Bayesian statistics 4/4

Performance, model evaluation, reporting

Oussama Abdoun (MEng, PhD) – oussama.abdoun@pm.me

Bayesian Statistics – CRNL – dec 2024

9 pragmatics reason to go bayesian

Parameter estimation

1. To obtain a measure of uncertainty that can be interpreted intuitively

2. To improve precision & decrease variance (for small samples in particular)

Beneficial “side effects”

7. Thinking deeply about statistical models

8. Learning and reasoning about effect sizes

9. Having an opportunity to read the literature!

Hypothesis testing

3. To test hypotheses of interest formally

4. To obtain a quantitative measure of evidence (BF)

5. To obtain evidence in favor of the null

6. To eliminate the multiple comparison problem

1.

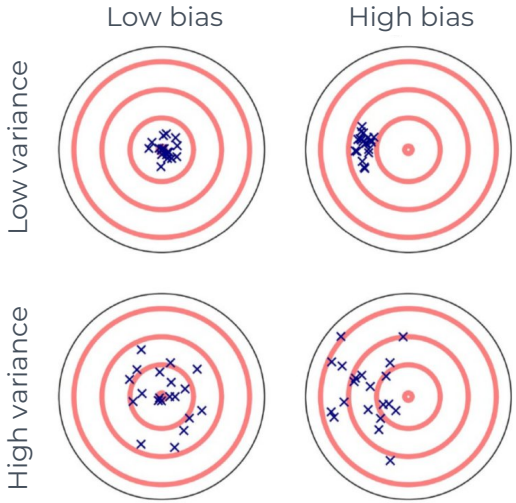
Performance of bayesian inference

Bias-variance tradeoff

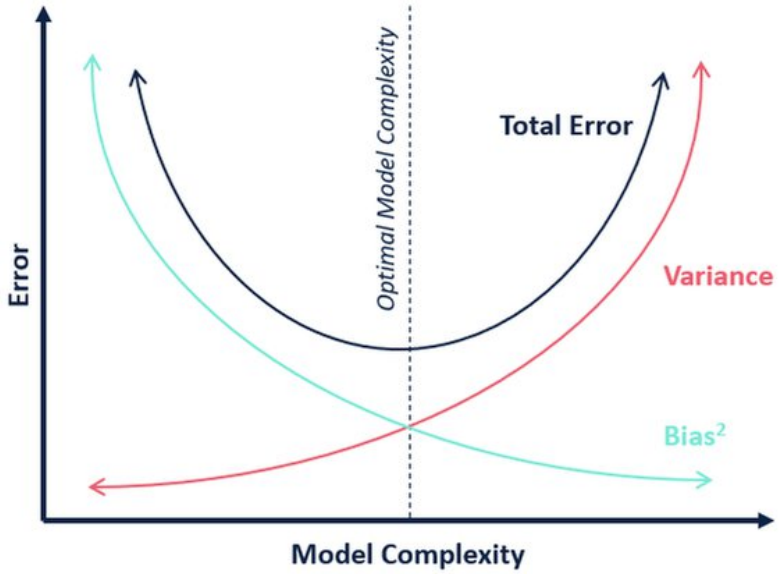
Bias = systematic error due to inadequate model (**underfitting**)

Variance = sensitivity to small fluctuations in the data (**overfitting**)

→ variability of parameter estimates across replications



Target center = true value
Crosses = model predictions



Bayes Factor

As an Occam's razor

$$BF_{21} = \frac{p(y|M_2)}{p(y|M_1)} = \frac{\int p(y|\theta_2)\,p(\theta_2)d\theta_2}{\int p(y|\theta_1)\,p(\theta_1)d\theta_1}$$

Model complexity is automatically **penalized** by the Bayes Factor:

— the more parameters, the more the prior is spread out over "irrelevant" regions, the more "diluted" the predictive power of the model

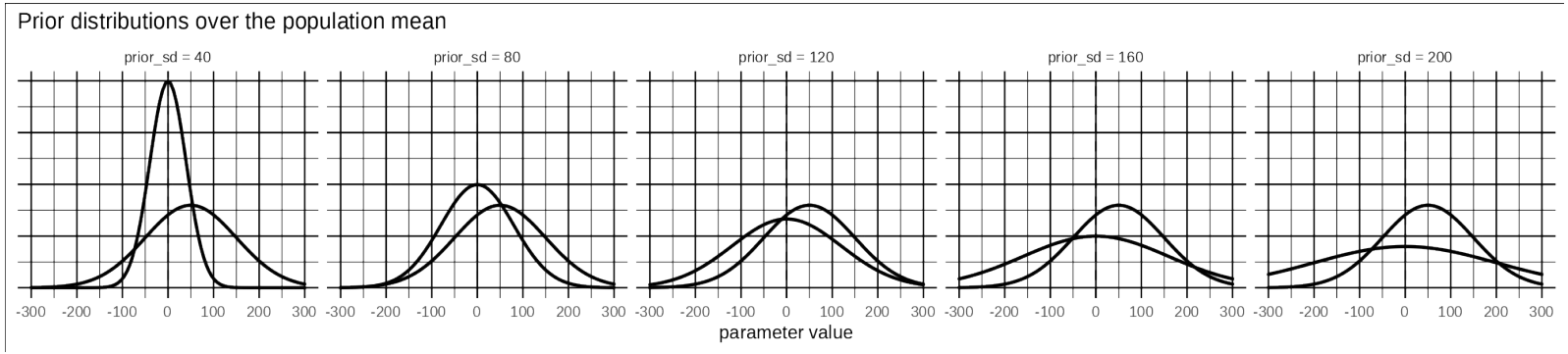
— diffuse priors follow the same logic

Bayesian vs. Frequentist

A simulation study

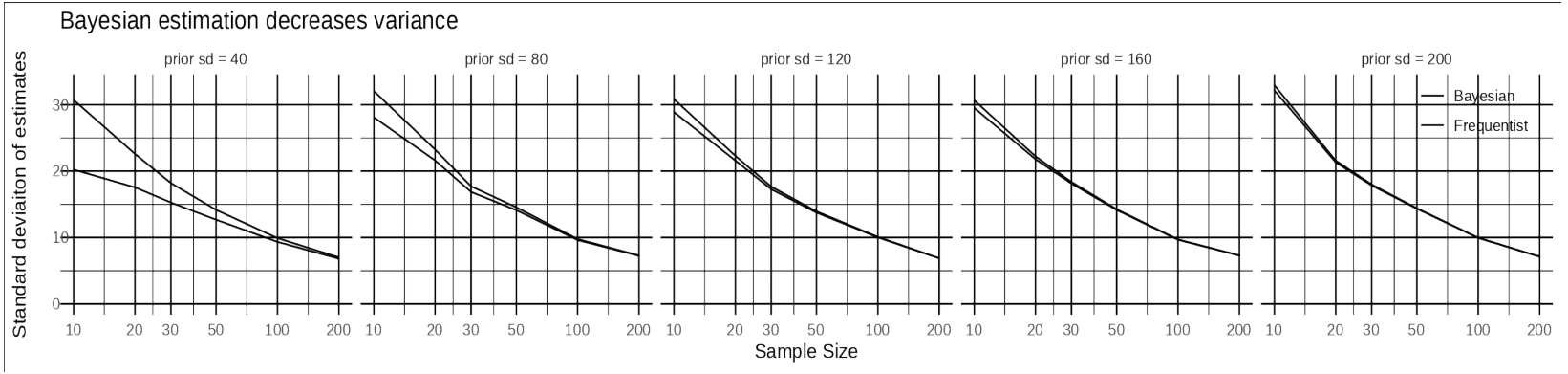
Let’s model a population with true mean = 50 and std = 100 (for example, the difference in reaction times between 2 groups or conditions). The values correspond to a Cohen’s d of 0.5.

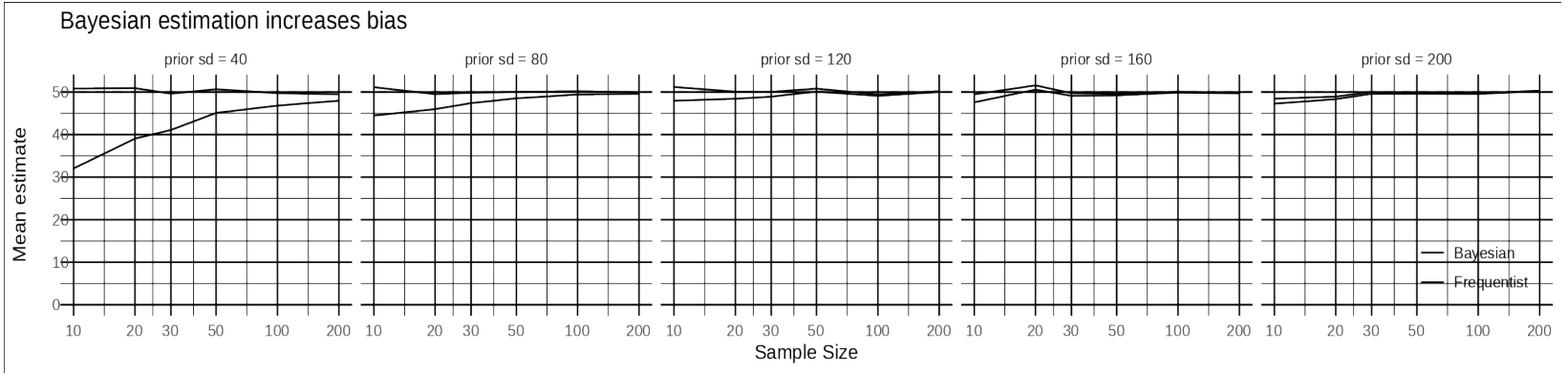
We will compare frequentist and bayesian results systematically, for a range of prior variance and sample sizes. For each combination of simulation parameters, we generate 1000 samples.



Bayesian vs. Frequentist

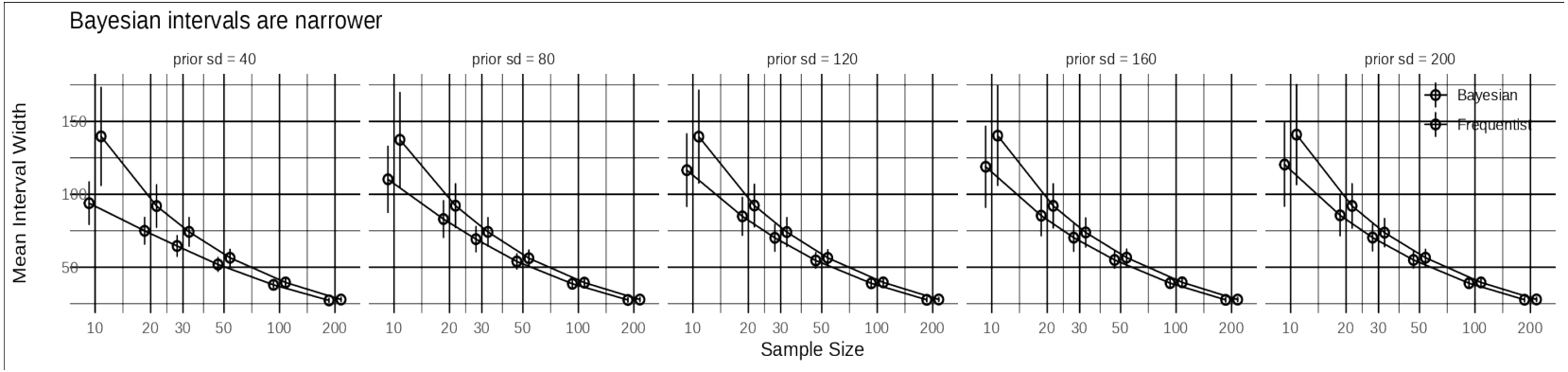
Bias & variance

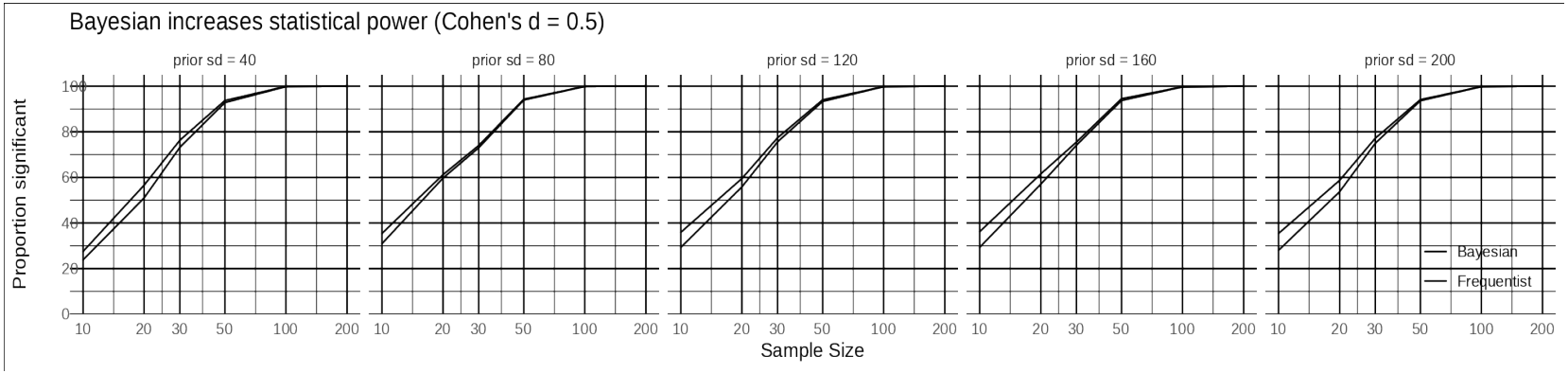




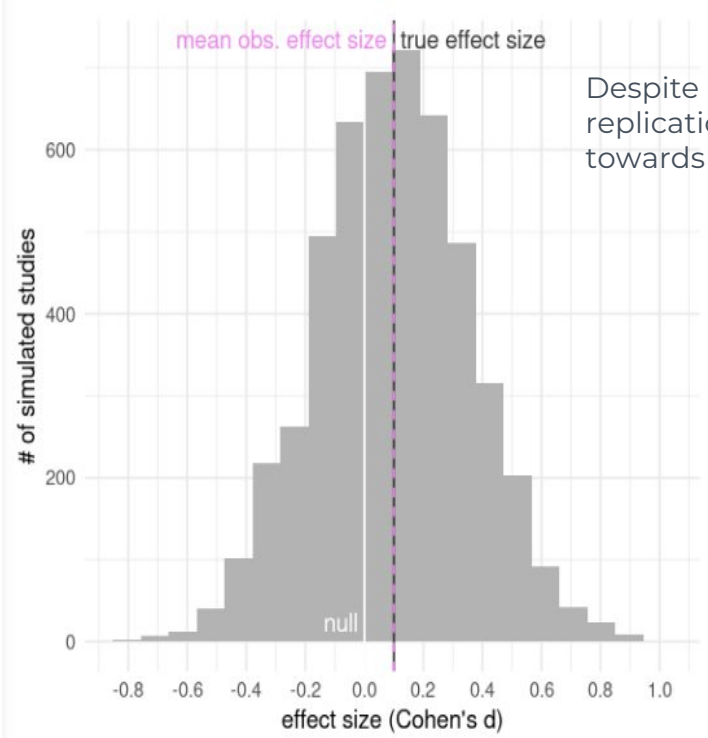
Bayesian vs. Frequentist

Improved precision



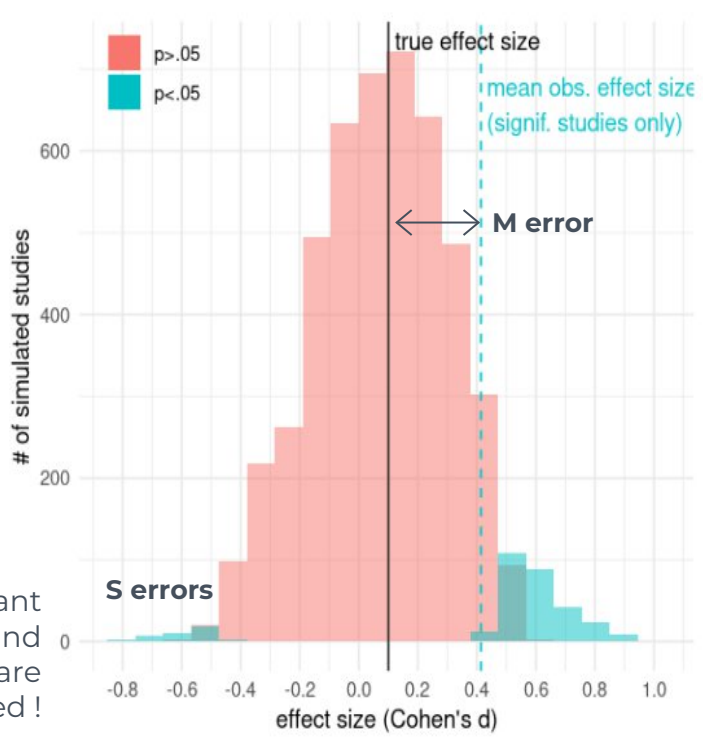


Type M and type S errors



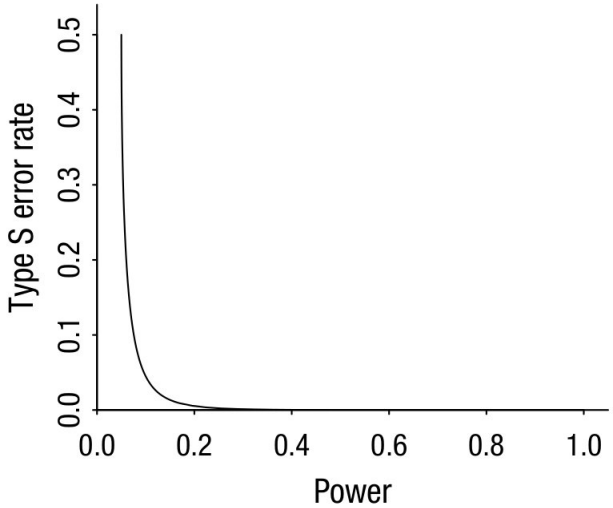
Despite low power, replications will converge towards the true effect size...

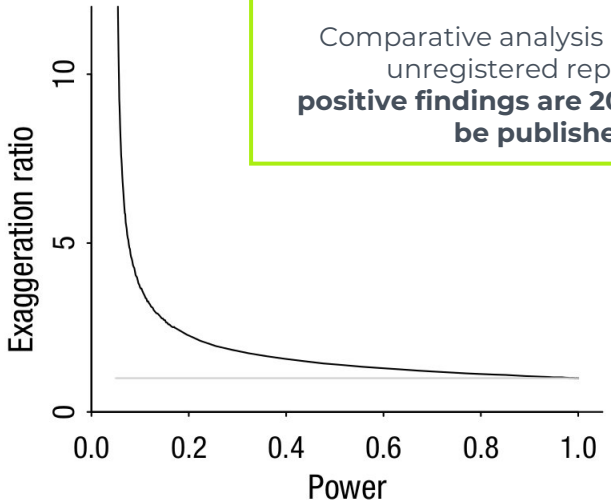
...unless non-significant findings are ignored and only significant ones are considered !



Type M and type S errors

The previous analysis can be extended to a range of true statistical powers :





These figures assume that only significant results enter meta-analyses, which implies extreme **selective reporting** (*aka* **publication bias**).

Comparative analysis of registered and unregistered reports suggest that **positive findings are 20x more likely to be published** than null ones.

Gelman & Carlin (2014). *Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors* [link]

Bayesian vs. Frequentist

The multiple comparison problem

If the prior distribution matches the true distribution of effect sizes:

— no type M error

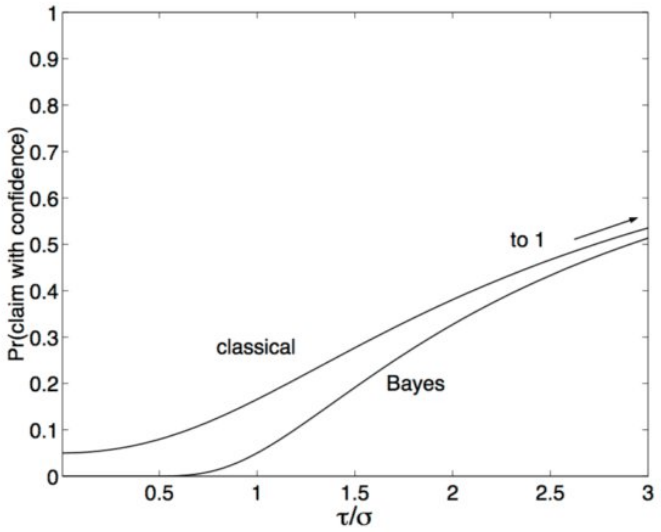
— type S error always < 2.5%

— no need of a correction for multiple comparison

If the prior variance **overestimates*** the true variance of effect sizes, we start loosing the beneficial effects of the Bayesian inference.

If the prior variance **underestimates*** the true variance of effect sizes, we start loosing statistical power with Bayesian inference.

*by a factor 3 or more to have visible effects



***small** effects and/or sample size* ⟷ ***large** effects and/or sample size*

Gelman (2016). *Bayesian inference completely solves the multiple comparisons problem* [link]

Bayesian vs. Frequentist

Sequential hypothesis testing / optional stopping

The standard procedure in the frequentist paradigm:

— postulate an effect size *d*

— calculate sample size needed to reject H_0 under such an effect (N)

— collect data… and pray

But we could obtain $p < .05$ with less data, especially if true effect size > *d*

Unfortunately, we can't run the stats after each new data point and stop when $p < .05$
⇒ substantial increase of the risk of false positives!

What about Bayesian inference?

We can legitimately run the analysis after each data point and stop collecting data when…

— a certain BF-threshold has been reached (e.g. BF_{10} or BF_{01} = 10, or 30): Schönbrodt et al. 2015 [link]

— a certain precision of the posterio distribution has been reached: Kruschke 2013 [link to blog post]

Rouder (2014). Optional stopping: No problem for Bayesians [link]

What about significance hacking?

Questionable research practices that increase the risk of type I error include any change in the analysis pipeline performed after looking at the results:

— data sifting: excluding participants, data points, conditions...

— model tweaking: adding or removing interactions, predictors, moderators...

— selective reporting of a subset of dependent variables, experimental factors...

In frequentist statistics, they are incentivized by the dichotomous rule of “p < .05”

In bayesian statistics, treating BF and other measures as continuous measures of evidence (as it should) deflates the (conscious or unconscious) urge for significance hacking. You can further protect yourself with the **usual measures**:

— think of all the details of your analyses ahead of time (**preregistration**)

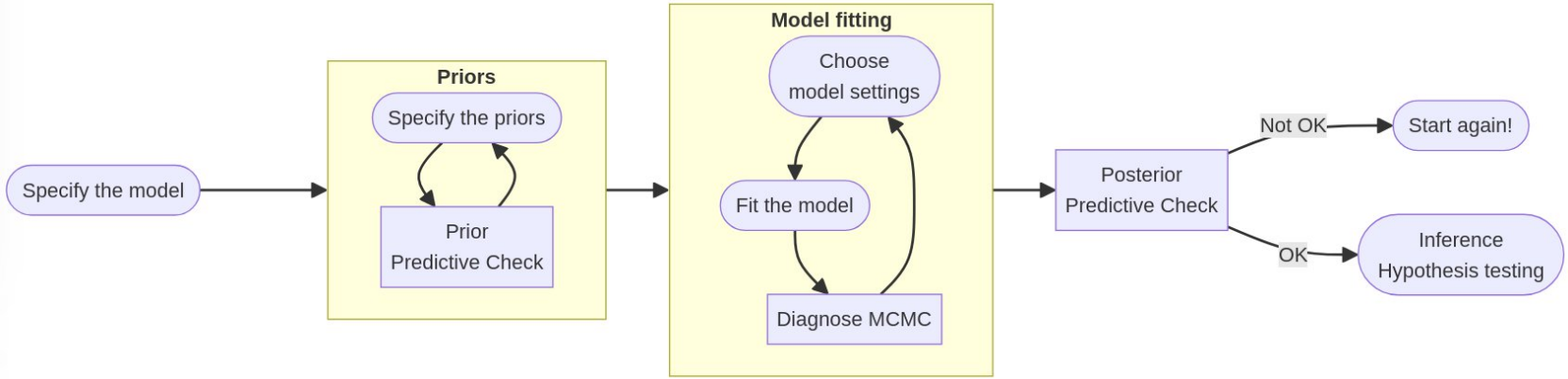
— **validate** your model *before* doing hypothesis testing

— **report everything** you did/tried

2.

Model checking & diagnosis

Workflow of bayesian modeling



See Schad, Betancourt & Vasishth (2021) for a more sophisticated workflow [link]

Prior predictive checks

(A) Spe Specify prior distributions for each parameter of the model

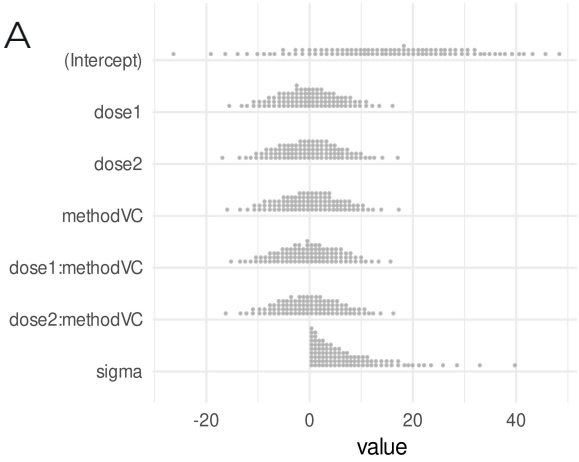
(B) Randomly draw a set of parameters from the priors

(C) Plug the values into the model to generate a synthetic dataset

Repeat (B) and (C) many times (a few tens or hundreds).

Adjust priors if needed and restart from (A).

A



B

C

5 $y = -9.4 + 2.1 \times dose1 + 11.9 \times dose2 + 1.1 \times methodVC + 7.8 \times dose1 : methodVC - 8.7 \times dose2 : methodVC + \varepsilon,\ \varepsilon \sim N(0, 16.7)$

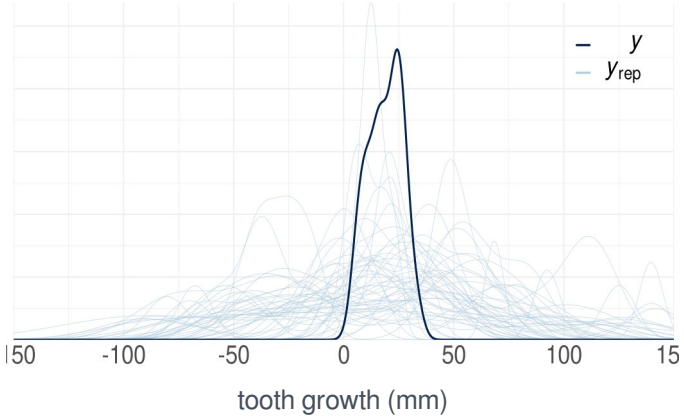
4 $y = 28.3 - 2.2 \times dose1 - 9 \times dose2 + 11.1 \times methodVC + 4.2 \times dose1 : methodVC + 7.3 \times dose2 : methodVC + \varepsilon,\ \varepsilon \sim N(0, 3)$

3 $y = 27.9 - 2.9 \times dose1 + 10.3 \times dose2 + 9.5 \times methodVC - 0.2 \times dose1 : methodVC - 1.4 \times dose2 : methodVC + \varepsilon,\ \varepsilon \sim N(0, 7.2)$

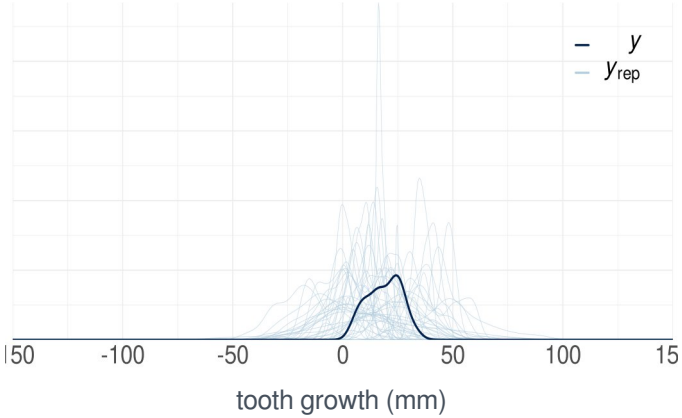
2 $y = 10.6 + 0.4 \times dose1 + 3.4 \times dose2 + 13.9 \times methodVC + 0.5 \times dose1 : methodVC + 3.2 \times dose2 : methodVC + \varepsilon,\ \varepsilon \sim N(0, 5.9)$

1 $y = 25.7 - 5.1 \times dose1 - 7.7 \times dose2 - 0.1 \times methodVC + 1.2 \times dose1 : methodVC - 1.3 \times dose2 : methodVC + \varepsilon,\ \varepsilon \sim N(0, 13.7)$

Prior predictive checks



rstan's **default** priors



custom **informative** priors

Prior predictive checks

Example with the effect of meditation state on mismatch negativity

Fucci, Poublan, Abdoun & Lutz (2022). *No effect of FA and OM meditation on EEG auditory MMN* [link]

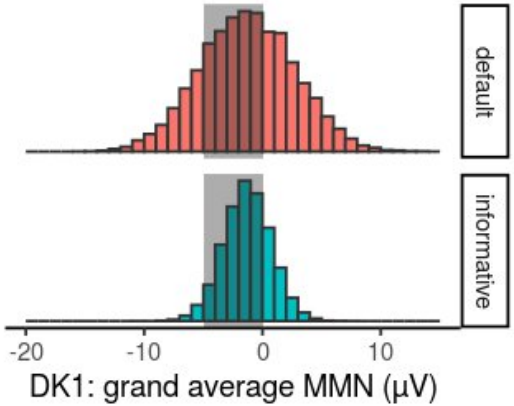
	rstanarm's default	informative
Fixed effects : intercept	$N(-1.3, 4.1^2)$	$N(-1.3, 2^2)$
Fixed effects : coefficients	$N(0, 5.8^2)$ to $N(0, 9.9^2)$	$N(0, 2^2)$
Random effects : variance of by-subject intercept	$exp(1)$	$exp(1)$
Residuals variance	$exp(0.6)$	$exp(1)$

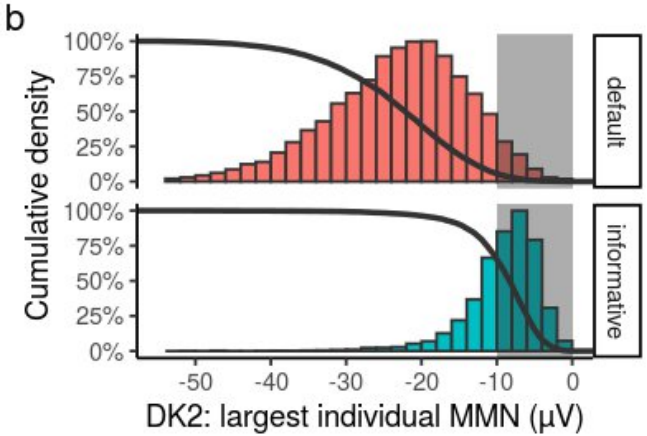
Prior predictive checks

Consistency with domain expertise using summary statistics

"**DK1:** The grand average MMN amplitude is in the order of -1 to -5µV."

"**DK2:** Individual MMN amplitude only rarely exceeds -10µV (see for example figure 3 in Pekkonen et al. 1995, figure 1 in Escera et al. 2000, figure 4 in Näätänen et al. 2012, figure 2 in Kim et al. 2020)."



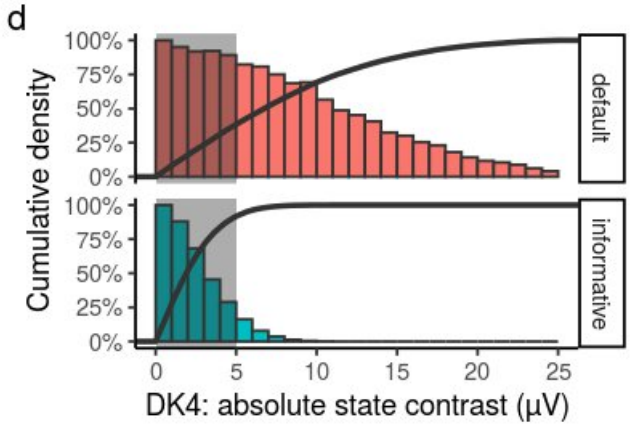


Prior predictive checks

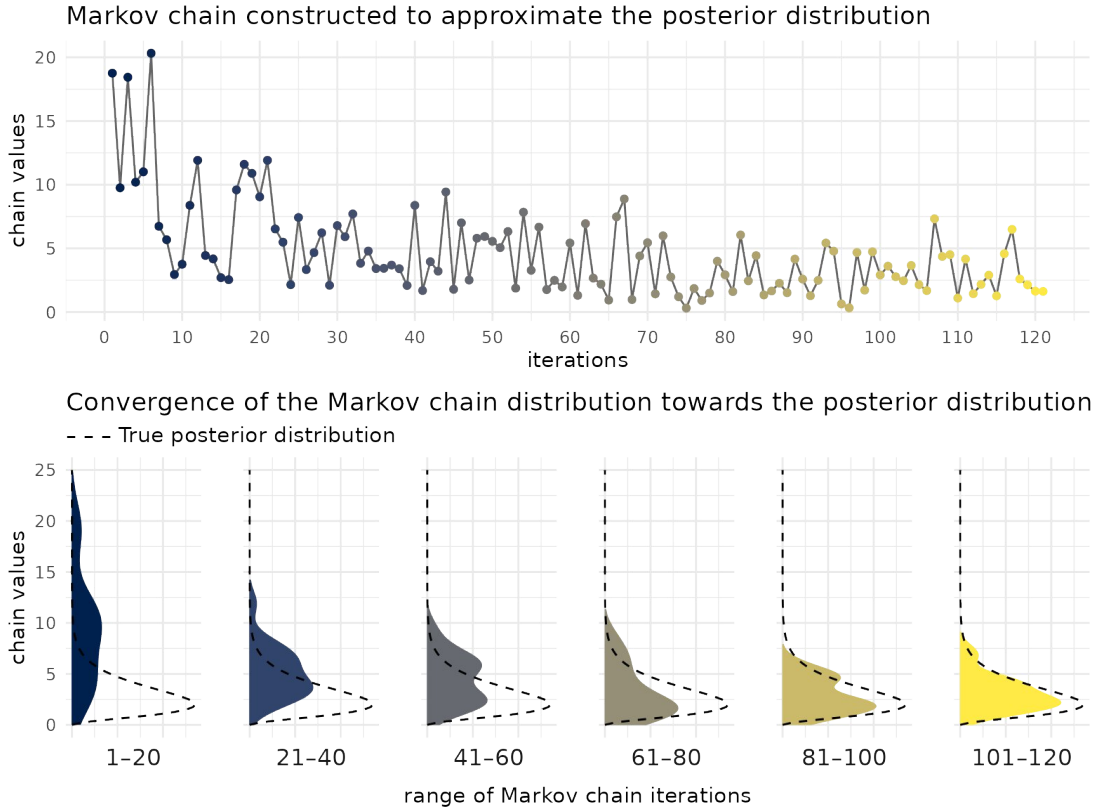
Consistency with domain expertise using summary statistics

"**DK3:** (...)"

"**DK4:** Effects of experimental conditions and differences between populations are in the order of 1µV and rarely exceed 3µv (see for example the studies reported in reviews by Näätänen et al. 2007 and Kujala & Leminen 2017). Importantly, contrasts between experimental conditions and/or populations cannot exceed the magnitude of MMN in absolute value, and therefore are typically smaller than 5µV."



Numerical simulation of the posterior: MCMC



When prior distributions are **not conjugate distributions** of the likelihood, we don't have an explicit expression of the posterior distribution anymore and we need to calculate it **numerically**. We use **Markov chain Monte Carlo** (MCMC) techniques, a family of algorithms sharing the same basic procedure:

1. A **Markov chain** (= random process where each sample depends probabilistically on the previous one) is created such that it, *in the long run*, its distribution converges towards the true posterior distribution.

2. A large number of **samples** (several to tens of thousands) are generated **iteratively** from the Markov chain.

3. Initial samples (typically 1000) are considered as not converged yet and rejected ("**warm up**" phase); the rest of the samples is used as an **approximation of the posterior** distribution.

Diagnosis of MCMC convergence

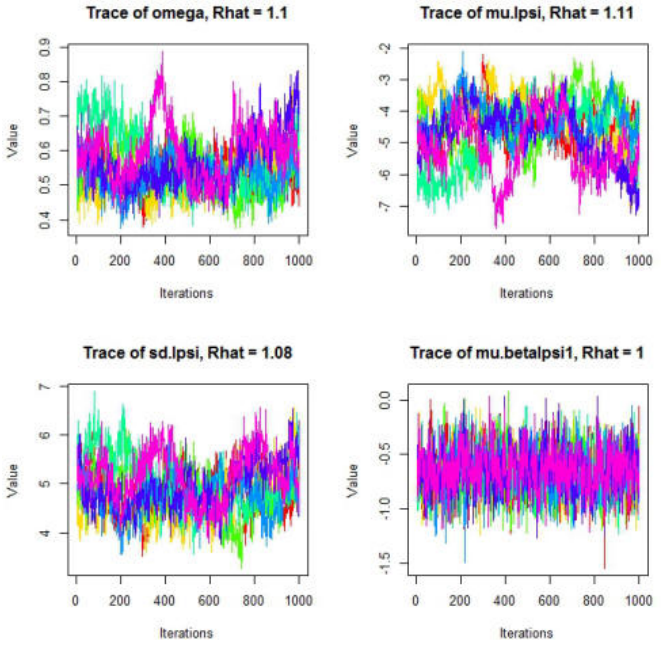
To check whether the MCMC converges properly, we start multiple independent chains (typically 4) and check:

— that they reach **stationarity**

— that they converge to the **same distribution** (mean and variance) *aka* "**chain mixing**"

⇒ numerical assessment = $\hat{R}$ (*aka the* Gelman-Rubin statistic)
~ between-chain variance / within-chain variance

Recommendation: $\hat{R} \leq 1.1$

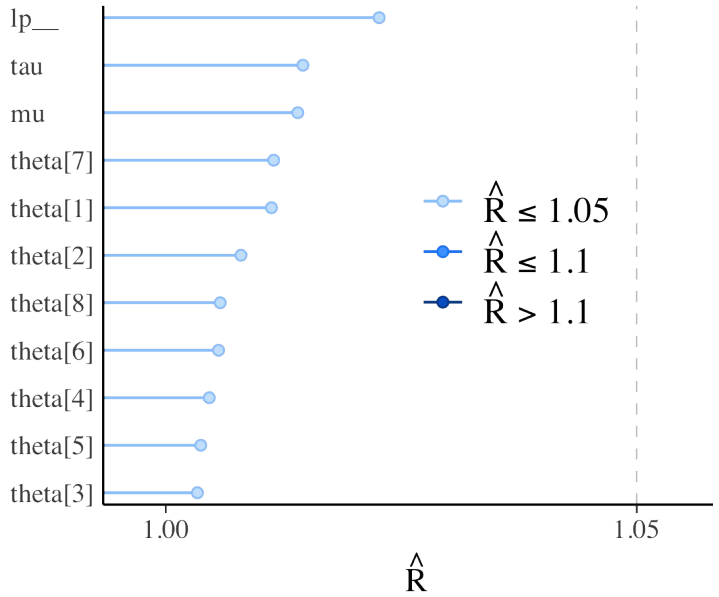


Diagnosis of MCMC convergence

MCMC convergence should be checked for **each parameter** of the model.

What to do when convergence is poor?

1. Use more informative priors
2. Collect more data
3. Increase the number of MCMC iterations
4. Simplify the statistical model



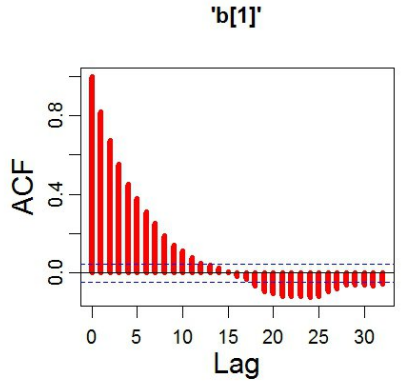
Diagnosis of MCMC convergence

Markov processes have a memory of the previous iteration. Therefore, Markov chains have **autocorrelation** and some of the information that they contain is “redundant”.

⇒ numerical assessment = **effective sample size** (n_{eff})

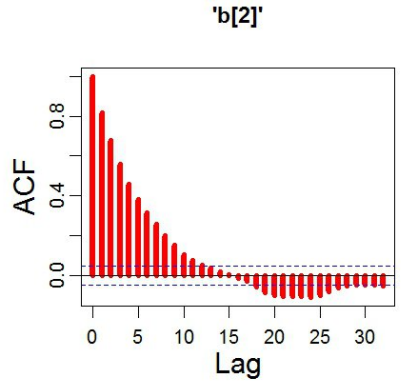
= number of independent draws from the posterior distribution

Recommendation: $n_{eff} \geq 500$ for parameters of interest



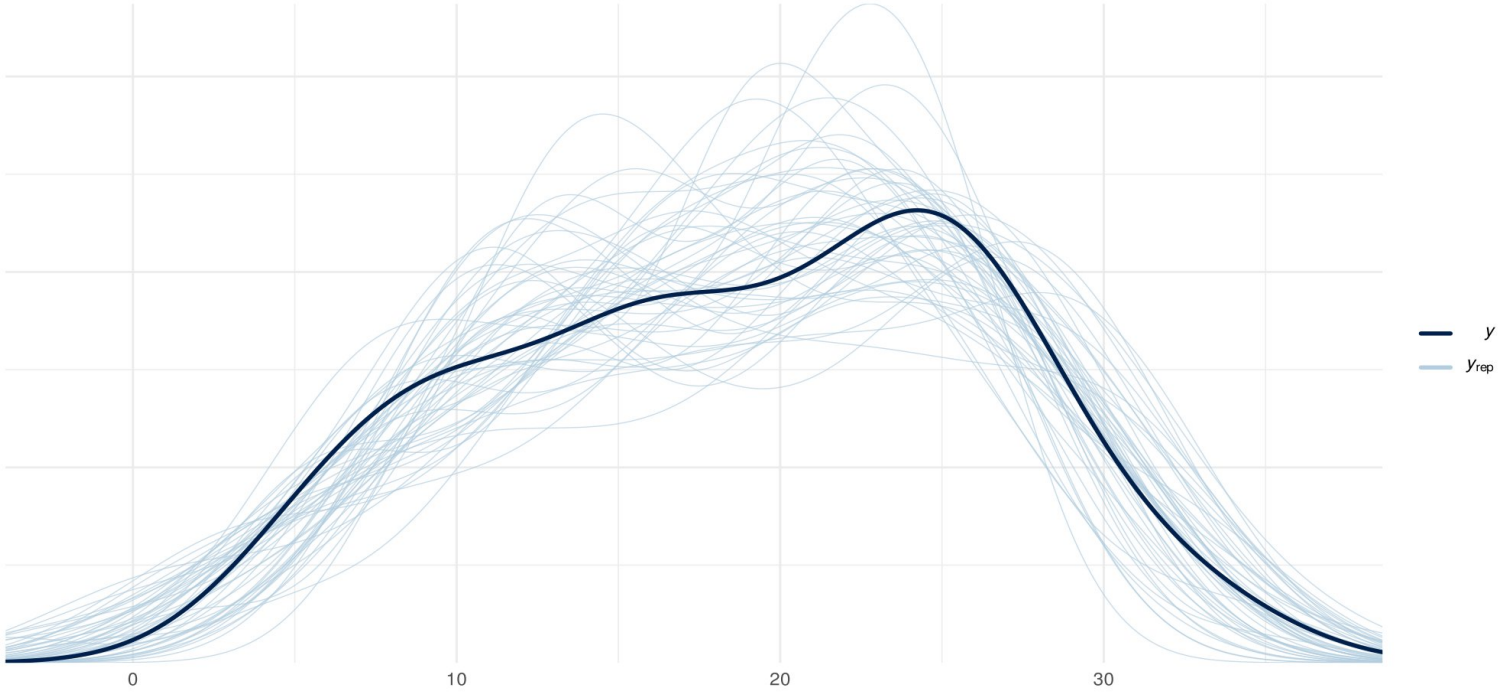
What to do when there are not enough effective samples?

The same as with insufficient mixing, especially increasing the number of MCMC iterations.



Posterior predictive checks

Similar procedure as the prior predictive check



3.

Reporting

Bayesian Analysis Reporting Guidelines

Kruschke (2021). *Bayesian Analysis Reporting Guidelines* [link]

Common mistakes : a test

		prevalence in the literature
	*"Since both the confidence interval (−.09 and .67) and the BF_{10} (1.2) do not point towards a true difference, this can be considered a very small or non-existent effect."*	***<4%***
	*"The Bayesian test showed a positive, but small, effect of Load on tracking pupil size ($BF_{incl.}$=7.506)"*	***~4%***
	*"For 6-year-olds, there was no difference between environments (t(52)=1.0, p=.31, d=0.3, BF=0.22)"*	***35%***
	*"A BF_{10} = 10 means that the H_1 is ten times more likely to be true than the H_0"*	***28%***
	*"These analyses revealed a Bayes factor of $BF_{1,0}$ = 0.19 in the attention condition, supporting the null hypothesis."*	***62%***

Common mistakes

	Example	Explanation	Correction
1,6—Describing or interpreting the BF as posterior odds	*"A BF_{10} = 10 means that the H_1 is ten times more likely to be true than the H_0"*	**Not rigorous:** BF are likelihood ratios. They are equal to posterior odds only when prior odds equal to 1	State that prior odds equal to 1.
3a—Missing explanation for the chosen priors	Using default priors without justification.	**Not transparent**	
3b—No mention of the priors used		**Not reproducible**	
3c—Incomplete info regarding the priors used	Mentioning using a Cauchy distribution without specifying the value of the scale parameter.	**Not reproducible**	
4—Not referring to the comparison of models	*"These analyses revealed a Bayes factor of $Bf_{1,0}$ = 0.19 in the attention condition, supporting the null hypothesis."*	**Not rigorous:** support for one model depends strongly on the model it is compared to.	Be more verbose, or explain in the Methods once for all how to interpret the BF.

Tendeiro et al. (2024). *Diagnosing the Misuse of the Bayes Factor in Applied Research* [link]

Common mistakes

	Example	Explanation	Correction
5—Making absolute statements	*"For 6-year-olds, there was no difference between environments (t(52)=1.0, p=.31, d=0.3, BF=0.22)"*	**Wrong:** the BF is a continuous measure of evidence, it can not prove anything with absolute certainty.	*"For 6-year-olds, there was weak evidence for an absence of difference ..."*
7—Considering the BF as effect size	*"The Bayesian test showed only positive, but smaller, effect of Load on tracking pupil size ($BF_{incl.}$=7.506)"*	**Wrong:** Erroneous association between statistical and practical significance.	
9—Inconclusive evidence as evidence of absence	*"Since both the confidence interval (−.09 and .67) and the BF_{10} (1.2) do not point towards a true difference, this can be considered a very small or non-existent effect."*	**Wrong:** Bayes factors close to 1 imply that the evidence for either model under comparison is about the same.	*"Since..., the data is inconclusive with respect to the null or alternative hypotheses."*
10—Interpreting ranges of BF values only	*"Evidence for greater disgust in the experimental group was strong (BF_{10} > 10), but there was only weak evidence for a difference in other emotions (BF_{10} 's < 3)"*	**Not rigorous:** BF is a continuous measure of evidence	Report BF values precisely, and interpret them individually.