one that best aligns with their research goal. In the following section, we will discuss considerations for using different priors.

Default priors

brms provides default priors that are *weakly regularizing*, which means that they somewhat constrain the possible parameter space to rule out vastly implausible parameter values, but do not comprise much commitment about the specific parameter values that we would expect. Using default priors is generally safe to do and they will not provide you with wrong conclusions in most cases. Using them might be a good idea if there is *no* information about the parameter space.

*Please be aware that you must **not** use the default priors if you want to compute Bayes factors; you need informative priors for that (see section 1.4.1 and the Appendix for a more elaborate discussion on why this is the case, and why some of us generally do not use Bayes Factors)*

Custom priors

If there is anything that one can *a priori* say about the parameter space (which in most cases is possible and easier than it might appear), it is often a good idea to specify custom priors, which can be tested for their implications by performing prior predictive checks. Specifying custom priors is especially useful when a previous study already provided data (such as in direct replication studies), in which case the posterior of the previous study can be used as the prior of the new study. In principle, custom priors are chosen to be conjugate to the likelihood function. The conjugation prior means the prior distribution and posterior distribution is identical for the likelihood distribution. For example, in Gaussian linear regression, the conjugate prior for the mean parameter is Gaussian, while for the standard deviation parameter, it is the inverse gamma distribution. For fixed effects, normally distributed priors are often a good choice, while for random effects, priors with heavier tails (e.g., Cauchy or Student-*t* distributed) might be more appropriate.

In our lab, we have different opinions about the use of default versus custom priors. Therefore, we prefer not to commit generally to one or the other and will specify this in the individual study pre-registrations. As a general rule, if in doubt, we use weakly regularizing priors (e.g., the default priors in **brms**). If we use custom priors, we check whether the different prior specifications lead to different results by comparing them to weakly regularizing (default) priors.

3.3.4 MCMC Iteration number for Bayesian approach

In MCMC sampling, samples exhibit a high degree of correlation with neighboring samples. The Effective Sample Size (ESS) represents the true number of independent samples, adjusting for this correlation. Determining an adequate ESS for accurately and reliably representing the posterior distribution is still debated. John Kruschke in "Doing Bayesian Data Analysis" suggests that an ESS of about 10,000 might be necessary for a precise

estimate of a posterior distribution, potentially requiring upwards of 50,000 raw MCMC samples. However, for estimating central tendencies—such as the mean, median, or mode—of the MCMC samples, such a substantial ESS might not be essential. Thus, thousands of MCMC iterations may suffice for accurate parameter estimation. Yet, for significance testing involving the use of the Highest Density Interval, more than 50,000 MCMC iterations are recommended. If time (and computational ressources) allow, we try to go for relatively large number of iterations, typically.

4. Model Diagnostics

In terms of diagnostics, there are many things one could possibly do. As a rule of thumb, we typically, at minimum, look at the following plots: qq-plots and/or density plots of residuals, and predicted versus observed values to check for things like outliers, violations of linearity, homoskedasticity, and normality (where appropriate). Additional diagnostics aspects often worth checking include multicollinearity (e.g., checking the variance inflation factors) and influence diagnostics (e.g., via **lme4's** influence function). Since in mixed models we have grouped data, it is also a good idea to check, e.g., things like normality not just overall, but also per "group" (e.g., per participant), which can help identify, e.g., whether some participants seem to have unusual data, etc.
Note that we always perform our diagnostics on the *model residuals*, not the raw data.
If there are statistical (numeric) versus visual ways to inspect the data, we usually prefer visualization. For example, commonly used tests like Kolmogorov-Smirnov tests are not appropriate for large enough datasets, and small *p*-values in such tests might be misleading when testing assumptions. For Bayesian models, residuals are perhaps a bit of an unusual concept, but one can compute residuals for brms models and then do the same diagnostics as for an **lme4** model, if one wishes to do so (except for the VIF, a recent search showed that **brms** developer Paul-Christian Buerkner recommended to fit the model in **lme4** to compute the VIF, since it's not implemented in **brms**; this seems a valid suggestion, since the VIF is concerned with the predictors, not the DV or other model aspects).

Typically, we first address convergence and singularity warnings, before we look into the typical model diagnostics. That being said, as recommended by authors such as the lme4 developer team or Dale Barr, it can also make sense to first check whether estimation problems might be caused by, e.g., unusual data points. We recommend to check diagnostics in the following order, since fixing the former ones will often also fix the latter ones (based on a suggestion by Ben Bolker):

1. Outliers and influential cases (in case we remove data due to them being outliers or influential, as a default we report the results both with and without these data points and discuss discrepancies in the conclusions)
2. Non-linearity
3. Homoscedasticity
4. Normality
5. Plot fitted vs. observed

For more details on how these are implemented in code, check the Appendix.
For the very handy package **performance**, containing many automated plots for model diagnostics, see, e.g., this vignette.

Note that if we run generalized linear mixed-effects models, some of the assumptions may not apply. For instance, for a model with a binomial or Bernoulli family, we do not have the normality and homoskedasticity assumption. We do check for influential cases and for the difference between the fitted and observed data. We also check for linearity, but only for the *transformed* relation between the dependent and independent variable (which for the Bernoulli family is on the log-odds scale). Not all diagnostics functions for linear mixed-effects models work for generalized linear mixed-effects models. The **DHARMa** package works for some generalized linear mixed-effects models; see the linked vignette for a list of models and for example code. To the best of our knowledge, however, for some other models, such as ordinal models, we currently have rather limited diagnostics options. It is always possible, however, to compare the fit between fitted and observed data (e.g., using `pp_check()` in **brms**).

5. Inferring significance (*p*-values, *CIs*, Bayesian)

5.1. Frequentist approach (ML/ REML)

When using a frequentist approach, we typically obtain Type-III *p*-values in one of the following ways (see also, e.g., Luke, 2017; but see also Barr et al., 2013 showing that likelihood ratio tests seem trustworthy). In the *pre-registration* of an individual project, we determine beforehand which method we are using. Since methods sometimes fail, it might make sense to pre-register a *decision tree*, e.g., "we plan to use method x to determine *p*-values; if that fails for technical reasons, then we use method y as fallback; etc.". If we had to recommend one specific method, then most of us would recommend KR *F*-tests.

- *F*-test with Kenward-Roger approximation for degrees of freedom:
  Run using either the `Anova()` function of the package **car** (Fox & Weisberg, 2019) or using the `mixed()` function of the package **afex** (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2019) with option `method = "KR"` (if you use `afex::mixed()`, then adding the argument `test_intercept = TRUE` means `car::Anova` is used in the background; otherwise, it will use `lmerTest`; at least some of us have a strong preference for `car::Anova` over `lmerTest` as we have observed odd and obviously incorrect results from `lmerTest` in the past). These functions in turn call the `KRmodcomp()` function of the package **pbkrtest** (Halekoh & Højsgaard, 2014).
- *F*-test with Satterthwaite approximation for degrees of freedom:
  Run using the `mixed()` function of the package **afex** with option `method = "S"`, which in turn calls the package **lmerTest** (Kuznetsova, Brockhoff, & Christensen, 2017).
- (Bootstrapped) Likelihood Ratio Tests:
  Run using the `mixed()` function of the package **afex** with option `method = "LRT"`. If bootstrapped with option `method = "PB"`, this calls the function `PBmodcomp()` of the package **pbkrtest**. Note that LRTs are the only available option (other than t-as-z and Wald chi-square tests; both of which we try to avoid) to directly obtain *p*-values for models fit with `glmer()`.

- 95% confidence intervals:
  CIs can be used by inspecting whether the interval includes 0 or not. These should be based preferably either on bootstrapping or profiling the likelihood (both available via **lme4**). If necessary, *CI*s can then be turned into *p*-values (e.g., if a 95% *CI* does not include zero, this can be used to derive that the *p*-value is < .05)

Note: whenever possible, we do **not** use t-as-z approaches, nor Wald chi-square tests (as implemented, e.g., in the `Anova()` function of the package **car**).

5.2. Bayesian approach

When using a Bayesian approach, we use the function `brm()` from the **brms** package (Bürkner, 2017) which provides an interface to Stan (Carpenter et al., 2017). A Bayesian model does not work with *p*-values to base the statistical significance of predictors on. There are several ways to compute null hypothesis significance testing (NHST) in a Bayesian framework, including the following:

- Computing 95% posterior credible intervals (CIs) either via **brms** default method based on quantiles or HDI (Highest Density Interval; note that this is also sometimes referred to as Highest Posterior Density [HPD] interval). The latter are available, e.g., via packages **sjstats**, **tidybayes**; **HDInterval, bayestestR**; see this vignette by Makowski et al. Please be aware that **emmeans** computes HDI CIs, see below; it probably makes sense to decide on one method to compute CIs *a priori* and then use that same method throughout all the analyses of a study/project). As our decision rule, we check whether the CI includes 0. The quantile-based approach and HDI generally result in highly similar conclusions. The HDI approach is preferable in cases where the posterior distribution of coefficients is highly skewed. In such cases, the quantile-based approach may produce atypical results.
- Computing a Bayesian "*p*-value" or so-called "probability of direction" based on the proportion of posterior samples larger or smaller than 0. This approach is strongly correlated with the frequentist *p*-value. Please think about whether you want to compute a one-sided or two-sided test and accordingly use the appropriate proportion of samples fulfilling that criterion (we typically favor two-sided tests as one-sided tests require ignoring even a very strong effect if it is in the unexpected direction).
- *Looic* (leave-one-out information criterion) version of Bayes factor: Bayes factors are based on the model comparison between the full model and a nested model. As mention further up, Bayes factors require properly specified informative priors. However, the Looic is less sensitive to the prior and can be used as an approximation of a Bayes factor. This *Looic* version of a Bayes factor is often called the pseudo Bayes factor. To compute a pseudo Bayes factor, we need to compute the full model and the nested model *looic* using `loo()` function and then compared two models' looic with `loo_compare()` function. Note that a full model incorporates all potential predictors deemed relevant for the analysis, while a nested model includes a subset of these predictors, implying a more simplified or specific version of the full model.
- Some more hands-on in **brms**: we can use the command `summary('model-name')` to get the 95% credible interval (CI) by default. More specifically: per

predictor, we get a coefficient, its estimated error, and the lower and upper end of the 95% CI range. If the 95% CI does not include 0, we deem an effect "significant" (i.e., we get a probability distribution of true values for a specific parameter; and if the 95% range of that distribution does not include 0, we deem it likely "enough" that the true value does not include 0 and call the effect significant). If we are interested in estimating trend effects or doing one-tailed tests (or computing any other CIs), we can obtain the 90% (or any other) CI by specifying `summary('model-name', prob=.90)`.

- If using the package **emmeans**, for pairwise comparisons/simple effects of the model (e.g., to find out for an interaction which levels significantly differ; Lenth, 2019), we get as output 95% HDIs, which work the same way: if the 95% HDI does not include 0, the pairwise comparison or simple effect is significant.

There are other ways to test significance or find support for a hypothesis (see, e.g., for a discussion of several approaches Makowski et al., 2019). These methods also include Bayes Factors. For different approaches of how to compute Bayes Factors for mixed models, see. e.g., this tutorial by Jonas Lindeløv. However, we are currently not using Bayes Factors as a default method in our lab, as some of us are quite skeptical. For critical discussions, including many code examples, see:

- The above-mentioned tutorial by Jonas Lindeløv
- A series of blog posts by Richard Morey, see especially Part 2
- A series of blog posts by Uri Simonsohn: http://datacolada.org/78a
- Dance of the Bayes Factors by Daniel Lakens
- The absurdity of mapping p-values to Bayes factors by Stephen R. Martin
- An explanation of the default Cauchy prior width of r = .707 used in JASP and the BayesFactor package by Eric-Jan Wagenmakers
- Why psychologists should not change the way they analyze their data: The devil is in the default prior by Ulrich Schimmack
- Wagenmakers' default prior is inconsistent with the observed results in psychological research by Ulrich Schimmack

For more information on indices of effect existence and significance in the Bayesian framework, see Makowski et al. (2019).

6. Post-hocs, follow-ups, simple slopes

Sometimes, to better understand the result patterns, we further investigate main effects or interactions by running additional analyses. In general, we use one of two approaches for additional analyses, post-hoc tests or follow-up models (for some pros and cons of each, see end of this section).

6.1. Post-hoc tests

The post-hoc tests that we use typically depend on the type of our predictors:

- For a significant categorical predictor with > 2 levels, we use the command `emmeans()`

- For a significant interaction between a categorical and continuous predictor, we use the commands `emtrends()` and `contrast(emtrends(), "pairwise", by = NULL)`.
- For a significant interaction between two categorical predictors, we use the commands `contrast(emmeans(), 'pairwise')` and `contrast(contrast(emmeans(), 'pairwise'), 'pairwise', by=NULL)`.

For more details and code specifics, see the Appendix. You can also specify yourself which contrasts you want to test/compare, see, e.g., this vignette on how to use **emmeans**. Also see this vignette on interactions in **emmeans**.

Note that **emmeans** (Lenth, 2019) can be used for **`lme4::glm()`** / **`afex::mixed()`** outputs as well as for Bayesian models (**brms**). It returns estimated marginal means per simple effect and can compute contrasts between them: For Bayesian models, it uses 95% highest posterior density or HPD intervals, while for **lme4**-type models, it provides *p*-values, which can be adjusted for multiple comparisons or not (adjustments for multiple tests are currently not available for **brms** models; if we want to adjust for multiple tests in **brms** models, we implement our own adjustment). Note that for **brms** models, the median instead of the mean is provided as default point estimate. The `hpd.summary()` function can be used to obtain the mean (see the Appendix for example code). For FAQs of **emmeans**, see the respective vignette.

6.2. Follow-up models

Another way to further investigate main effects or interactions is to run separate follow-up models. For example, if we find an interaction between a factor with 2 levels and/or several covariates, one can run 2 models, one per factor level. However, if we have an interaction that includes a factor with more than 2 levels, it would be necessary to run models where the more-than-two levels are restricted to just two levels, which means that multiple models will be run. Whether we adopt such a strategy of follow-up models or rather a post-hoc approach will be determined in the individual study pre-registration.

6.3. General advice

- We only run the follow-up/post-hoc tests that are relevant. We find it often sufficient to interpret the pattern of the interaction based on figures showing the pattern, rather than running many possible additional tests. In our opinion and experience, the main model is typically the most important one for drawing conclusions.
- **emmeans** uses the model estimates for post-hoc tests, not the raw data. Therefore, we always check with raw data or other methods whether the results/conclusions from our post-hocs seem reasonable.
- Correction for multiple comparisons can be done automatically in **emmeans** for ***lme4*** and ***afex*** models. This is *not the case for* ***brms****! Thus, if adjustment for multiple tests is desired for* ***brms*** *post-hoc tests, we do this ourselves.*
- When fitting separate models for different DVs, some kind of correction for multiple comparisons is often warranted. In such a situation it is worth considering

approaches that might mitigate inflated Type 1 errors by means other than adjusting *p*-values: Gelman, Hill, and Yajima (2012) describe a solution where the identity of the DV (e.g., different items or subscales in a questionnaire; when using DVs on different scales, it is appropriate to *standardize* those first) are used as a *grouping variable*. The shrinkage applied to the levels of this grouping variable will automatically adjust for multiple comparisons while retaining higher power. Another option to consider are multivariate mixed-effects models, which are quite easy to run in brms (and very flexible in that they allow the combination of DVs from different distributions, and also allow different predictors for different DVs). It is worth looking at the respective **brms** vignette.

6.4. More considerations

6.4.1. Omnibus vs. targeted tests

Although this does depend on our research question, in general we are interested in specific effects, and thus we strive to run targeted tests (i.e., planned comparisons or contrasts) and not just omnibus tests (also see Baguley, 2012, for a discussion of omnibus tests versus planned comparisons or contrasts). That being said, this might be different for different projects/research questions and thus the individual project's pre-registration will specify the testing strategy.

6.4.2. Contrasts

In general, it is often possible to modify the contrast coding (using custom contrasts) in such a way that the model directly tests the desired comparisons. This could make post-hoc and follow-up tests obsolete. For a nice treatment and tutorial, see Schad, Vasishth, Hohenstein, and Kliegl (2020).
For a tutorial of how to compute contrasts with **brms**, see this blog post by Matti Vuorre. We also have our own materials (lecture slides, example *R* scripts) on how to generate do-it-yourself contrasts (quite similar to the Schad et al., 2020 paper); if you are interested, ask Bernd. In general, we think such custom contrasts are still underused (in our lab, and in general).

7. Reporting results

7.1. In Writing

Our reports include a description of the following parts (also see Meteyard & Davies, 2019; Barr et al., 2013):

- Model specification, including:
  - Dependent variable, and all fixed and random effects (intercepts, slopes, correlations), both in words and possibly also by providing the model equation/ *R*-pseudo code (so-called Wilkinson notation)
  - Transformation of variables, e.g., standardizing or centering variables
  - Contrast coding (typically sum-to-zero coding)

- Inference:
    - Description of how *p*-values were obtained (in case of a frequentist approach) or what other (Bayesian) decision rule was used for inference.
    - Description of what post-hoc or follow-up tests were performed
    - Any convergence or singularity issues that may arise while running the model (in particular if they require adjustments in the model specification) and how they were dealt with should be described, as well as the subsequent adjustments that were made.
- Model output, at minimum the following:
    - Model results: (un)standardized regression coefficients, standard errors and/or confidence/credible intervals, test statistics, degrees of freedom, *p*-values (note that the latter three are only relevant in case of a frequentist analysis).

7.2. Plotting

One question when plotting is how to compute the correct standard errors from the raw data (as there seems to be no generally accepted solution for all cases). One can thus either decide to plot the model-based results, or decide to plot the raw data. Importantly, these two approaches do not always give the same impressions. Moreover, it might not be possible to compute appropriate standard errors/CIs for plotting the raw data.
Here are some options:

- For plotting regression coefficients (several of us find this a most informative plot, because it allows for comparisons across magnitudes and uncertainties of the different observed effects):
    - Use SEs/CIs from the model output.
- For plotting group/condition means:
    - If plotting the raw data (single data points), do not plot any indicator of uncertainty (i.e., no CI or SE indicator), unless there is an appropriate way to calculate it.
        - If aggregating raw data per condition, compute the SEs of the mean like in an ANOVA.
        - When plotting the raw data for within-subjects SEs, mind that between-subjects variability could/should be subtracted first and an appropriate correction for the potential bias performed (Morey, 2008). This is already implemented in the `summarySEwithin()` command from package **Rmisc** (see e.g. this blog post by Niklas Johannes, this blog post by Matt Craddock on visualizing ERPs, and an associated discussion on an MNE Python github issue). Please be aware that this is not a universally accepted approach.
    - Use model-based plots instead of plotting the raw data, e.g.:
        - The **effects** package (Fox, 2003).
        - The `conditional_effects()` function in **brms**.
            - Note here that when the model contains multiple categorical predictors, and one only wants to plot the effect of a single predictor, **brms** does *not* aggregate across levels of the other predictors. Instead, it will plot the effect of the predictor of

interest at the reference level of the other categorical predictors. If you do want to aggregate across the other categorical predictors, you can use the following code: `plot(brms::conditional_effects(modelname, effects = "predictor_of_interest", conditions = data.frame("categoricalpredictor_not_of_interest" = NA)))`

- Also note that by default, **brms** plots the medians instead of the means. To plot the means, add `robust = FALSE)`

- The function `emmip()` in the **emmeans** package can be used to plot interactions. See this vignette for examples. When working with a generalized linear mixed-effects model that uses a link function to transform the dependent variable from the response scale to a different scale (e.g., the log-odds scale), the **emmeans** package can back-transform estimated means and SEs/CIs to the response scale. This might facilitate interpretation of the effects. Note, however, that this back-transformation option is currently restricted to certain link function (e.g., logit), and might thus not work for all link functions.

Note that a distinction can be made between *conditional* and *marginal* effects. More information can be found here from slide 321 (PDF page 346) onwards.

7.3. A note on effect sizes

There are no generally accepted ways to compute standardized effect sizes for mixed effects models, but different variants have been proposed (such as Pseudo-R^2; variants of Cohen's *d*, etc). Individual pre-registrations will specify if they want to report standardized effect sizes, and if so, which (and how they compute them). In addition, standardized effect sizes can be handy in the context of a priori power analyses.
In the context of a recent meta-analysis (Powers, Schaefer, Figner, & Somerville, 2023), some of us took a deep-dive into computing standardized effect sizes in the context of mixed-effects models. We won't reiterate everything here (for details, see page 791 of this Powers et al. paper, under section "Calculation of effect sizes" --> "Mixed-effects models"). For Cohen's *d*-like effect sizes for mixed models, one approach is to standardize the group mean difference by the square root of the sum of all the random effect variances (including the residual variance), see Westfall et al. (2014) and Brysbaert and Stevens (2018); see also Pustejowski (2016). For models with a binary DV, one can transform odds ratios into Cohen's *d* based on the formulas 7.1 and 7.2 in Borenstein et al. (2009). For further details, we could recommend the OSF page associated with the meta-analysis paper, where we have compiled the different formulas, etc: https://osf.io/t6xpb

References

[Bambinos]. (n.d). *Bayesian model-building interface in Python (Bambi)*. Github. https://bambinos.github.io/bambi/

[Chris]. (2012, September 20). *What is the minimum recommended number of groups for a random effects factor?* [Online forum post]. StackExchange. https://stats.stackexchange.com/questions/37647/what-is-the-minimum-recommended-number-of-groups-for-a-random-effects-factor

[EasyStats]. (n.d.). *Performance*. Github. https://easystats.github.io/performance/index.html

[JuliaStats]. (n.d.). *MixedModels.jl*. Github. https://github.com/JuliaStats/MixedModels.jl/

[palday]. (n.d.). *JellyMe4.jl*. Github. https://github.com/palday/JellyMe4.jl

[RePsychLing]. (2020). *MixedModels-lme4-bridge*. Github. https://github.com/RePsychLing/MixedModels-lme4-bridge/blob/master/using_jellyme4.ipynb

Algermissen, J., & Quandt, J. [julianquandt]. (2019*). One to rule them all: A beginner's guide to fitting Bayesian mixed-effects models in Stan using brms @ SIPS 2019*. Github. https://github.com/julianquandt/brms-intro-SIPS2019

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

Astivia, O. L. O., & Kroc, E. (2019). Centering in multiple regression does not always reduce multicollinearity: How to tell when your estimates will not benefit from centering. *Educational and Psychological Measurement*, *79*(5), 813-826. https://doi.org/10.1177/0013164418817801

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44*, 158-175. https://link.springer.com/article/10.3758/s13428-011-0123-7

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(3), 295–314. https://doi.org/10.1037/met0000337

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, 328. https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00328/full

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. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models (Version 1). *ArXiv Preprints arXiv: 1506.04967*. https://arxiv.org/abs/1506.04967

Bolker, B. (2014). *Simulation-based power analysis for mixed models in lme4*. RPubs by RStudio. https://rpubs.com/bbolker/11703

Bolker, B. (2015). *Lme4 convergence warnings: Troubleshooting.* RPubs by RStudio. https://rpubs.com/bbolker/lme4trouble1

Bolker, B. (2016, January 3). *Best way to deal with heteroscedasticity* [Comment on the online forum post *Stats.StackExchange Cross Validated*]. https://stats.stackexchange.com/q/189116

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. http://doi.wiley.com/10.1111/psyp.13049

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. https://doi.org/10.5334/joc.72

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1). http://www.journalofcognition.org/articles/10.5334/joc.10/

Bürkner, P. (2017). Advanced Bayesian multilevel modeling with the R package brms (Version 2*). ArXiv Preprint arXiv: 1705.11123*. https://doi.org/10.48550/arXiv.1705.11123

Bürkner, P. (2024, March 19). *Define custom response distributions with brms*. Cran.R-project. https://cran.r-project.org/web/packages/brms/vignettes/brms_customfamilies.html

Bürkner, P. (2024, March 19). *Estimating multivariate models with brms*. Cran.R-project. https://cran.r-project.org/web/packages/brms/vignettes/brms_multivariate.html

Bürkner, P. [paul-buerkner]. (2016, October 12). *Default priors #131* [Comment on the online forum post *paul-buerkner/brms*]. Github. https://github.com/paul-buerkner/brms/issues/131#issuecomment-253301079

Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology, 73*(3), 420-451. https://doi.org/10.1111/bmsp.12195

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Centre for Multilevel Modelling. (n.d.). *Sample sizes for multilevel models*. University of Bristol. https://www.bristol.ac.uk/cmm/learning/multilevel-models/samples.html#mlpowsim

Chalmers, P. (2024). *SimDesign: Structure for organizing Monte Carlo simulation design* (Version 2.14) [Computer Software]. Cran.R-project. https://cran.r-project.org/web/packages/SimDesign/index.html

Chalmers, R., & Adkins, M. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Comparisons and contrasts in emmeans. (n.d.). Cran.R-project. https://cran.r-project.org/web/packages/emmeans/vignettes/comparisons.html

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small','medium', and 'large' for power analysis. *Trends in Cognitive Sciences, 24*(3), 200-207. https://www.sciencedirect.com/science/article/abs/pii/S1364661319302979

Craddock, M. (2016, November 28). *ERP visualization: Within-subject confidence intervals*. https://www.mattcraddock.com/blog/2016/11/28/erp-visualization-within-subject-confidence-intervals/

Cumming, G. (2018, June 22). Precision for planning: Great new developments. *Introduction to the New Statistics.* https://thenewstatistics.com/itns/2018/06/22/precision-for-planning-great-new-developments/

de Heide, R., & 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

DeBruine, L. M., & Barr, D. J. (2023, July 20). *Understanding mixed effects models through data simulation*. OSFHome. https://osf.io/3cz2e

Donahue, M. C. (2024). *Longpower: Sample size calculations for longitudinal data* (Version 1.0.25) [Computer Software]. Cran.R-project. *Longpower*HYPERLINK "https://cran.r-project.org/web/packages/longpower/index.html" *: Sample size calculations for longitudinal data* (Version 1.0.25) [Computer Software]. Cran.R-project. https://cran.r-project.org/web/packages/longpower/index.html

Dunlap, W. P., & Kemery, E. R. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin, 102*(3), 418–420. https://doi.org/10.1037/0033-2909.102.3.418

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*(2), 121. https://doi.org/10.1037/1082-989X.12.2.121

FAQ for emmeans. (n.d.). Cran.R-project. https://cran.csiro.au/web/packages/emmeans/vignettes/FAQs.html

Farmus, L., Arpin-Cribbie, C. A., & Cribbie, R. A. (2019). Continuous predictors of pretest-posttest change: Highlighting the impact of the regression artifact. *Frontiers in Applied Mathematics and Statistics, 4*, 64. https://doi.org/10.3389/fams.2018.00064

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1–27. http://www.jstatsoft.org/v08/i15/

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage publications. https://www.john-fox.ca/Companion/index.html

Gelman, A. (2005). Two-stage regression and multilevel modeling: A commentary. *Political Analysis, 13*(4), 459–461. https://doi.org/10.1093/pan/mpi032

Gelman, A. (2019, August 21). *You should (usually) log transform your positive data*. Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2019/08/21/you-should-usually-log-transform-your-positive-data/

Gelman, A. (2023, December 12). *Who cares about the normal assump? I don't!* Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2023/12/12/who-cares-about-the-normal-assump-i-dont/

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189-211. https://doi.org/10.1080/19345747.2011.618213

Goldfeld, K. (2018, January 28*). Have you ever asked yourself, "how should I approach the classic pre-post analysis?"*. ouR Data Generation. https://www.rdatagen.net/post/thinking-about-the-run-of-the-mill-pre-post-analysis/

Goldfeld, K.,& Wujciak-Jens, J. (2023). Simstudy: Illuminating research methods through data generation. *Journal of Open Source Software, 5*(54), 2763. https://doi.org/10.21105/joss.02763

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498. https://doi.org/10.1111/2041-210X.12504

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software, 59*, 1-32. https://doi.org/10.18637/jss.v059.i09

Hartig, F., (2022, September 8). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. Cran.R-project. https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html

Interaction analysis in emmeans. (n.d.). Cran.R-project. https://cran.r-project.org/web/packages/emmeans/vignettes/interactions.html

Isager, P. M. (2019, November 25). *Mixed model equivalence test using R and PANGEA*. Pedermisager.org. https://pedermisager.org/blog/mixed_model_equivalence/

Jobe, T. (2009, September 18). *Power analysis for mixed-effect models in R*. Computational Ecology. https://toddjobe.blogspot.com/2009/09/power-analysis-for-mixed-effect-models.html

Johannes, N. (2022, June 24). *Calculating and visualizing error bars for within-subjects designs*. https://www.niklasjohannes.com/post/calculating-and-visualizing-error-bars-for-within-subjects-designs/

Jolly, E. (2019, February 18). *Comparing common analysis strategies for repeated measures data*. https://eshinjolly.com/2019/02/18/rep_measures/

Sassenhagen, J. [jona-sassenhagen]. (2018, December 20). *Within-subject confidence intervals #5812* [Online forum post]. Github. https://github.com/mne-tools/mne-python/issues/5812

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods, 23*(2), 226-243. https://doi.org/10.1037/met0000127

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). https://nyu-cdsc.github.io/learningr/assets/kruschke_bayesian_in_R.pdf

Kruschke, J. K. (2013, November 13). *Optional stopping in data collection: P values, Bayes factors, credible intervals, precision.* Doing Bayesian Data Analysis. http://doingbayesiandataanalysis.blogspot.com/2013/11/optional-stopping-in-data-collection-p.html

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

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270-280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K. (n.d.). *Bayesian estimation supersedes the t test*. https://jkkweb.sitehost.iu.edu/BEST/

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*, 178-206. https://link.springer.com/article/10.3758/s13423-016-1221-4

Kubinec, R. (2022). Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *Political Analysis*, 1–18. https://doi.org/10.1017/pan.2022.20

Kurz, S. (2019, February 10). Bayesian robust correlation with brms and why you should love student's t. https://solomonkurz.netlify.app/blog/2019-02-10-bayesian-robust-correlations-with-brms-and-why-you-should-love-student-s-t/ https://solomonkurz.netlify.app/blog/2019-02-10-bayesian-robust-correlations-with-brms-and-why-you-should-love-student-s-t/

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTestlmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. package: Tests in linear mixed effects models. HYPERLINK "https://doi.org/10.18637/jss.v082.i13" *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2016, July 18). *Dance of the Bayes factors*. Daniel Lakens. https://daniellakens.blogspot.com/2016/07/dance-of-bayes-factors.html?m=1

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7-31. http://journals.sagepub.com/doi/10.1177/0265407517710342

Lenth, R. (2019). *Emmeans package: Estimated marginal means, aka least-squares means* (Version 1.3.5.1) [Computer Software]. Cran.R-project. https://cran.r-project.org/web/packages/emmeans/emmeans.pdf

Liao, J. G., Midya, V., & Berg, A. (2021). Connecting and contrasting the Bayes factor and a modified ROPE procedure for testing interval null hypotheses. *The American Statistician, 75*(3), 256-264. https://doi.org/10.1080/00031305.2019.1701550

Lindeløv, J. K. (2018, February). *How to compute Bayes factors using lm, lmer, BayesFactor, brms, and JAGS/stan/pymc3*. RPubs by RStudio. https://rpubs.com/lindeloev/bayes_factors

Lindeløv, J. K. (2024, May 1). *Reaction time distributions: An interactive overview* [Internet App]. http://lindeloev.net/shiny/rt/

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494-1502. https://link.springer.com/article/10.3758/s13428-016-0809-y

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541

Makowski, D., Ben-Shachar, M. S., Chen, S. H., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02767/full

Marquardt, D. W. (1980). Comment: You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, *75*(369), 87–91. https://doi.org/10.1080/01621459.1980.10477430

Martin, S. R. (2018, February 13). *The absurdity of mapping p-values to Bayes factors*. https://srmart.in/absurdity-mapping-p-values-bayes-factors/

Masur, P. (2018, May 23). *How to center in multilevel models*. https://philippmasur.de/2018/05/23/how-to-center-in-multilevel-models/

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. https://doi.org/10.1016/j.jml.2017.01.001

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61-64. http://www.tqmp.org/RegularArticles/vol04-2/p061/p061.pdf

Morey, R. D. (2018, January 1). *Redefining statistical significance: The statistical arguments*. Medium. https://richarddmorey.medium.com/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91

Morey, R. D. (2020, June 12). *Power and precision.* Medium. https://richarddmorey.medium.com/power-and-precision-47f644ddea5e

Murayama, K., Usami, S., & Sakaki, M. (2022). Summary-statistics-based power analysis: A new and practical method to determine sample size for mixed-effects modeling. *Psychological Methods, 27*(6), 1014–1038. https://doi.org/10.1037/met0000330

Powers, K. E., Schaefer, L., Figner, B., & Somerville, L. H. (2022). Effects of peer observation on risky decision-making in adolescence: A meta-analytic review. *Psychological Bulletin, 148*(11-12), 783–812. https://doi.org/10.1037/bul0000382

Powers, K. E., Schaefer, L., Figner, B., & Somerville, L. H. (2020, December 29). *Formulas used for effect size calculation*. OSFHome. https://osf.io/t6xpb

Pustejovsky, J. E. (2016). Alternative formulas for the standardized mean difference. https://www.jepusto.com/alternative-formulas-for-the-smd/

Quandt, J. (2020, July 1). *Power analysis by data simulation in R – Part II.* Julian Quandt. https://julianquandt.com/post/power-analysis-by-data-simulation-in-r-part-ii/

Quandt, J. (2020, July 1). *Power analysis by data simulation in R – Part III.* Julian Quandt. https://julianquandt.com/post/power-analysis-by-data-simulation-in-r-part-iii/

Quandt, J. (2020, July 1). *Power analysis by data simulation in R – Part IV.* Julian Quandt. https://julianquandt.com/post/power-analysis-by-data-simulation-in-r-part-iv/

Quandt, J. (2022, November 9). *Power analysis by data simulation in R – Part I*. Julian Quandt. https://julianquandt.com/post/power-analysis-by-data-simulation-in-r-part-i/

Raudenbush, S.W., Spybrook, J., Bloom, H., Congdon, R., Hill, C., & Martínez, A. (2011). *Optimal design software for multi-level and longitudinal research* (Version 3.0) [Computer software]. William T. Grant Foundation. https://wtgrantfoundation.org/optimal-design-with-empirical-information-od

Rizopoulos, D. (n.d.). *Statistical analysis of repeated measurements data* [PowerPoint slides]. Department of Biostatistics, Erasmus University Medical Center. https://www.drizopoulos.com/courses/EMC/CE08.pdf

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301-308. https://link.springer.com/article/10.3758/s13423-014-0595-4

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19-26. https://journals.sagepub.com/doi/full/10.1177/2515245917745058

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356-374. https://doi.org/10.1016/j.jmp.2012.08.001

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. *ArXiv Preprint arXiv:1904.12765*. http://arxiv.org/abs/1904.12765

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2023). Workflow techniques for the robust use of Bayes factors. *Psychological Methods, 28*(6), 1404–1426. https://doi.org/10.1037/met0000472

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*, 104038. https://doi.org/10.1016/j.jml.2019.104038

Schaefer, L., Ikink, I., Woyke, I., Heuvelmans, V., Roelofs, K., & Figner, B. (2022, August 18). No evidence for a causal effect of exogenous testosterone on risky decision-making in women: An experiment and meta-analysis. *Decision*, *9*(4). https://decision-lab.org/wp-content/uploads/2023/02/Schaefer-Ikink-Woyke-Heuvelmans-Roelofs-Figner-2022-Decision.pdf

Schimmack, U. (2015, May 9). *Why psychologists should not change the way they analyze their data: The devil is in the default prior*. Replicability-Index. https://replicationindex.com/2015/05/09/why-psychologists-should-not-change-the-way-they-analyze-their-data-the-devil-is-in-the-default-piror/

Schimmack, U. (2016, June 30). *Wagenmakers' default prior is inconsistent with the observed results in psychologial research*. Replicability-Index. https://replicationindex.com/2016/06/30/wagenmakers-default-prior-is-inconsistent-with-the-observed-results-in-psychologial-research/

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128-142. https://link.springer.com/article/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E. J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322. https://psycnet.apa.org/record/2015-56330-001

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569. https://doi.org/10.1177/0956797614567341

Simonsohn, U. (2019, September). *Drop that Bayes: A Colada series on Bayes factor*. Data Colada. https://datacolada.org/78

Simpson, D. (2019, March 19). *Maybe it's time to let the old ways die; or We broke R-hat so now we have to fix it.* Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2019/03/19/maybe-its-time-to-let-the-old-ways-die-or-we-broke-r-hat-so-now-we-have-to-fix-it/

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Schachar, M. S. (2020). *Afex: Analysis of factorial experiments* [Computer Software]. Cran.R-project. https://CRAN.R-project.org/package=afex

Singmann, H., Kellen, D., Spieler, D. H., & Schumacher, E. (2019). *New methods in cognitive psychology* (1st ed.). Routledge.

Smeet, L., & van de Schoot, R. (2019, August 21). *WAMBS R tutorial (using brms)*. Rens van de Schoot. https://www.rensvandeschoot.com/tutorials/wambs-checklist-in-r-using-brms/

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In: B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Volume 3). Wiley. https://www.stats.ox.ac.uk/~snijders/PowerSampleSizeMultilevel.pdf

Stan Development Team. (2022, March 10). *Runtime warnings and convergence problems*. Stan. https://mc-stan.org/misc/warnings.html#bulk-ess

Stan Development Team. (n.d.). *Stan reference manual*: *Convergence* (Old version). Stan. https://mc-stan.org/docs/2_20/reference-manual/convergence.html

Statsmodels. (2023, December 14). *Linear mixed effect models*. https://www.statsmodels.org/stable/mixed_linear.html

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods, 24*(6), 774–795. https://doi.org/10.1037/met0000221

University of York. (n.d.). *Power contour estimation* [Internet app]. https://shiny.york.ac.uk/powercontours/

Viechtbauer, W. (2024). Metafor: Meta-analysis package for R [Computer Software]. Cran.R-project. https://cran.r-project.org/web/packages/metafor/index.html

Vuorre, M. (2019, February 18). *How to analyze visual analog (slider) scale data? A reasonable choice might be the zero-one-inflated beta model*. Matti's Homepage. https://mvuorre.github.io/posts/2019-02-18-analyze-analog-scale-ratings-with-zero-one-inflated-beta-models/

Vuorre, M. (2020, February 6). *How to calculate contrasts from a fitted brms model: Answer more questions with your estimated parameters, without refitting the model*. Matti's Homepage. https://mvuorre.github.io/posts/2020-02-06-how-to-calculate-contrasts-from-a-fitted-brms-model/

Wagenmakers, E. J. [EJ]. (2015, November). Cauchy prior widths [Comment on the online forum post *Cognitive Science and More*]. http://www.cogsci.nl/forum/index.php?p=/discussion/1725/cauchy-prior-widths

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology, 60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wang, T., & Merkle, E. C. (2018). merDeriv: Derivative computations for linear mixed effects models with application to robust standard errors. *Journal of Statistical Software*, *87*, 1-16. https://doi.org/10.18637/jss.v087.c01

Westfall, J. (N.A). *PANGEA (v0.2): Power ANalysis for GEneral Anova designs*. https://jakewestfall.shinyapps.io/pangea/

Westfall, J. (N.A). *Power analysis with crossed random effects*. https://jakewestfall.shinyapps.io/crossedpower/

Westfall, J. (N.A). *Power analysis with random targets and participants*. https://jakewestfall.shinyapps.io/two_factor_power/

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020. https://psycnet.apa.org/doi/10.1037/xge0000014

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Zeileis, A., & Lumley, T. (2023). *Sandwich: Robust covariance matrix estimators* [Computer Software]. Cran.R-project. https://cran.r-project.org/web/packages/sandwich/sandwich.pdf

Appendix

Connections between mixed-effects model and commonly used statistical tests

Commonly used tests like *t*-test and ANOVA can be regarded as special cases of mixed-effects models. Here is a tutorial about the relationship between commonly used tests and linear regression.
Note that in experiment psychology, mixed-effects models are commonly applied to entire datasets (i.e., the unaggregated raw data), while ANOVA is typically used for data aggregated per participant.
We list several of them (some connections are not included in the tutorial),

- *T*-test is equivalent to a linear regression without any random effects.
- Welch's *t*-test is equivalent to a linear regression with varying standard deviation.
- Paired-wise *t*-test is equivalent to a mixed-effects model with random intercept.
- Between-subjects design ANOVA is equivalent to a linear regression without any random effects.
- Within-subject design ANOVA is equivalent to a mixed effect-model with random slope and random intercept.

Diagnostics

Outliers

- We save the standardized residuals

```
sum(abs(resid(model, scaled = TRUE)) > value) /
length(resid(model))
```

- As a check, some of us used the following expectation about what pattern could be expected (based on a normal distribution):
    - o No values larger than +/- 3 (or 3.5)
    - o Max. 1 % larger than +/- 2.5
    - o Max. 5 % larger than +/- 2

However, a recent paper suggested that assuming a parameters based on a normal distribution (i.e., mean and SD) to check whether a distribution is a normal distribution might not be optimal. Consistent with this suggestion, we agree that it often makes more sense to use a criterion such as median +/- 2.5 (or 3) MAD (median absolute deviation) to check this assumption. Note that when more than 50% of the data have identical values, the MAD value will be computed as zero (https://eurekastatistics.com/using-the-median-absolute-deviation-to-find-outliers/). Although this scenario may not arise very often, some of us have encountered this in their data analysis. We recommend checking the distribution of data before using MAD to identify outliers.

Auto-correlation

- Use the function `acf()`: We would expect no significant correlations across any lags (no bars more extreme than the dotted horizontal lines). Beware, though, that—as far as we know—this function relies just on the order of the observations in the data frame and is thus not "aware" of (and thus cannot take into account) the clustering of observations in groups (such as trials clustered within participants). We have not found an existing function or solution to this problem (i.e., one would have to compute the auto-correlation within each, e.g., participant, and then average this correlation across participants).
    - `library(lme4)`
    - `plot(acf(sleepstudy$Reaction)) # pretty dramatic autocorrelation in the raw data`
    - `m1 <- lmer(Reaction ~ Days + (1 + Days | Subject), data = sleepstudy)`
    - `plot(acf(resid(m1))) # no serious autocorrelation in the residuals`

Homoscedasticity

- Plot of fitted values vs. residuals to check for homo/heteroskedasticity (optional: fitted vs. observed values)
    - `plot(model, type = c('p', 'smooth'))`
- Check the ratio between the highest and lowest variance (by visual inspection, called *Fmax*).
- For ungrouped data (i.e., continuous predictors), heteroscedasticity is not fatal: “The linear relationship between variables is captured by the analysis, but there is even more predictability if the heteroscedasticity is accounted for. If it is not, the analysis is weakened, but not invalidated” (Tabachnick & Fidell, 2013, p. 85).
- For group data (i.e., factors), for equal cell sizes (up to a ratio of 1:4), an *Fmax* of up to 10 is acceptable (Tabachnick & Fidell, 2013, p. 86). If cell sizes are very uneven (say 1:9) and variance larger in smaller cells than bigger cells, *Fmax* as small as 3 can be associated with increased Type 1 error rates (Milligan, Wong, & Thompson, 1987)

Normality

- Density plot or qq-plots of residuals to check for normal distribution:
    - `densityplot(resid(model, scaled = TRUE))`
    - `qqmath(model, scaled = TRUE)`
    - `qqPlot(resid(model))`

Influential cases

We like to use the function: **`lme4::`**`influence` (package **dharma** for generalized models) to get influence statistics for formal inspection:

- `inf_model <- influence(model, "grouping factor")`
- `str(inf_model)`

To check for problematic values

- Cook's distance: `cooks.distance(inf_model)`
  - values larger than 1
  - values larger than 4/N (grouping units)
  - Points that stand out
    - `plot(inf_model, which = 'cook', sort=T)`

- Dfbeta: `dfbetas(inf_model)`
  - Values larger than 1
  - Values larger than 2/sqrt(N)
  - Points that stand out
    - `plot(inf_model, which = 'dfbetas')`

Additional quantitative and visual checks

- Check distributions of raw data and residuals per cell (factor levels):
  - `with(dataframe, densityplot(~y | factor))`
  - `with(dataframe, densityplot(~ res_model | factor))`
- Create xy plots for regressors separately over groups:
  - `xyplot(res_model ~ regressor, data = dataframe, type = c('p', 'r', 'smooth'))`
- Screen groups separately:
  - `xyplot(y ~ regressor | grouping factor, data = df, type = c('p', 'r'))`
  - `xyplot(res_model ~ regressor | grouping factor, data = dataframe, type = c('p', 'r'))`

Bayesian

Generally, we examine the same diagnostics as those described above for frequentist models (outliers, auto-correlation, normality, influential cases). Below, we describe a few additional checks that we sometimes do when running Bayesian analyses.

Posterior predictive checks

Some of us use posterior predictive checks to examine whether the predictions made by the fitted model are in line with the observed data. We often use the function `pp_check()` from the **brms** package for this.

Influential cases

- We use the function **`loo::`**`loo` to get influence statistics for formal inspection.
- We start with:
  - `loo_model <- loo(model)`
  - `print(loo_model)`
    `Computed from 16000 by 1758 log-likelihood matrix`

```
         Estimate    SE
elpd_loo  -6917.9 115.4
p_loo       135.6  20.0
looic     13835.8 230.8
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     1745  99.3%   1053
 (0.5, 0.7]   (ok)          8   0.5%   617
   (0.7, 1]   (bad)         4   0.2%   17
   (1, Inf)   (very bad)    1   0.1%   10
See help('pareto-k-diagnostic') for details.
```

- In the above, we see that there are 5 bad and very bad (i.e. influential) observations. If there are only a few of these (less than 10), we can test their influence directly by refitting the model once for each observation using `reloo = TRUE`. This can take a lot of time if the model-fitting takes a long time:
  - `loo_new <- loo(model, reloo = TRUE, reloo_extra_args = list(cores = n_cores, chains = n_chains)`
- Again we will get a table like the one above. If the resulting `Monte Carlo SE of elpd_loo` is small compared to the other SEs in the table, the influence of these observations is not too strong.
- If we have too many influential observations (more than 10), loo will tell you that approximate loo might not work well anymore and k-fold cross validation should be used instead.
- Alternatively, if we want to check robustness of our results without however many influential cases, we can exclude all of them at once the following way (if d is the data that was used during model-fitting)
  - `influential_cases <- pareto_k_ids(loo_model, threshold = .7)`
  - `d_new <- d[-influential_cases, ]`
  - `model_new <- update(model, newdata = d_new)`

  Now we can see whether conclusions stay the same

Post-hoc tests or planned comparisons code

Below is some code that we often use to do post-hoc tests or planned comparisons. More examples and code can be found in the interactions vignette from **emmeans**.

- For significant categorical predictors with >2 levels, we use the command `emmeans(model-name, pairwise ~ factor_with_e.g.3levels)`:
  - Returns estimated marginal means (EMMs) per factor level, the pairwise comparisons between the 3 factor levels (e.g. level 1-2, level 1-3, and level 2-3), returning estimates, and lower/upper end of 95% HPD intervals.
  - When using a **brms** model as input, the median is provided as default point estimate. To obtain the means, the following code can be used: `hpd.summary(emmeans(model-name, pairwise ~`

`factor_with_e.g.3levels)$emmeans, point.est = mean)`. To obtain the contrasts between the means, `$emmeans` should be replaced by `$contrasts`.
    - To get 90% HPD intervals, we use the command `confint(emmeans(model-name, pairwise ~ factor_with_3levels), level = .90)`.
- If using a response transformation, results are on the transformed scale as well. But if responses are on the log or logit scale (e.g., such as when using binary dependent variables), we can 'back-transform' them to the original scale using the command `emmeans(model-name, pairwise ~ factor_with_e.g.3levels, type='response')`. However, note that back-transforming should be delayed until the end (i.e., right before reporting estimates; for more information see this **emmeans** vignette). Also note that, as described above, back-transformations are not (yet) available for all types of model families and link functions.
- For a significant interaction between two categorical predictors, we can use the following commands.
    - `emmeans(model-name, pairwise ~ factor1 | factor2))`. It returns per level of factor 2 the significance of factor 1 (e.g., is the effect of factor 1 significant for each separate level of factor 2)
    - To test whether the interaction between the two factors is significant in the first place (which is only relevant if the direct summary output does not provide one with this information), we use: `contrast(emmeans(model-name, pairwise ~ factor1 | factor2)[[1]], interaction = 'pairwise', by=NULL)`.
    - If one wants to do a planned comparison between specific combination of factor levels, we use `emmeans(model-name, pairwise ~ factor1 * factor2))`.
- For a significant interaction between a categorical and continuous predictor, we use the following commands.
    - `emtrends(model-name, 'factor_with_xlevels', var = 'continuous_predictor')`. It returns simple slopes of the continuous predictor per factor level, and their significance (e.g., is the continuous predictor significant per factor level)
    - `contrast(emtrends(model-name, ~ 'factor_with_xlevels', var = 'continuous_predictor'), "pairwise", by = NULL)`. It returns pairwise comparisons between the factor levels for the continuous predictor effect (e.g., do the slopes differ significantly between the factor levels, comparing slope 1-2, slope 1-3, etc.)
    - If we want to test the effect of the categorical predictor at different levels of the continuous predictor, we first determine the points of the continuous predictor for which we want to test this. This could for instance be the mean, 1*SD* below the mean, and 1 *SD* above the mean. We then create the following list: `mylist <- list(continuous predictor=c(1sdbelow, meanlevel, 1sdabove), categorical_predictor = c("factorlevel1","factorlevel2"))`. Next, we run the following code to get the simple slopes: `contrast(emmeans(model-name, ~`

```
continuous_predictor * categorical_predictor, at =
mylist), "pairwise", by = "continuous _predictor")
```

ROPE test versus Bayes Factors to support a null effect

As described in section 1.4.1, some of us prefer ROPE tests over Bayes Factors as method to quantify evidence for a null effect. Here, we elaborate on our reasons for this preference. A ROPE test implies defining a region of parameter values that one considers to be practically equivalent to the null value (i.e., a ROPE). One subsequently estimates the proportion of the posterior distribution or the Highest Density Interval (HDI) that falls within the ROPE, i.e., that is practically equivalent to the null. This is a form of Bayesian *parameter estimation*, meaning that we derive the posterior distribution to estimate the most credible values of a parameter (Kruschke, 2015). Bayes Factors, in contrast, are a form of Bayesian *model comparison*. In model comparison, the focus is not on estimating the parameter value, but on comparing the evidence the data provide for one model over the other (e.g., the null model over the alternative model). More specifically, one compares the marginal likelihood that the observed data have occurred under one model (e.g., the null model) versus another (e.g., the alternative model). The marginal likelihood for each model is derived by, for all possible parameter values, computing the likelihood that the data were generated by the candidate model, weighting the likelihoods by the prior credibility of these parameter values, and then integrating them. Crucially, because each likelihood is weighted by the prior, the prior is as important as the likelihood, and therefore strongly influences the marginal likelihood, even when the data are strongly informative (Schad et al., 2022). Therefore, Bayes Factors are extremely sensitive to the choice of prior distribution. This contrasts with Bayesian parameter estimation, in which the prior has a regularizing function, but is easily overwhelmed by the data. Therefore, parameter estimation, including the ROPE approach, is generally much more robust against changes in the prior (Kruschke, 2013; Wagenmakers et al., 2010).

Thus, in order to use Bayes Factors in an informative manner, the priors should be selected carefully. For the null model, we may select a spike-shaped prior at an effect of zero. For the alternative model, a broader prior distribution should be selected that spreads over parameter values in a plausible way. Not having a theoretically informed expectation regarding the prior distributions, however, complicates the selection of a prior. One tempting option would be to use a diffuse, uninformative prior, that spreads prior plausibility of parameter values evenly across all or a wide range of values, reflecting minimal prior knowledge. Such noninformative priors exert only a weakly regularizing influence on the posterior distribution, and hence, on parameter estimation — which is why we often use such priors for our statistical models. In contrast, however, when computing Bayes Factors, noninformative priors exert a strong negative influence on the marginal likelihood of the alternative model. The reason for this is that a diffuse prior implies that (almost) all parameter values are equally plausible, including implausible values. The likelihood for implausible values will naturally be low, and because the marginal likelihood is computed by integrating the likelihood across all parameter values, this marginal likelihood will also decrease. Therefore, noninformative, diffuse priors will lead to a Bayes Factors that are biased towards the null model (Rouder et al., 2012; Schad et al., 2022; Tendeiro & Kiers, 2019; Wagenmakers et al., 2010). Several default priors have been developed in an attempt to avoid this issue, such as the unit-information prior and the Jeffreys-Zellner-Siow (JZS) prior (Rouder et al., 2009). However, such priors are not informed by the researcher's study

and expectations, and because of their complex, non-intuitive nature, can be misleading (Tendeiro & Kiers, 2019). Therefore, it has been advocated to only use Bayes Factors if one can meaningfully translate one's hypothesis into a prior distribution, and otherwise use the ROPE approach (Kruschke & Liddell, 2018; Liao et al., 2021; Makowski et al., 2019) or even simply plot the posterior distribution (Tendeiro & Kiers, 2019).

We acknowledge that although the ROPE approach is robust to the choice of prior distribution, it is highly sensitive to the choice of ROPE limits (Kruschke, 2011). Therefore, we acknowledge that the ROPE approach, to some extent, suffers from issues similar to the Bayes Factors approach, and does not provide a perfect solution to quantifying evidence for the null value. However, in many cases, both for researchers and readers, translating one's expectations about plausible effect sizes to ROPE limits that are explicitly and transparently communicated (e.g., in a figure) may be more straightforward than translating these expectations to a density function used as prior, especially when the ROPE limits are specified on the raw response scale. Nevertheless, some arbitrariness remains in deciding on these ROPE limits, and we therefore often create figures displaying the proportion of the posterior inside the ROPE as function of the ROPE width (see the right-hand figures in this blog post by John Kruschke). This allows readers to decide for themselves how convincing they find the evidence in support of a null effect.

Finally, although both ROPE tests and Bayes Factors are influenced by estimation precision or uncertainty, we believe the ROPE test to provide a more transparent representation of estimation uncertainty compared to Bayes Factors. The proportion of the posterior distribution falling within the ROPE varies with sample size. Lower precision or higher uncertainty, caused by, e.g., fewer data points and/or higher sampling noise, will be reflected by a wider posterior distribution, and hence a lower proportion of the posterior falling within the ROPE. By plotting the full posterior distribution in relation to the ROPE, we try to communicate the precision of the estimate in an explicit, transparent manner. This allows the reader to evaluate the confidence with which to draw conclusions for themselves. In Bayes Factors, although low precision or high uncertainty should in theory be reflected by Bayes Factors indicating inconclusive evidence, they have instead been found to result in a bias towards the null model, favouring the null model when they should not (Kruschke, 2013; Tendeiro Kiers, 2019). As sample size increases, the Bayes Factors are more likely to favour the alternative model (Baguley, 2012; Tendeiro & Kiers, 2019). Therefore, Bayes Factors have been criticized to conceal information about the uncertainty of the effect (Kruschke, 2013).

End of Appendix