Consistency of Bayes Factor estimates in Bayesian Analysis of Variance

Roland Pfister

Trier University, Germany

This is the final author manuscript, not the official publication in *Psychological Method*s. See https://doi.org/10.1037/met0000703

Correspondence

Prof. Dr. Roland Pfister

General Psychology

Trier University

Johanniterufer 15

54290 Trier, Germany

Email: mail@roland-pfister.net

Author note: R.P. is funded by the Heisenberg Programme of the German Research Foundation (PF 853/10-1, 490925504).

Abstract

Factorial designs lend themselves to a variety of analyses with Bayesian methodology. The de-facto standard is Bayesian Analysis of Variance (ANOVA) with Monte Carlo integration. Alternative, and readily available methods, are Bayesian ANOVA with Laplace approximation as well as Bayesian *t*-tests for individual effects. This simulation study compared the three approaches regarding ordinal and metric agreement of the resulting Bayes Factors for a 2x2 mixed design. Simulation results indicate remarkable disagreement of the three methods in certain cases, particularly when effect sizes are small and studies include small sample sizes. Findings further replicate and extend previous observations of substantial variability of ANOVAs with Monte Carlo integration across different runs of one and the same analysis. These observations showcase important limitations of current implementations of Bayesian ANOVA. Researchers should be mindful of these limitations when interpreting corresponding analyses, ideally applying multiple approaches to establish converging results.

Introduction

Analysis of Variance (ANOVA) is a particularly flexible tool to analyze studies with multi-factor designs. It is also among the most widely used techniques in psychological science, and has been for decades (Blanca et al., 2017; Counsell & Harlow, 2017; Skidmore & Thompson, 2010; Zhou & Skidmore, 2017; see also Edgington, 1964). A particularly appealing property of classical ANOVA is its ability to accommodate complex factorial designs by assessing main effects and especially interactions among factors. These factors can further represent within-subject and between-subjects variables so that any factorial design lends itself to analysis within the ANOVA framework.

Proponents of Bayesian methods of statistical inference have built on the obvious selling points of ANOVA methodology by constructing Bayesian equivalents of this analytical procedure (Rouder et al., 2012; van den Bergh et al., 2020). Current Bayesian approaches to ANOVA indeed appear to implement many of the elegant properties of classical ANOVA at first sight (Rouder et al., 2017). Closer inspection of the available methods points towards important differences, however, that are only slowly beginning to be addressed (Oberauer, 2022; Pfister, 2021). Here I report on an unwanted behavior of Bayesian ANOVA that has critical implications for the reliability of current methods. In short, there is substantial disagreement between different available methods to calculate Bayes Factors for main effects and interactions. The evidence for a particular empirical effect thus critically depends on the (arbitrary) choice between seemingly comparable and equivalent computational methods.

The goal of documenting this instance of unwanted behavior for Bayesian ANOVA should not be seen as trying to argue for or against a certain analytical approach, such as Bayesian rather than frequentist methods of statistical inference (for an overview of this discussion see, e.g., Harlow et al., 2017; Tendeiro & Kiers, 2019, 2022a; Van de Schoot et al., 2017; Wagenmakers, 2007). I believe, however, that applying statistical methods is only possible when being aware of

critical pitfalls of different methodologies. Two such pitfalls have recently been discussed in the context of Bayesian ANOVA.

A first pitfall of Bayesian ANOVA concerns the specification of the error term in within-subject designs (Oberauer, 2022). Within-subject designs are prevalent in psychology because they allow for removing variance associated to a particular person. This property, in turn, helps to assess group-level effects. Within-subject designs, therefore, are particularly helpful when overall variability between people is large whereas experimental effects are relatively small. One way to remove inter-individual variability is to remove a person's overall mean from all their observations, in order to center each participant on the sample mean. This approach is known as the random intercept model. Its critical assumption is that any (fixed) effect in the design is equal across participants so that any remaining variability across conditions is treated as noise. Another way to remove inter-individual variability is to allow for different (fixed) effects for each participant. This approach is known as random slopes, and it corresponds to the common practice of using subject-by-condition error terms in classical ANOVA (e.g., Franz & Loftus, 1994). Whether the random intercept approach or the random slope approach is more suited for analyzing a dataset at hand depends on how different sources of error variance are distributed across the dataset. However, not accounting for random slopes may at times affect results substantially (Barr et al., 2013). Common implementations of Bayesian ANOVA, such as the R package BayesFactor (Morey & Rouder, 2022) and JASP (JASP Team, 2022) used to include only random intercepts, however. This choice has been shown to yield systematically increased rates for statistical decision errors, particularly false positives (Type I errors). This state of affairs is currently beginning to change with random slopes being incorporated into the JASP software package (van den Bergh et al., 2022). Researchers should therefore use this updated methodology when conducting Bayesian repeated measures ANOVA to avoid biased results.

A second pitfall concerns the variability of Bayes Factor estimates across different iterations of one and the same analysis (Pfister, 2021). Standard implementations of Bayesian ANOVA and JASP use iterative procedures to estimate Bayes Factors (Rouder et al., 2012).[1] Repeatedly running the same analysis on a constant dataset therefore produces variable results on each run. In rare instances, this may even produce a Bayes Factor supporting the null hypothesis on one iteration and a Bayes Factor supporting the alternative hypothesis on another iteration (Pfister, 2021). Researchers should therefore take measures to safeguard against potential outlier results, e.g., by repeatedly running each analysis to assess the consistency of the resulting Bayes Factors.

Iterative approaches to Bayes Factor computation are only one of several possible methods, though. Crucially, there are also deterministic approximations such as the Laplace approximation (Long et al., 2013; Ryan et al., 2016; Schillings et al., 2020), and these methods are also implemented in current software packages (Morey & Rouder, 2022). How Bayes Factors resulting from such deterministic approximations compare to those obtained from iterative Monte Carlo integration has not yet been studied empirically to date, though. Moreover, many study designs with two factor levels can be analyzed by a series of (Bayesian) *t*-tests instead. This is especially true for full within-subjects designs, where even interactions of two-level factors can be assessed via *t*-tests for paired differences. In classical, frequentist ANOVA, there is a clear relation in this case with the *F*-statistic of an ANOVA being directly related to the *t*-statistic of the

[1] In a previous publication, I had stated that these algorithms would be of the Markov Chain Monte Carlo (MCMC) variety (Pfister, 2021). This statement was motivated by corresponding documents suggesting that this would be the case, such as: “result is a BFmcmc object” in the documentation of the “compare” function of the BayesFactor package for R (Morey & Rouder, 2018) or: “When running the ANOVA analysis [...] the results are likely to be very slightly different to the ones in this presented chapter. This is because the analyses are based on numerical algorithms like Markov chain Monte Carlo (MCMC)” (Goss-Sampson et al., 2020; see also van den Bergh et al., 2020; van Doorn et al., 2020). Richard Morey (personal communication), clarified that this is not the case, however, in response to this article. Current implementations build on Monte Carlo integration instead (Rouder et al., 2012). This applies at least to the BayesFactor package. Matters are more difficult to evaluate for JASP, because available documentation of the software explicitly state that they use the algorithms of the BayesFactor package while also maintaining that Bayes Factors are based on MCMC methodology.

corresponding *t*-test with $\sqrt{F(1, df)} = t(df)$, with df = degrees of freedom. The same holds true for mixed designs where the main effect of the between-subjects factor and the interaction in a 2x2 design are identical to corresponding *t*-tests. For Bayesian inference, such a relation cannot exist because there are different ways to compute Bayes Factors for Bayesian ANOVAs and also for Bayesian *t*-tests. I still find it instructive to ask how much these two Bayesian procedures agree for a given dataset when using their most common algorithmic implementation.

The main goal of the present study therefore was to assess the agreement of three ways to test specific effects in a factorial design: Bayesian *t*-tests, Bayesian ANOVA with Laplace approximation and Bayesian ANOVA with Monte Carlo integration. This was done in a set of simulation studies.

Data and Hypotheses

As a model case I used a 2x2 mixed design with one within-subject factor and one between-subjects factor (Pfister, 2021). Such a design obviously comes with three potential effects of interest (excluding the intercept): The main effect of the within-subject factor, the main effect of the between-subjects factor, and their interaction. I simulated different datasets for such a design, using different sample sizes, population effect sizes, and effect size priors. I then analyzed each dataset with (1) separate Bayesian *t*-tests for main effects and the interaction, (2) Bayesian ANOVA with Laplace approximation, and (3) Bayesian ANOVA with Monte Carlo integration. Note that the first two analyses return deterministic results whereas the third option returns varying results for repeated analyses of the same dataset. To compensate for the latter variability, I used the median of 100 re-runs of the same analysis for every dataset.

The main question was how much the Bayes Factors resulting from the three methods would agree. This question has two parts. The first part concerns *ordinal agreement* between the measures. That is, do these measures agree on a rank order of how much different datasets

support either the null hypothesis of no effect (H_0) or the alternative hypothesis (H_1)? The second part concerns *metric agreement* in terms of how strongly a particular dataset supports either hypothesis. I had initially expected high ordinal agreement but wondered about potentially relevant patterns in terms of metric agreement. The results were surprising, however, in that they showed substantial disagreement already at an ordinal level. I therefore focus on this aspect in the following while only presenting a short overview on metric agreement.

Simulation Method

Figure 1 provides an overview of the simulation procedure. All simulations were implemented in custom R code, running R version 4.0.2. The code used for simulating the data, for assessing the validity of the simulation settings, and for running all following analyses is available on the Open Science Framework (https://osf.io/4u9sx/). This study was not preregistered.

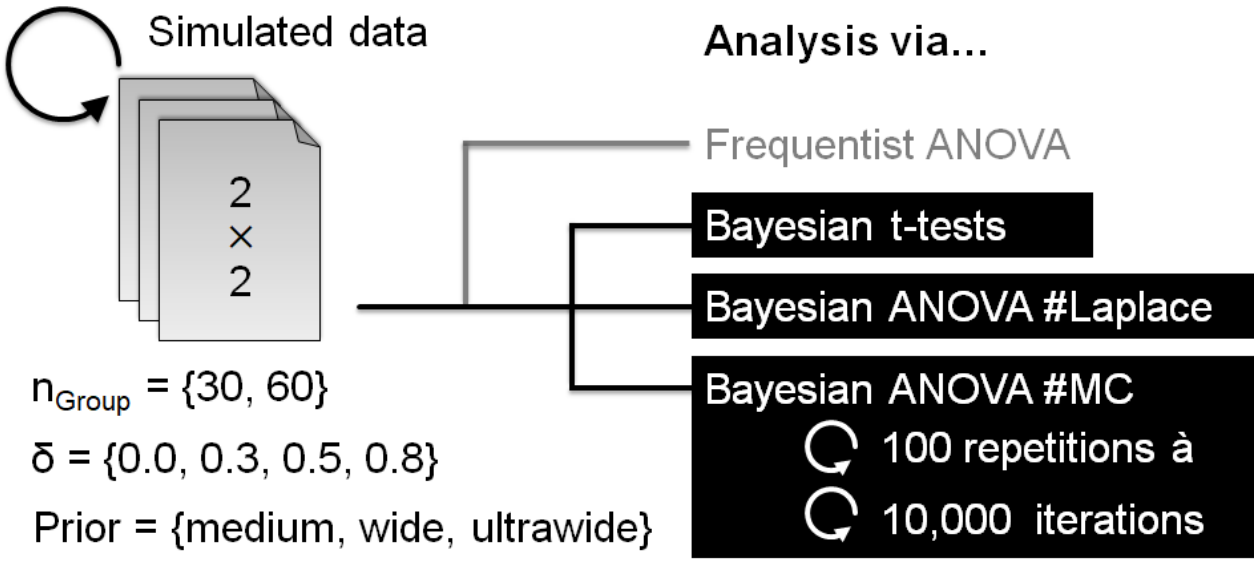

Figure 1. Overview of the simulation procedure. To compare the estimates of different Bayesian methods for analyzing data from factorial designs, I sampled independent datasets following a 2x2 mixed design for each combination of sample sizes (n for each of two groups), population effect sizes (δ) and prior settings for Bayesian analyses. Each dataset was analyzed by frequentist Analysis of Variance (ANOVA) to assess the accuracy of the sampling, followed by corresponding Bayesian tests. These tests comprised (1) separate Bayesian *t*-tests for each main effect and the interaction, as well as Bayesian ANOVAs either (2) with Laplace approximation or

(3) with Monte Carlo (MC) integration. The latter ANOVAs used 10,000 MC iterations and they were repeated 100 times for each dataset to arrive at a stable estimate of the resulting median Bayes Factor. I then compared the results of the three Bayesian methods for *ordinal agreement* by computing Spearman rank correlations across datasets as well as for *metric agreement* by assessing the mean of the (log-transformed) Bayes Factors across design cells. Initial analyses included 100 datasets per design cell whereas selected observations were reproduced with an increased sample of 10,000 datasets.

Simulation rationale

The simulations involved an iterative procedure that I will summarize in the following. For each combination of sample size, effect size, and prior setting (see the section on *Simulation parameters* below) I sampled 100 datasets of a 2x2 mixed design. Data was drawn from normal distributions using (pseudo-)random sampling.

For each of the datasets, I computed a standard, frequentist ANOVA using Type III sum of squares, and stored the resulting *F*- and *p*-values for each main effect and the interaction. I further computed a paired *t*-test to approximate the main effect of the within-subjects factor (note that I did not compute *t*-tests for independent samples for the remaining two effects, because their results are necessarily identical to the results of the main effect of the between-subjects factor and the interaction of the frequentist ANOVA).

I then assessed all three effects of the mixed design with three Bayesian methods each: (1) Bayesian *t*-tests, (2) a Bayesian ANOVA with Laplace approximation, and (3) a Bayesian ANOVA with Monte Carlo integration. All tests used the *ttestBF* function or the *anovaBF* function as implemented in the BayesFactor package version 0.9.12-4.2. I extracted Bayes Factors with evidence for the alternative hypothesis in the numerator (BF_{10}), and I will describe these critical measures in the following.

Bayesian *t*-tests. For the main effect of the between-subjects factor, I averaged the individual scores across the within-subjects factor, computed a Bayesian *t*-test for independent samples to compare the two groups, and saved the corresponding Bayes Factor. For the main effect of the within-subjects factor, I pooled the data across the between-subjects factor, and

computed a Bayesian *t*-test for paired samples. For the interaction, I computed pairwise differences across the within-subjects factor for each simulated subject, computed a Bayesian *t*-test for independent samples to compare these paired differences between both groups, and saved the corresponding Bayes Factor.

Bayesian ANOVA with Laplace approximation. I computed a Bayesian ANOVA, coding both main effects and the interaction term as fixed factors, subject as random effect (i.e., random intercept; using a "nuisance" prior equal to 1), and specified the estimation method as "laplace".[2] I then extracted and stored Bayes Factors for both main effects and the interaction, coding the Bayes Factor of the interaction relative to a null model assuming additive factors. Cases in which Laplace ANOVAs returned missing values were excluded from all analyses (about 20% of the cases when using the medium prior at relatively small sample sizes).

Bayesian ANOVA with Monte Carlo integration. I computed a Bayesian ANOVA, coding both main effects and the interaction term as fixed factors, subject as random effect (using a "nuisance" prior equal to 1), and relied on the default method of Monte Carlo sampling with 10,000 Monte Carlo iterations. To compensate for the variability of the sampling procedure (Pfister, 2021), I repeated each ANOVA 100 times per dataset and stored the median of the resulting Bayes Factors, accompanied by the range of the resulting Bayes Factors, and the ratio of the largest relative to the smallest Bayes Factor across re-runs (BF_{ratio}).

Simulation parameters

To cover a sufficiently large parameter space of the analyses, I simulated datasets for different sample sizes (n = 30 per group and n = 60 per group of the 2x2 mixed design), effect sizes (population-level δs corresponding to Cohen's *d* values of 0.0, 0.3, 0.5, and 0.8), as well as different priors of the Bayesian analyses (using the default values of *medium*, *wide*, and *ultrawide*

[2] Updated methodology with random slopes was not easily available at the time of conducting these simulations as it is only now being implemented in standard software packages (Oberauer, 2022; van den Bergh et al., 2022).

priors of the *ttestBF* and the *anovaBF* function). Several considerations guided these choices as described in the following.

Sample size (n = 30 and n = 60). Sample sizes were chosen in accordance with current standards in psychological experiments (e.g., Brysbaert, 2019) while at the same time allowing for a coverage of different power levels for between-subjects and within-subject comparisons alike. For the smaller sample size, the power of the between-subjects comparisons spans a range from about 20% to almost 90%, whereas the power of the within-subject comparisons spans a range from about 60% to > 99% (see Table 1). For the larger sample size, the power of the between-subjects comparisons spans a range from about 40% to almost > 99%, whereas the power of the within-subject comparisons spans a range from about 90% to > 99%. Because the following results were remarkably consistent across sample sizes, I chose to report only the results of the smaller sample size in the main text (considering that it spans a larger range of power levels) whereas I provide the results relating to the larger sample size in Appendix A.

Effect sizes (δ = 0.0, 0.3, 0.5, and 0.8). Including a situation without any effects was a natural starting point, given that the ability to quantify evidence for the absence of an effect is often considered a major strength of Bayesian approaches (e.g., Tendeiro & Kiers, 2019). The remaining effect sizes were chosen in accordance with Cohen's (1977) classification scheme with 0.5 and 0.8 indicating the lower boundaries of medium and large effects, respectively. For the smallest non-zero effect size, I settled on 0.3 as a compromise between the typically recommended lower boundary of a small effect (0.2) and the median effect size in psychological experiments, which is often reported to fall in the range of 0.3-0.4 (Open Science Collaboration, 2015; Stanley et al., 2018).

Effect sizes were implemented in an all-or-none fashion in the design to keep the results compact. That is, both main effects as well as the interaction effect were always set to the same value of either 0.0, 0.3, 0.5, or 0.8. To allow for power calculations, I further used the effect size variant that is tailored to each effect. For both between-subjects comparisons, i.e., for the main

effect of the between-subjects factor and for the interaction, I therefore used Cohen's d_s, whereas I used Cohen's d_z for the main effect of the within-subjects factor.[3]

Prior scale parameter (*medium*, *wide*, and *ultrawide*). I used the three suggested settings of the *ttestBF* and the *anovaBF* function to implement a reasonable range of scale parameters for the Bayesian prior (though this range is certainly non-exhaustive). In case of the Bayesian *t*-test, these values correspond to scale parameters of the Cauchy prior of $\sqrt{2}/2$, 1, and $\sqrt{2}$ for *medium*, *wide*, and *ultrawide* settings, respectively. For the Bayesian ANOVA, these values correspond to scale parameters for fixed-effects terms of 0.5, $\sqrt{2}/2$, and 1, for *medium*, *wide*, and *ultrawide* settings, respectively. To credit the fact that for actual analyses, priors should necessarily be defined a priori, I decided to simulate independent datasets for each prior setting rather than re-running Bayesian ANOVAs with different priors on the same dataset.

Results

Table 1 presents a manipulation check of the conducted simulations for group size n = 30. To this end, I computed the relative frequency of significant results of the frequentist ANOVA against the nominal α level of 0.05 in the case of an assumed population effect of 0, and against the expected power for all non-zero effects. The results closely align with the expectations based on the simulation parameters, indicating that the simulations worked as intended. The same held true at group sizes of n = 60 as shown in Table A1 in Appendix A.

Table 1. Results of the simulation check. I had sampled 100 datasets for each combination of effect size and assumed prior for the Bayesian analyses. Each dataset involved a 2x2 mixed design. To determine that the simulations had worked as intended, I calculated classical Analyses of Variance (ANOVAs) on each dataset and counted how many of the 100 ANOVAs returned a significant effect at α = 0.05. I compared this relative frequency against the expected power given

[3] Cohen's d_s refers to the difference between two means from independent samples, divided by the pooled standard deviation. Cohen's d_z refers to the mean difference for two dependent samples, divided by the standard deviation of the paired differences.

the effect size (at n = 30 per group). Note that the effect size applied to all three effects at the same time, so that, e.g., for an effect size of 0.3, the main effect of the between-subjects factor (ME Between), the main effect of the within-subject factor (ME Within), and the interaction were all sampled to come with an effect of 0.3. Effect sizes were modeled as Cohen's d_s for both between-subjects comparisons (ME Between, Interaction) and as Cohen's d_z for the within-subject comparison (ME Within).

	Power		Proportion of significant tests		
Effect size	(Between/Within)	subset	ME Between	ME Within	Interaction
0.0	N/A	all	0.04	0.06	0.05
		medium	0.05	0.07	0.08
		wide	0.05	0.03	0.01
		ultrawide	0.03	0.07	0.06
0.3	0.21 / 0.63	all	0.18	0.62	0.22
		medium	0.22	0.62	0.25
		wide	0.19	0.60	0.23
		ultrawide	0.14	0.64	0.17
0.5	0.48 / 0.97	all	0.45	0.96	0.50
		medium	0.46	0.94	0.51
		wide	0.46	0.96	0.49
		ultrawide	0.44	0.99	0.51
0.8	0.86 / >0.99	all	0.83	1.00	0.85
		medium	0.84	1.00	0.82
		wide	0.79	1.00	0.87
		ultrawide	0.86	1.00	0.86

Following the successful manipulation check, I probed for ordinal agreement of the different Bayesian methods. Ordinal agreement was assessed by computing pairwise Spearman rank correlations across datasets. As a reference value for such correlations I correlated the rank-ordered BF_{01} for Bayesian *t*-tests with the rank-ordered *p*-value of a classical (frequentist) ANOVA. These rank correlations were $\rho = 1.00$ for the main effect of the between factor as well as for the interaction, and $\rho > .94$ for the main effect of the within-subjects factor.

Table 2 shows the main results of the present simulations in terms of bivariate rank correlations (Spearmans ρ) for all three pairs of Bayesian approaches at group sizes of n = 30 (see Table A2 for the results for group sizes of n = 60). A first observation is that all approaches agreed almost perfectly in case of large effects in the population. For lower population effects many correlations still suggest sufficient reliability, particularly regarding the agreement of

Bayesian *t*-tests and Bayesian ANOVAs with Monte Carlo integration. Even here, however, several correlations are unexpectedly low (and lower than correlations between Bayesian *t*-tests and classical ANOVA), with one instance of $\rho < .900$. Bayesian ANOVAs with Laplace approximation, by contrast, did not agree consistently with the remaining two methods at lower effect sizes. A particularly strong outlier was present for the main effect of the between-subjects factor at $\delta = 0$, with a correlation of $\rho < .037$ with the rank-ordered results for Bayesian *t*-tests.

Table 2. Ordinal agreement of Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Each method was applied to all effects of a 2x2 mixed design, i.e., the main effect of the between-subjects term, the main effect of the within-subjects term, and the interaction effect. Correlations were computed across 100 simulated datasets with n = 30 participants per group and varying effect sizes and Bayesian priors as indicated in the table. Bayes Factors with MC sampling were computed as the median Bayes Factor of 100 iterations of a Bayesian ANOVA with 10,000 MC iterations.

			Spearman's ρ		
Effect size	Prior	Term	t | Laplace	t | MC	Laplace | MC
0.0	medium	ME Between	0.037	0.929	0.158
		ME Within	0.998	0.999	0.997
		Interaction	0.993	0.999	0.993
	wide	ME Between	0.653	0.882	0.890
		ME Within	0.998	0.999	0.997
		Interaction	0.999	0.998	0.997
	ultrawide	ME Between	0.861	0.932	0.981
		ME Within	0.998	0.999	0.998
		Interaction	1.000	0.998	0.998
0.3	medium	ME Between	0.652	0.979	0.705
		ME Within	1.000	1.000	1.000
		Interaction	0.998	0.998	0.998
	wide	ME Between	0.895	0.983	0.954
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	0.999
	ultrawide	ME Between	0.977	0.987	0.997
		ME Within	1.000	1.000	1.000
		Interaction	1.000	0.999	0.999
0.5	medium	ME Between	0.954	0.995	0.967
		ME Within	1.000	1.000	1.000
		Interaction	0.998	0.999	0.999
	wide	ME Between	0.990	0.996	0.997
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
	ultrawide	ME Between	0.997	0.997	1.000
		ME Within	1.000	0.999	1.000
		Interaction	1.000	1.000	1.000
0.8	medium	ME Between	1.000	0.999	0.999

	ME Within	1.000	1.000	1.000
	Interaction	1.000	1.000	1.000
wide	ME Between	0.999	0.999	1.000
	ME Within	0.999	0.999	1.000
	Interaction	0.999	1.000	1.000
ultrawide	ME Between	0.999	0.998	1.000
	ME Within	1.000	1.000	1.000
	Interaction	1.000	1.000	1.000

To get a better grasp on the lower correlations reported in Table 2, Figure 2 shows scatterplots for the first three rows of the table. This corresponds to a population effect of δ = 0, and a medium scale prior at a sample size of n = 30 per group. The plots suggest relatively high agreement at the upper end of the scale, i.e., for datasets clearly speaking towards the null hypothesis, whereas disagreement arises especially at the lower end of the scale, i.e., for datasets that lean towards the alternative hypothesis. This is at least the case for both main effects, whereas no clear pattern emerged for the interaction. Even though this pattern is quite remarkable, one might argue that the situation shown in Figure 2 is implausible for an actual empirical situation: If researchers were using a medium prior for their study, then they would have expected a medium or small effect. If expecting a relatively small effect, however, the study would likely have included more than 30 participants per group. I therefore re-ran the simulations for a more plausible parameter setting – a null effect in the population with a wide prior. As wide priors are the default setting of the algorithm's current implementation, this situation is more likely to occur in an actual study. Here, I simulated 10,000 datasets but reduced the number of re-iterations of the Monte Carlo integration to 9 to save computation time. All other settings were as for the main simulation. Figure 3 shows the resulting scatterplots, which replicated the main observations.

Table 3 shows metric agreement as quantified by the effect size d_z of a pairwise difference between two methods (at n = 30 per group; see Table A3 for corresponding data on n = 60 per group). Note that excessively high values of up to d_z = 23.51 result from small but highly systematic differences. Figure 4 shows corresponding descriptive statistics in terms of log-transformed Bayes Factors (here: BF_{10}). While all three methods converged for the interaction term, there were

marked differences for both main effects. Here, it appears as if Bayesian *t*-tests had systematically lower values for small effect sizes as compared to the remaining two methods, whereas this pattern reversed at larger effect sizes. Moreover, marked differences between the three methods emerged when assessing the percentage of correct decisions when using a $BF_{10} > 3$ as a cutoff for accepting the alternative hypothesis and $BF_{10} < 1/3$ as a cutoff for accepting the null hypothesis.[4] Table 4 summarizes these statistics for both group sizes, indicating that the Bayesian *t*-test outperformed the remaining two methods for null effects at groups of n = 30.

[4] I thank an anonymous reviewer for suggesting this analysis.

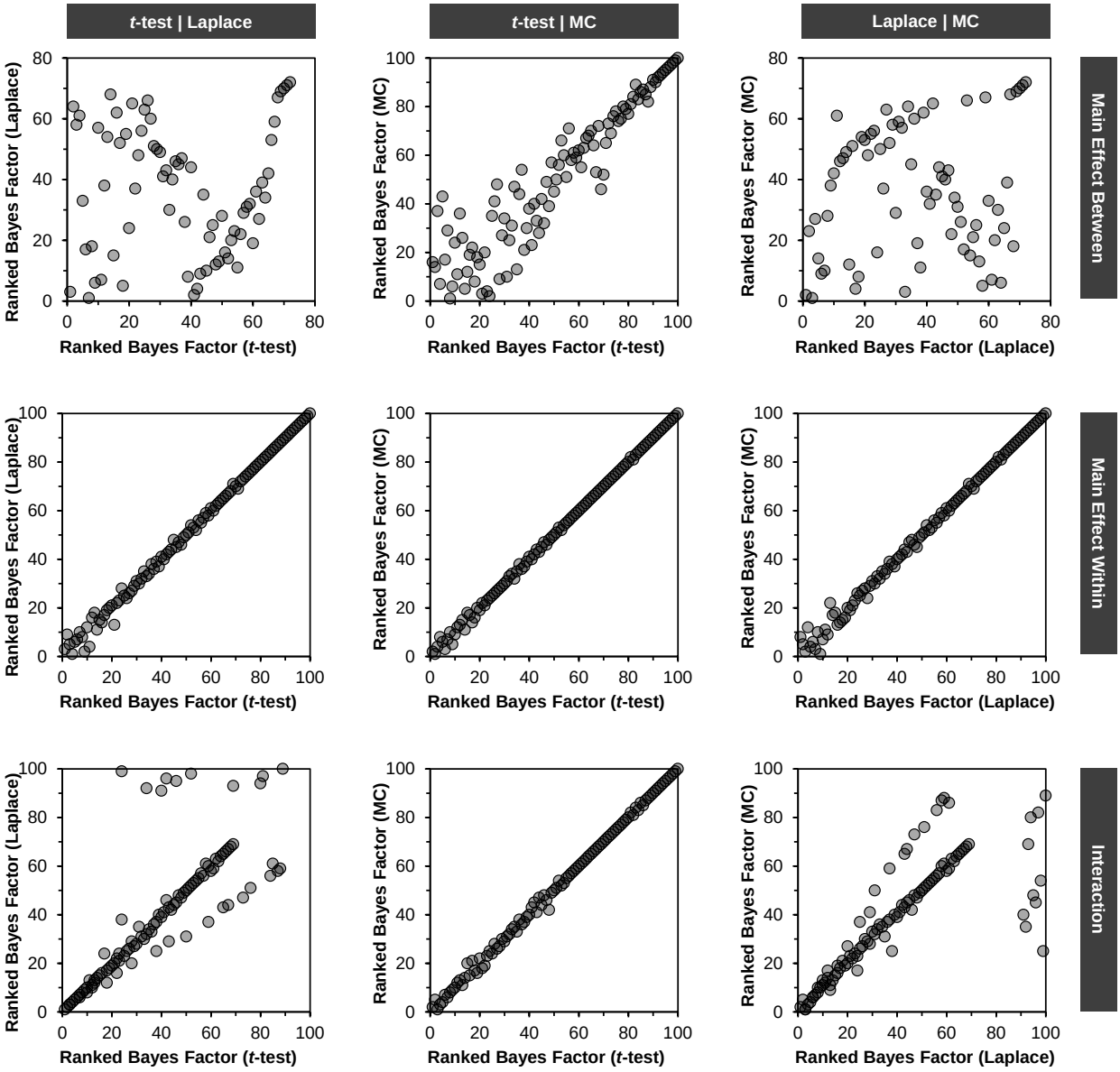

Figure 2. Ordinal agreement of ranked Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Scatterplots showcase the first three rows of Table 2 (see the corresponding caption for details). Each dataset was simulated with n = 30 participants per group. Note that the Laplace approximation could not be computed for a limited number these sets so that the corresponding scatterplots show fewer points than the scatterplots for *t*-tests and MC integration.

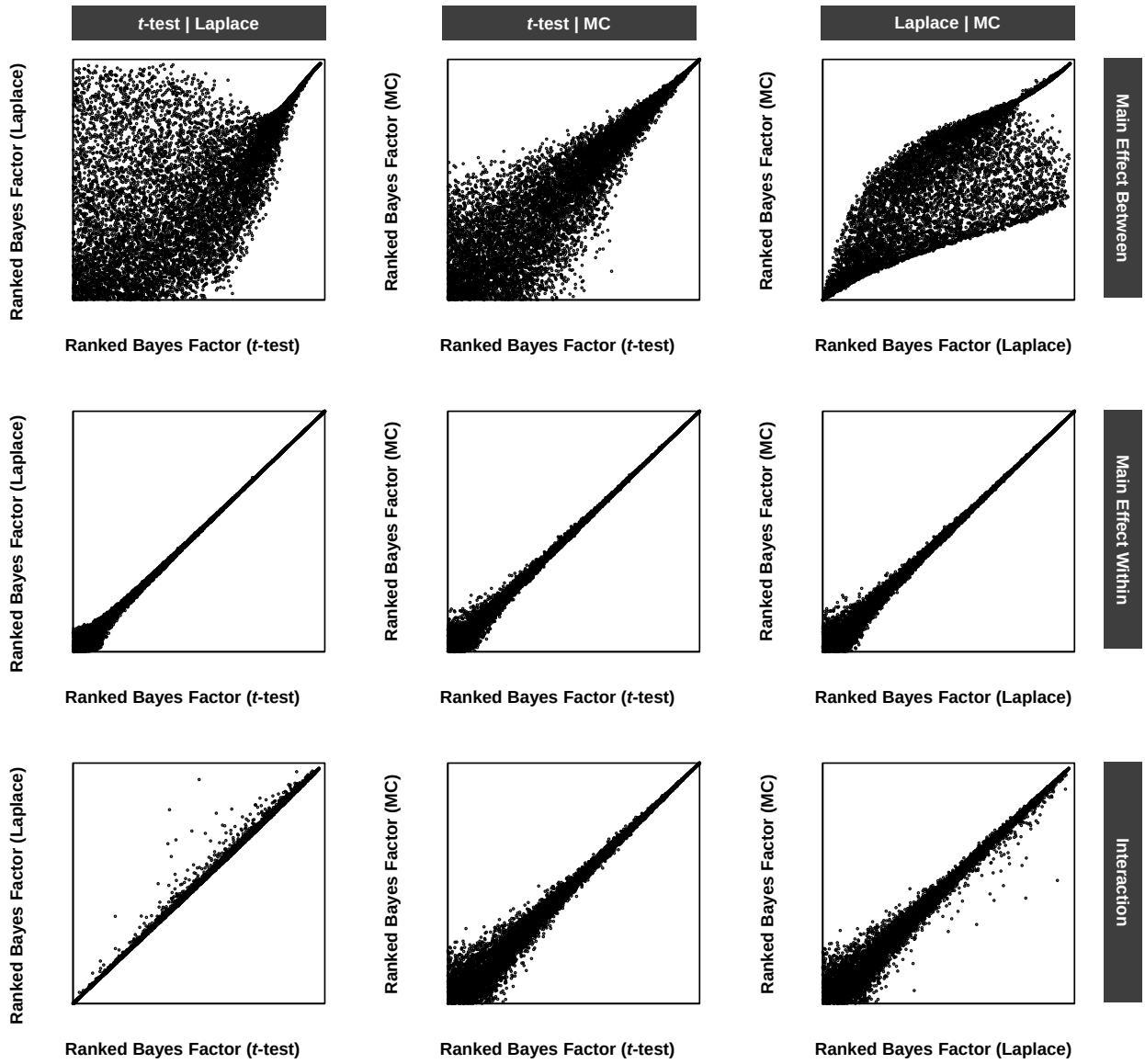

Figure 3. Ordinal agreement of ranked Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Scatterplots showcase the same settings as in Figure 2, but with a wide scale parameter and 10,000 instead of 100 datasets. Rank correlations correspond to ρ = .621, .915, and .845, in the upper row, ρ = .999, .998, and.997, in the middle row, and ρ = .999, .992, .991, in the lower row.

Table 3. Metric agreement of Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Each method was applied to all effects of a 2x2 mixed design, i.e., the main effect of the between-subjects term, the main effect of the within-subjects term, and the interaction effect. Bayes Factors were computed across 100 simulated datasets with n = 30 participants per group and varying effect sizes and Bayesian priors. Bayes Factors with MC integration were computed as the median Bayes Factor of 100 iterations of a Bayesian ANOVA, each with 10,000 MC iterations. Values represent standardized differences (Cohen's d_z) for pairwise differences of the resulting Bayes Factors of the three computational methods.

			ΔBF (d_z)		
Effect size	Prior	Term	t – Laplace	t – MC	Laplace – MC
0.0	medium	ME Between	-1.08	-3.08	0.68
		ME Within	-4.75	-6.76	-19.11
		Interaction	0.86	0.43	-0.77
	wide	ME Between	-1.34	-3.19	0.5
		ME Within	-10.59	-14.69	-16.73
		Interaction	2.55	0.96	-2.52
	ultrawide	ME Between	-2.27	-3.61	-0.15
		ME Within	-8.02	-10.71	-19.47
		Interaction	5.35	0.64	-7.18
0.3	medium	ME Between	-0.45	-1.83	0.2
		ME Within	-0.4	-1.12	-18.68
		Interaction	0.26	0.7	-0.19
	wide	ME Between	-1.12	-2.7	0.16
		ME Within	-1.14	-1.98	-18.49
		Interaction	0.78	0.47	-0.73
	ultrawide	ME Between	-2.26	-3.6	-0.95
		ME Within	-2.06	-3.21	-23.51
		Interaction	3.88	0.39	-6.58
0.5	medium	ME Between	-0.1	-1.41	-0.3
		ME Within	0.71	0.13	-18.46
		Interaction	0.76	1.2	-0.52
	wide	ME Between	-0.7	-2.07	-0.78
		ME Within	0.25	-0.49	-18.13
		Interaction	1.86	1.13	-1.76
	ultrawide	ME Between	-1.37	-2.51	-3.11
		ME Within	-0.46	-1.45	-19.25
		Interaction	2.5	0.21	-4.8
0.8	medium	ME Between	1.75	-0.4	-2.58
		ME Within	2.51	1.89	-19.69
		Interaction	2.23	1.81	-1.93
	wide	ME Between	0.36	-0.57	-3.04
		ME Within	1.67	1.11	-18.92
		Interaction	1.97	1.4	-2.61
	ultrawide	ME Between	-0.03	-1.02	-9.02
		ME Within	1.38	0.74	-19.19
		Interaction	2.64	1.29	-6.91

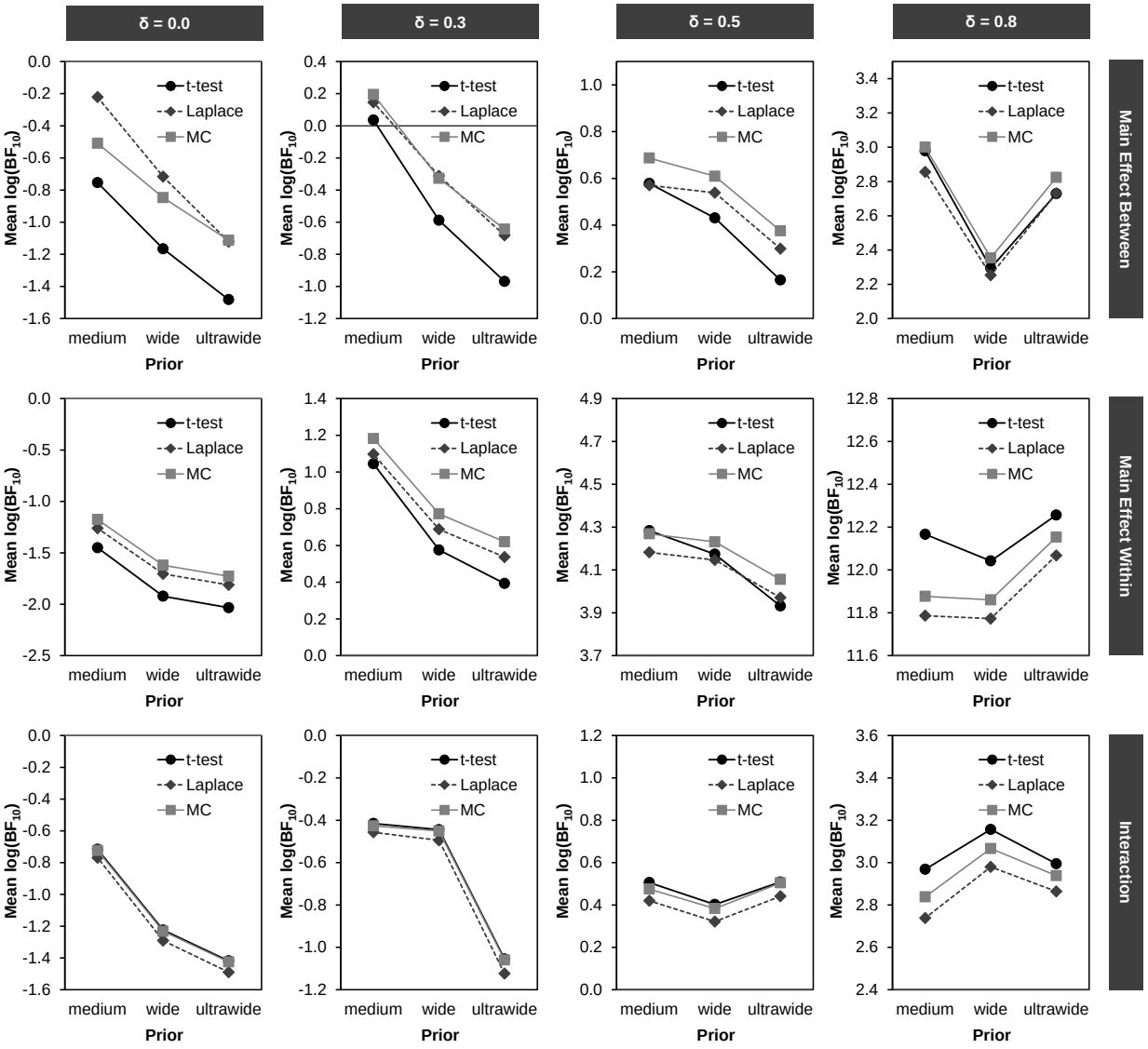

Figure 4. Metric agreement of Bayes Factors (BF_{10}) as computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Y axes are scaled to maximize readability of the individual patterns of means.

Table 4. Percentage of correct decisions against the ground truth of the simulated effect sizes. A decision was counted as correct with $BF_{10} > 3$ for non-zero effect sizes and with $BF_{10} < 1/3$ for an effect size of zero. Percentages were computed across all scale parameters and all three terms (main effect between, main effect within, interaction).

		% correct		
n	Effect size	t	MC	Laplace
30	0	74	63	60
	0.3	18	19	18
	0.5	49	50	48
	0.8	81	81	80
60	0	86	80	74
	0.3	35	37	36
	0.5	71	72	71
	0.8	97	97	97

Finally, I aimed at re-assessing previous reports of substantial variability for Bayes Factors as computed via Monte Carlo integration (Pfister, 2021). I therefore extracted the ratio of the highest and the lowest Bayes Factor for each dataset of the simulation study. Table 5 shows mean ratios as well as the maximum ratio for the 100 datasets in each design cell (at group sizes of n = 30; see Table A4 for group sizes of n = 60). I further assessed the relative frequency of datasets for which the largest and the smallest Bayes Factor fell on the same side of the neutral value of 1. The results replicate previous observations in terms of substantial variability. This was particularly evident for the interaction term, whereas agreement was higher for both main effects.

Table 5. Variability of Bayes Factors computed via Bayesian Analysis of Variance (ANOVA) with Monte Carlo (MC) integration. A sample of 100 datasets was simulated for each combination of effect sizes and assumed priors (n = 30 per group), and for each of these datasets I re-ran the same Bayesian ANOVA 100 times. BF_{ratio} denotes the ratio of the largest relative to the smallest Bayes Factor resulting from these 100 iterations, whereas $\%_{sameside}$ denotes the percentage of datasets for which the largest and smallest Bayes Factor were located on the same side of the neutral value of 1.

			Measure of consistency		
Effect size	Prior	Term	Mean BF_{ratio}	Maximum BF_{ratio}	$\%_{sameside}$
0.0	medium	ME Between	1.32	7.87	93
		ME Within	1.57	7.49	97
		Interaction	2.76	28.09	80
	wide	ME Between	1.23	4.16	89
		ME Within	1.48	4.36	99
		Interaction	2.85	22.00	87
	ultrawide	ME Between	1.35	3.79	96
		ME Within	1.70	10.37	94
		Interaction	5.01	167.05	91
0.3	medium	ME Between	1.77	34.63	89
		ME Within	1.89	33.66	88
		Interaction	3.39	68.76	77
	wide	ME Between	1.26	3.13	93
		ME Within	1.50	11.72	97
		Interaction	3.49	27.54	78
	ultrawide	ME Between	1.51	13.32	97
		ME Within	3.37	174.44	93
		Interaction	3.36	26.57	89
0.5	medium	ME Between	1.40	3.68	95
		ME Within	1.40	3.85	100
		Interaction	4.59	52.77	73
	wide	ME Between	1.53	11.69	93
		ME Within	2.75	133.86	93
		Interaction	3.84	39.09	69
	ultrawide	ME Between	1.37	3.44	93
		ME Within	2.47	100.20	98
		Interaction	4.05	50.00	81
0.8	medium	ME Between	1.55	9.83	95
		ME Within	1.54	13.47	100
		Interaction	3.88	64.74	84
	wide	ME Between	1.81	23.85	93
		ME Within	1.45	5.00	100
		Interaction	8.00	362.70	92
	ultrawide	ME Between	1.56	17.17	96
		ME Within	1.83	39.42	100
		Interaction	6.38	206.31	88

Conclusions

The present simulations aimed at assessing the agreement – i.e., consistency – of different Bayesian methods for analyzing factorial designs. While Bayesian *t*-tests and Bayesian ANOVA with Monte Carlo integration showed good to excellent ordinal agreement for most (but not all) datasets, there were striking differences to Bayesian ANOVA with Laplace approximation. This was true for different priors, sample sizes and also emerged for large number of simulated studies. Note that ordinal agreement relates to a rank order of how strongly a simulated dataset allegedly supported a hypothesis of a main or interaction effect. There were also notable differences in metric agreement in that Bayesian *t*-tests compared to their ANOVA counterparts were more prone to support the null hypothesis of no effect at low effect sizes or with non-existent population effects. Finally, the present simulations also replicated previous observations suggesting that Bayesian ANOVA with Monte Carlo integration has limited agreement with itself (Pfister, 2021). This included several cases in which smallest and largest Bayes Factor across 100 iterations fell on different sides of the neutral value of 1, i.e., pointing towards evidence for the null hypothesis of no effect on one iteration while pointing towards evidence for the alternative hypothesis on another iteration. Surprisingly, such behavior occurred even at large population effect sizes, albeit less frequently as compared to smaller effects. Reporting median Bayes Factors across multiple runs thus appears to be important to safeguard against possible outliers introduced by Monte Carlo integration.

Limits of the present study concern the use of simulated data as well as restricted parameter sets for the simulations. To establish that the observed divergence of Bayesian methods also occurs for empirical data, I therefore re-analyzed an existing dataset (Pfister et al., 2016). This analysis is reported in Appendix B and replicates the main observations of the simulation approach. Restrictions on the parameter set of simulations concern the use of a single factorial design, i.e., a 2x2 mixed design. Straightforward extensions to pure within-subject or pure between-subjects designs as well as factors with three or more levels are easily possible with the

code underlying the reported results (https://osf.io/4u9sx/). Additional extensions relate to correlations of within-subject data, with a constant correlation of about r = .7 throughout the reported results. The consistency of the present data across re-runs with different sample sizes and different numbers of datasets, however, suggest that the main observations outlined above may indeed generalize.

Researchers are thus well advised not to rely exclusively on any particular method in the context of Bayesian ANOVA. Validating any empirical result by means of alternative approaches can help to establish trust in a given conclusion. Especially when probing for null effects with small sample sizes, the present simulations indicate Bayesian *t*-tests to outperform both types of Bayesian ANOVA so that this methodology can be employed to increase trust in the outcome of a Bayesian ANOVA. For non-zero population effects, all three methods produced a similar percentage of correct decisions, despite marked disagreement on which datasets actually represented evidence in favor of the alternative hypothesis. This state of affairs makes cross-validation with different techniques particularly relevant. In addition to the approaches discussed here, there are conceptually similar possibilities via Bayesian linear models (Oberauer, 2022) or Bayesian estimation (Rouder et al., 2018; Tendeiro & Kiers, 2022b). Likewise, using insights from non-Bayesian methodology may help to establish converging evidence for what to take from any empirical dataset.

Appendix A: Results for group size n = 60

Results for the datasets with n = 60 per group replicated the observations at group sizes of n = 30 as shown in the main text. This concerned the manipulation check as shown in Table A1, the main simulation results in terms of ordinal agreement as shown in Table A2, and metric agreement as shown in Table A3. Variability of the Monte Carlo method was again particularly pronounced for the interaction term as shown in Table A4.

Table A1. Results of the simulation check. I had sampled 100 datasets for each combination of effect size and assumed prior for the Bayesian analyses. Each dataset involved a 2x2 mixed design. To determine that the simulations had worked as intended, I calculated classical analyses of variance (ANOVAs) on each dataset and counted how many of the 100 ANOVAs returned a significant effect at α = 0.05. I compared this relative frequency against the expected power given the effect size (at n = 60 per group). Note that the effect size applied to all three effects at the same time, so that, e.g., for an effect size of 0.3, the main effect of the between-subjects factor (ME Between), the main effect of the within-subjects factor (ME Within), and the interaction were all sampled to come with an effect of 0.3. Effect sizes were modeled as Cohen's d_s for both between-subjects comparisons (ME Between, Interaction) and as Cohen's d_z for the within-subjects comparison (ME Within).

	Power		Proportion of significant tests		
Effect size	(Between/Within)	subset	ME Between	ME Within	Interaction
0.0	N/A	all	0.03	0.07	0.04
		medium	0.03	0.02	0.03
		wide	0.03	0.11	0.04
		ultrawide	0.03	0.07	0.05
0.3	0.37 / 0.90	all	0.38	0.87	0.37
		medium	0.43	0.90	0.39
		wide	0.34	0.91	0.38
		ultrawide	0.38	0.81	0.35
0.5	0.78 / >0.99	all	0.76	1.00	0.77
		medium	0.71	1.00	0.81
		wide	0.75	1.00	0.77
		ultrawide	0.81	1.00	0.73
0.8	0.99 / >0.99	all	0.99	1.00	1.00
		medium	0.99	1.00	0.99
		wide	1.00	1.00	1.00
		ultrawide	0.98	1.00	1.00

Table A2. Ordinal agreement of Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Each method was applied to all effects of a 2x2 mixed design, i.e., the main effect of the between-subjects term, the main effect of the within-subjects term, and the interaction effect. Correlations were computed across 100 simulated datasets with n = 60 participants per group and varying effect sizes and Bayesian priors as indicated in the table. Bayes Factors with MC sampling were computed as the median Bayes Factor of 100 iterations of a Bayesian ANOVA with 10,000 MC iterations.

			Spearman's ρ		
Effect size	Prior	Term	t | Laplace	t | MC	Laplace | MC
0.0	medium	ME Between	0.222	0.947	0.336
		ME Within	1.000	0.999	0.998
		Interaction	0.998	0.996	0.996
	wide	ME Between	0.829	0.936	0.964
		ME Within	1.000	1.000	1.000
		Interaction	1.000	0.998	0.998
	ultrawide	ME Between	0.905	0.928	0.997
		ME Within	1.000	1.000	0.999
		Interaction	1.000	0.996	0.996
0.3	medium	ME Between	0.939	0.996	0.952
		ME Within	1.000	1.000	1.000
		Interaction	0.999	1.000	0.998
	wide	ME Between	0.992	0.996	0.999
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
	ultrawide	ME Between	0.995	0.997	0.999
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
0.5	medium	ME Between	0.998	1.000	0.999
		ME Within	1.000	1.000	1.000
		Interaction	0.997	1.000	0.998
	wide	ME Between	0.999	1.000	1.000
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
	ultrawide	ME Between	0.999	0.999	1.000
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
0.8	medium	ME Between	1.000	1.000	1.000
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
	wide	ME Between	1.000	1.000	1.000
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000
	ultrawide	ME Between	1.000	1.000	1.000
		ME Within	1.000	1.000	1.000
		Interaction	1.000	1.000	1.000

Table A3. Metric agreement of Bayes Factors computed via Bayesian *t*-tests, Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. Each method was applied to all effects of a 2x2 mixed design, i.e., the main effect of the between-subjects term, the main effect of the within-subjects term, and the interaction effect. Bayes Factors were computed across 100 simulated datasets with n = 60 participants per group and varying effect sizes and Bayesian priors. Bayes Factors with MC sampling were computed as the median Bayes Factor of 100 iterations of a Bayesian ANOVA, each with 10,000 MC iterations. Values represent standardized differences (Cohen's d_z) for pairwise differences of the resulting Bayes Factors of the three computational methods.

			ΔBF (d_z)		
Effect size	Prior	Term	t – Laplace	t – MC	Laplace – MC
0.0	medium	ME Between	-1.29	-4.59	0.81
		ME Within	-11.34	-15.11	-28.34
		Interaction	2.07	1.22	-1.71
	wide	ME Between	-2.96	-5.23	0.93
		ME Within	-11.58	-15.27	-24.14
		Interaction	11.11	1.44	-8.34
	ultrawide	ME Between	-3.82	-5.35	-1.36
		ME Within	-21.15	-27.85	-26.83
		Interaction	15.81	1.24	-9.26
0.3	medium	ME Between	-0.50	-1.96	0.06
		ME Within	-0.72	-1.62	-25.30
		Interaction	0.96	0.94	-0.78
	wide	ME Between	-1.91	-3.82	-0.32
		ME Within	-2.28	-3.57	-25.86
		Interaction	2.65	0.78	-3.11
	ultrawide	ME Between	-2.88	-4.18	-2.65
		ME Within	-3.08	-4.43	-24.39
		Interaction	4.97	0.33	-9.87
0.5	medium	ME Between	-0.06	-1.45	-0.71
		ME Within	1.23	0.41	-28.20
		Interaction	1.01	1.85	-0.67
	wide	ME Between	-1.15	-2.96	-2.42
		ME Within	0.41	-0.50	-26.06
		Interaction	4.61	1.70	-5.90
	ultrawide	ME Between	-2.42	-4.00	-6.87
		ME Within	-0.36	-1.38	-27.55
		Interaction	5.71	0.80	-9.34
0.8	medium	ME Between	4.34	-1.80	-8.66
		ME Within	5.14	4.14	-30.44
		Interaction	3.19	2.33	-4.91
	wide	ME Between	0.65	-1.63	-18.57
		ME Within	3.76	2.85	-28.48
		Interaction	3.98	2.57	-7.88
	ultrawide	ME Between	-0.15	-1.61	-22.09
		ME Within	2.59	1.63	-31.52
		Interaction	5.35	2.56	-10.29

Table A4. Variability of Bayes Factors computed via Bayesian Analysis of Variance (ANOVA) with Monte Carlo (MC) integration. A sample of 100 datasets was simulated for each combination of effect sizes and assumed priors (n = 60 per group), and for each of these datasets I re-ran the same Bayesian ANOVA 100 times. BF_{ratio} denotes the ratio of the largest relative to the smallest Bayes Factor resulting from these 100 iterations, whereas $\%_{sameside}$ denotes the percentage of datasets for which the largest and smallest Bayes Factor were located on the same side of the neutral value of 1.

			Measure of consistency		
Effect size	Prior	Term	Mean BF_{ratio}	Maximum BF_{ratio}	$\%_{sameside}$
0.0	medium	ME Between	1.19	4.95	93
		ME Within	1.82	32.03	97
		Interaction	2.47	27.50	93
	wide	ME Between	1.33	6.76	96
		ME Within	1.79	25.22	95
		Interaction	3.31	31.23	95
	ultrawide	ME Between	1.60	12.06	96
		ME Within	2.82	103.12	94
		Interaction	3.57	28.76	92
0.3	medium	ME Between	1.36	4.75	95
		ME Within	1.53	9.02	99
		Interaction	3.93	44.38	82
	wide	ME Between	1.39	6.84	92
		ME Within	1.79	21.75	97
		Interaction	3.51	25.84	74
	ultrawide	ME Between	1.31	4.06	98
		ME Within	1.48	6.84	97
		Interaction	4.69	32.62	77
0.5	medium	ME Between	4.25	268.78	97
		ME Within	2.85	82.98	100
		Interaction	4.06	79.00	84
	wide	ME Between	1.56	8.04	97
		ME Within	1.56	9.41	100
		Interaction	3.41	12.45	80
	ultrawide	ME Between	1.47	7.45	96
		ME Within	1.49	8.10	100
		Interaction	4.20	32.43	89
0.8	medium	ME Between	1.38	2.99	100
		ME Within	1.56	7.45	100
		Interaction	4.41	48.09	97
	wide	ME Between	1.55	6.55	100
		ME Within	1.87	21.61	100
		Interaction	9.58	381.99	97
	ultrawide	ME Between	1.60	13.49	100
		ME Within	1.95	38.15	100
		Interaction	4.49	36.91	95

Appendix B: Extension to empirical data

To establish that the present simulation results are representative also for real-word data, I re-analyzed the data of Pfister et al. (2016) with Bayesian *t*-tests, Bayesian ANOVA with Laplace approximation and Bayesian ANOVA with Monte Carlo integration. I chose this dataset because it features a similar 2x2 mixed design as the simulation study while at the same time providing multiple dependent variables. In this study, participants performed a point-and-click task with the computer mouse so that we were able to analyze initiation time (IT), movement time (MT), maximum absolute distance (MAD) to a reference line, as well as area under the curve (AUC) for each movement. Table A5 shows the resulting Bayes Factors for each of these measures when conducting separate 2x2 mixed ANOVAs with the same settings as in the simulations, with the r scale parameter set to medium (also using 100 iterations for the ANOVA with Monte Carlo integration). Results show that Bayes Factors diverge by a factor of up to 2 for several analyses, suggesting limited agreement of the three methods. Table A6 further shows substantial variability of the Bayes Factor results across iterations of the Monte Carlo approach. This includes two cases in which smallest and largest Bayes Factor are on different sides of the neutral value of 1, as well as Bayes Factor ratios larger than 4 for the interaction term for several measures.

Table A5. Agreement of Bayes Factors computed via Bayesian t-tests (t), Bayesian Analysis of Variance (ANOVA) with Laplace approximation, and Bayesian ANOVA with Monte Carlo (MC) integration. The underlying dataset is a mouse-tracking study (Pfister et al., 2016) with a 2x2 mixed design that yielded the measures of initiation time (IT), movement time (MT), maximum absolute distance (MAD), and area under the curve (AUC). Numbers indicate BF_{10} for the three terms of the design.

Measure	Term	t	Laplace	MC
IT	ME Between	0.70	0.68	0.81
	ME Within	31488.35	18043.54	20034.25
	Interaction	799.95	384.78	558.78
MT	ME Between	0.46	0.62	0.65
	ME Within	39.21	32.31	35.26
	Interaction	1.39	1.19	1.36
MAD	ME Between	3.89	3.28	4.09

	ME Within	106.83	83.51	93.09
	Interaction	85.93	55.74	68.95
AUC	ME Between	1.58	1.29	1.75
	ME Within	74.97	59.03	65.63
	Interaction	37.27	24.82	31.61

Table A6. Variability of Bayes Factors computed via Bayesian Analysis of Variance (ANOVA) with Monte Carlo (MC) integration, as per the analysis shown in Table A5. I re-ran the same Bayesian ANOVA 100 times for each measure, i.e., initiation time (IT), movement time (MT), maximum absolute distance (MAD) and area under the curve (AUC). BF_{min} and BF_{max} show the smallest and largest resulting Bayes Factors for each design cell, whereas BF_{ratio} denotes the ratio of the largest relative to the smallest Bayes Factor.

Measure	Term	BF_{min}	BF_{max}	BF_{ratio}
IT	ME Between	0.78	1.03	1.33
	ME Within	19349.00	24644.01	1.27
	Interaction	307.10	1417.35	4.62
MT	ME Between	0.64	0.71	1.12
	ME Within	34.24	42.71	1.25
	Interaction	0.94	4.20	4.49
MAD	ME Between	3.91	4.92	1.26
	ME Within	89.17	137.15	1.54
	Interaction	50.11	125.07	2.50
AUC	ME Between	1.66	2.59	1.57
	ME Within	63.15	104.11	1.65
	Interaction	14.96	67.26	4.50

References

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*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current practices in data analysis procedures in psychology: What has changed? *Frontiers in Psychology*, *9*, 2558. https://doi.org/10.3389/fpsyg.2018.02558

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in Canadian journal articles in psychology. *Canadian Psychology / Psychologie canadienne*, *58*, 140–147. https://doi.org/10.1037/cap0000074

Edgington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist*, *19*, 202–203. https://doi.org/10.1037/h0039177

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: Generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic Bulletin & Review*, *19*, 395-404. https://doi.org/10.3758/s13423-012-0230-1

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Erlbaum.

JASP Team (2022). *JASP*. Retrieved from https://jasp-stats.org/

Long, Q., Scavino, M., Tempone, R., & Wang, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, *259*, 24-39. https://doi.org/10.1016/j.cma.2013.02.017

Morey, R. D., & Rouder, J. N. (2018). *Package 'BayesFactor'* (version 0.9.12-4.2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morey, R. D., & Rouder, J. N. (2022). *BayesFactor: Computation of Bayes Factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. https://doi.org/10.1126/science.aac4716

Pfister, R. (2021). Variability of Bayes Factor estimates in Bayesian analysis of variance. *The Quantitative Methods for Psychology*, *17*, 40-45. https://doi.org/10.20982/tqmp.17.1.p042

Pfister, R., Wirth, R., Schwarz, K., Steinhauser, M., & Kunde, W. (2016). Burdens of non-conformity: Motor execution reveals cognitive conflict during deliberate rule violations. *Cognition*, *147*, 93-99. https://doi.org/10.1016/j.cognition.2015.11.009

Rouder, J.N., Haaf, J.M., & Vandekerckhove, J (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113. https://doi.org/10.3758/s13423-017-1420-7

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

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*, 304–399. https://doi.org/10.1037/met0000057

Ryan, E. G., Drovandi, C. C., McGree, J. M., Pettitt, A. N. (2016). A review of modern computational algorithms for Bayesian optimal design. *International Statistical Review*, *84*, 128–154. https://doi.org/10.1111/insr.12107

Schillings, C., Sprungk, B., & Wacker, P. (2020). On the convergence of the Laplace approximation and noiselevel-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. *Numerische Mathematik*, *145*(4), 915–971. https://doi.org/10.1007/s00211-020-01131-1

Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*, *70*, 777-795. https://doi.org/10.1177/0013164410379320

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346. https://doi.org/10.1037/bul0000169

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

Tendeiro, J. N., & Kiers, H. A. L. (2022a). On the white, the black, and the many shades of gray in between: Our reply to Van Ravenzwaaij and Wagenmakers (2021). *Psychological Methods*, *27*, 466–475. https://doig.org/10.1037/met0000505

Tendeiro, J. N., & Kiers, H. A. (2022b). With Bayesian estimation one can get all that Bayes factors offer, and more. *Psychonomic Bulletin & Review.* Manuscript in press, doi: 10.3758/s13423-022-02164-3.

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: the last 25 years. *Psychological Methods*, *22*, 217–239. doi: 10.1037/met0000100

van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E. J., Derks, K., ... & Wagenmakers, E. J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Annee psychologique*, *120*, 73-96. https://doi.org/10.3917/anpsy1.201.0073

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. http://dx.doi.org/10.3758/BF03194105

Zhou, Y., & Skidmore, S. T. (2017). A reassessment of ANOVA reporting practices: A review of three APA journals. *Journal of Methods and Measurement in the Social Sciences*, *8*, 3-19. https://doi.org/10.2458/v8i1.22019