Journal of Mathematical Psychology 122 (2024) 102873

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Theoretical Note

A straightforward and valid correction to Nathoo et al.'s Bayesian within-subject credible interval

Steffen Zitzmann [a,*], Christoph Lindner [b], Martin Hecht [c]

[a] Medical School Hamburg, Germany
[b] Max Planck Institute of Psychiatry, Germany
[c] Helmut Schmidt University, Germany

ARTICLE INFO

Keywords:
Within-subject design
Random effects
Bayes
Credible interval
Degrees of freedom

ABSTRACT

The APA encourages authors to thoroughly report their results, including confidence intervals. However, considerable debate exists regarding the computation of confidence intervals in within-subject designs. Nathoo et al.'s (2018) recently proposed a Bayesian within-subject credible interval, which has faced criticism for not accounting for the uncertainty associated with estimating subject-specific effects. In this article, we show how Nathoo et al.'s within-subject credible interval can be easily corrected by utilizing the theory of degrees of freedom. This correction obviates the necessity for estimates of subject-specific effects that offer shrinkage. Instead, it involves a straightforward adjustment in degrees of freedom in both the interaction mean squares and the t-distribution used to compute the interval. Therefore, our proposed interval, being easily computable through a simple formula, eliminates the need for fully Bayesian approaches. It accurately represents uncertainty and offers the interpretational benefit of Bayesian intervals.

The American Psychological Association (APA) recommends that authors report the results of their statistical analysis in detail, including confidence intervals (American Psychological Association, 2019). These intervals serve to communicate uncertainty and helps decide whether the values within the intervals are substantial enough to be practically significant (e.g., Amrhein & Greenland, 2022; see also Zitzmann, Nagengast, Hübner, & Hecht, 2024). Confidence intervals, computable for any statistic, are particularly crucial in experimental psychology, where the most important statistics are level means and their differences. Whereas confidence intervals for means are undisputed in between-subject designs, where each person is exposed to only one level of the factor, there is a hot debate as to how confidence intervals should be computed when the same person is exposed to all levels (within-subject designs). This type of design, preferred for its higher statistical power and cost-effectiveness compared to a between-subject design, presents a challenge: standard confidence intervals may overlap in samples of typical size, potentially obscuring differences between levels, particularly when subjects differ to a large degree in their means across the levels. To address this issue, Loftus and Masson (1994) proposed a frequentist within-subject confidence interval. This interval is not a confidence interval in the usual sense (Loftus & Masson, 1994), but it works as a way to make significant differences between levels visible. The idea behind this interval is to minimize the impact of between-subject variability on the interval. Since Loftus and Masson (1994) introduced their method, there has been increasing interest in refining this interval, leading to various alternative calculation methods being proposed by scholars such as Baguley (2012), Cousineau (2005), or Hollands and Jarmasz (2010), among others. Whereas Loftus and Masson's original procedure is considered complex, subsequent methods like the one proposed by Cousineau (2005) have simplified the calculation by using a normalizing procedure, which minimizes differences between subjects in the data. At the heart of this procedure lies a centering approach, which consists of subtracting each subject's data points by the subject's mean. Morey criticized this interval as tending to be too narrow and proposed a simple correction to address this issue. Notably, Morey's corrected interval closely aligns with Loftus and Masson's within-subject confidence interval, except that it relaxes the sphericity assumption. More recent techniques, such as those by Cousineau (2019) and Tryon (2001), employ decorrelation strategies to preserve the uncorrelated components of the data, resulting in intervals similar to previous ones when the compound symmetry assumption holds.

In recent years, it has repeatedly been claimed that Bayesian credible intervals should be favored for their interpretational benefit. These intervals indicate the most likely values of parameters given the data, and this is more intuitive than what can be learned from frequentist confidence intervals (Hoekstra, Morey, Rouder, & Wagenmakers,

[*] Correspondence to: Medical School Hamburg, 20457 Hamburg, Germany.
E-mail address: steffen.zitzmann@medicalschool-hamburg.de (S. Zitzmann).

https://doi.org/10.1016/j.jmp.2024.102873
Received 23 February 2024; Received in revised form 21 June 2024; Accepted 8 July 2024
Available online 22 July 2024

Table 1
Motivating example.

Subject	Exposure duration		
	1 s	2 s	5 s
1	10	13	13
2	6	8	8
3	11	14	14
4	22	23	25
5	16	18	20
6	15	17	17
7	1	1	4
8	12	15	17
7	1	1	4
8	12	15	17
9	9	12	12
10	8	9	12

2014). Therefore, Nathoo, Kilshaw, and Masson (2018) proposed a Bayesian within-subject credible interval. Technically, this interval conditions on Maximum Likelihood (ML) estimates of subject-specific effects. However, Heck (2019) criticized the interval by pointing out that it treats the ML estimates as known values, thereby preventing the incorporation of the uncertainty in these estimates into the interval (see also Nathoo et al., 2018, who acknowledged this limitation themselves). Furthermore, he argued that ML estimates would not offer shrinkage, meaning they are not adjusted towards zero. As a consequence, a too large amount of between-subject variability would be removed, resulting in an understatement of uncertainty by the interval. To address these supposed shortcomings, Heck (2019) proposed a stepwise procedure that emulates a fully Bayesian approach. This procedure enables that uncertainty in estimates of subject-specific effects is included in the interval.

We welcome the renewed interest in within-subject intervals and commend Nathoo et al. (2018) for their innovative proposal, which marks an advancement in the field. Also, we appreciate Heck's critique. It prompted us to think again about Nathoo et al.'s within-subject interval, leading us to suggest an even more straightforward correction. Whereas we agree with Heck's primary argument, we question that estimates of subject-specific effects must be shrunken estimates. Specifically, we question the underlying assumption at the technical level (i.e., the level of implementation) that these effects stem from a common distribution or, in other words, that these effects are random effects. We wish to clarify that the concept of random effects is multifaceted. While not delving deeply into this complex topic, we propose to distinguish two different notions of the term "random effects". The first one has a sampling-theoretical interpretation, where the units that exhibit the effects are randomly drawn in accordance with a sampling design. One may refer to this interpretation as random by design. This sampling-theoretical interpretation may involve a distributional assumption at the conceptual level. The second notion represents a more technical aspect and relates to the distribution of subject-specific effects. The assumption that random effects stem from of a parametric distribution is very common in the field as indicated by various articles in this journal that have used GLMMs, JAGS, or Stan to model differences between subjects. We will argue and show that this latter "random-effects assumption" may not be necessary for deriving a correct within-subject credible interval. Based on our conclusion, we will justify a simple way to correct Nathoo et al.'s within-subject interval in such a way that the uncertainty in the estimates of subject-specific effects is accounted for. The proposed correction involves an adjustment in degrees of freedom in both the interaction mean squares and the t-distribution used to compute the interval. We will demonstrate that when uncertainty in the ML estimates is incorporated in this manner, the resulting Bayesian credible interval aligns with the frequentist counterpart as well as with Heck's fully Bayesian interval, which has been noted for its numerical similarity to the frequentist one.

1. Nathoo et al.'s Bayesian within-subject credible interval

Our discussion centers on the most basic within-subject design, featuring only one within-subject factor. This design has a sample size of N, and the number of factor levels is C. An example is presented by Table 1, which showcases the data used by Loftus and Masson (1994) in their seminal work and by various authors who have since expanded, criticized, or contributed to the debate about within-subject intervals. At the implementation level, the model for this design can be expressed as follows. The response of the ith subject under the jth level is:

$$Y_{ij} = \mu_j + b_i + \varepsilon_{ij} \tag{1}$$

where μ_j represents the mean across the responses under the jth level. The b_is stand for subject-specific effects. As they model differences between subjects, they are almost always of no substantive interest and thus typically regarded as nuisance parameters (Loftus & Masson, 1994). The ε_{ij}s are the residuals. Notice that contrary to common practice, we do not make what we have referred to as the random-effects assumption for reasons that we will explain later on, meaning that technically, we do not assume that subject-specific effects stem from a distribution. That is, we do not implement such a distribution. In fact, in line with other scholars, Nathoo et al. (2018) wrote about subject-specific effects being "random" throughout their article (even in the abstract). Yet, they clarified that no parametric distribution for these effects is assumed, leaving it ambiguous whether any distribution is required at all and what exactly is meant by random. It might be speculated that the term random effects was used synonymously with subject-specific effects, without any distributional implication.

Nathoo et al. (2018) developed their Bayesian within-subject credible interval to offer a Bayesian alternative to the within-subject confidence interval. The procedure used in the development of this Bayesian interval conditions on ML estimates of subject-specific effects. These estimates are the deviations of individual means across the different levels from the overall mean. In the statistical literature, this approach of conditioning on ML estimates is also widely recognized as empirical Bayes. Its usefulness has been explored by, for example, Liang and Tsou (1992). More specifically, employing aspecific uninformative prior distribution, the so-called Jeffreys prior, for the level means and the variance of the residuals, Nathoo et al. (2018) first derived a posterior distribution of a level mean. What is crucial to note is that this posterior conditions on the subject-specific effects. Similar to the normalization and decorrelation procedures used by others, conditioning on estimates of subject-specific effects effectively minimizes between-subject variability and thereby the impact of this type of variability on the interval. In a next step, Nathoo et al. (2018) made use of the empirical Bayes approach and substituted the subject-specific effects with their corresponding ML estimates. The posterior then took the form of a t-distribution (see Appendix of Nathoo et al., 2018, for details). From this distribution, Nathoo et al. (2018) obtained their interval. Note that each residual (ε_{ij}) in the model is essentially the interaction effect of subject i's being observed under level j (Loftus & Masson, 1994, Appendix A). Thus, their sum of squares (i.e., the sum of squares due to the subject × level interaction) is given by:

$$\mathrm{SS}_{\mathrm{int}} = \sum_{i=1}^{N} \sum_{j=1}^{C} \left(Y_{ij} - M_i + M - M_j \right)^2 \tag{2}$$

where M_i, M_j, and M are the ML estimates of the ith subject's mean, the jth level's mean, and the overall mean, respectively. Using this notation, Nathoo et al.'s within-subject interval is expressed as:

$$M_j \pm \sqrt{\frac{\mathrm{SS}_{\mathrm{int}}}{(N-1)\,C} \bigg/ N} \cdot t_{(N-1)C} \tag{3}$$

Heck's (2019) primary argument that this interval fails to convey the appropriate amount of uncertainty, as uncertainty in the estimates of subject-specific effects is not incorporated, is generally valid. However, we challenge his argument that uncertainty would be underestimated when ML estimates are used instead of shrunken estimates. This argument is based on the random-effects assumption. In many if not all articles on within-subject intervals, the premise is that subjects are randomly drawn from a larger population (random by design), which may involve a distributional assumption. What is less clear, however, is how this premise relates to the technical assumption that subject-specific effects stem from a common distribution. As one reviewer noted and in line with our argument, Heck (2019) assumed a parametric distribution, but Loftus and Masson (1994) as well as Cousineau (2005) and Morey (2008) did not. In mixed-effects modeling, the random-effects assumption is often made by specifying a parametric distribution. However, as Nathoo et al. (2018) pointed out on page 4, the assumption is not necessary and was indeed not used in their within-subject credible interval, justifying interpretation of this interval as a semiparametric interval. Echoing Nathoo et al. (2018), we question the necessity of the random-effects assumption. The fact that it is often made by researchers does not imply its necessity. In other words, avoiding the assumption does not introduce any extra amount of bias into Nathoo et al.'s (2018) within-subject credible interval apart from the bias resulting from ignoring the uncertainty in the estimates of subject-specific effects. It is interesting to note that Heck (2019) made the random-effects assumption and incorporated it in his analysis (see his Equation 6) in order "to actually fit the model" (p. 28). Specifically, he specified a normal distribution for all subject-specific effects, following common practice. In the following, we provide both a substantial and a formal argument in order to show that the random-effects assumption may not be necessary after all.

2. Why the random-effects assumption may *not* be necessary

First, we aim to clarify a common misconception regarding subject-specific effects. Heck (2019) appears to have operated under the assumption that subject-specific effects require that their estimates offer shrinkage and thus, technically, a common distribution be assumed. This random-effects assumption is indeed commonly made in the field. However, the rationale for considering subject-specific effects as random is typically not provided. Addressing the broader question of when effects should be classified as random, Searle, Casella, and McCulloch (1992) argued that this classification is appropriate if the focus is on their population characteristics, particularly their variance. However, it can be doubted whether researchers using within-subject designs are interested in inferring population variances of random effects. Rather, their interest lies in the within-subject factors and in their levels. Another possible justification of the random-effects assumption is the view that any Bayes method would involve some kind of parameter distribution or, in Bayesian terminology, a prior. This means that for the subject-specific effects, a prior would have to be specified. Despite the popularity of specifying one hierarchical prior for all subject-specific effects, however, an (independent) prior for each subject-specific effect could also be specified and justified. Consequently, there seems to be no entirely convincing reason to categorically treat subject-specific effects as random.

Further, given these considerations, one may ask whether uncertainty about factor levels is affected by the way subject-specific effects are treated. As we will argue, the answer is No, at least, when priors are assumed to be suitably uninformative. To see this, consider the mean squares associated with the sum of squares due to the subject × level interaction:

$$\mathrm{MS_{int}} = \frac{\mathrm{SS_{int}}}{\mathrm{df}} \tag{4}$$

df denotes degrees of freedom, a concept to which we will come back further below. First, let us examine the scenario where subject-specific effects are considered *fixed* rather than random. It is important to note that in this case, the interaction effects will also be fixed, because an interaction effect of two fixed effects is inherently fixed. Under these conditions, computing the expected value of the interaction mean squares simplifies to just removing the expectation operator (i.e., the E symbol). Consequently, the expected value of the mean squares is equal to the mean squares themselves:

$$\mathrm{E}\left(\mathrm{MS_{int}}\right) = \mathrm{MS_{int}} \tag{5}$$

The factor within the multiplicative term of the within-subject interval is obtained by dividing these mean squares by the sample size (N) and then taking the square root so that the formula for this interval reads:

$$M_j \pm \sqrt{\frac{\mathrm{MS_{int}}}{N}} \cdot t_{\mathrm{df}} \tag{6}$$

which is the classical within-subject confidence interval as proposed by Loftus and Masson (1994).

Now, suppose the subject-specific effects are *random*. Then, the interaction effects are random too. In this case, it can easily be shown (e.g., Searle et al., 1992) that the expected value of the mean squares is equal to the variance of the residuals in the model:

$$\mathrm{E}\left(\mathrm{MS_{int}}\right) = \sigma^2 \tag{7}$$

Applying the ANOVA method, which is a (formal) way to derive (co)variance estimates, we yield an estimate of this variance:

$$\hat{\sigma}^2 = \mathrm{MS_{int}} \tag{8}$$

Thus, under the random-effects perspective, the within-subject interval is:

$$M_j \pm \sqrt{\frac{\hat{\sigma}^2}{N}} \cdot t_{\mathrm{df}} \xrightarrow{\text{Eq. (7)}} M_j \pm \sqrt{\frac{\mathrm{MS_{int}}}{N}} \cdot t_{\mathrm{df}} \tag{9}$$

and thus exactly identical with the one obtained under the fixed-effects assumption. This is clear evidence that the assumption of subject-specific effects being random is indeed not necessary. It also means that a common distribution of subject-specific effects is not necessary, nor are shrunken estimates.

Next, we will shift focus to Heck's primary argument. Specifically, we will address the issue that Nathoo et al.'s within-subject credible interval does not accurately reflect the proper amount of uncertainty.

3. Correction to Nathoo et al.

Heck's critique regarding the lack of accounting for uncertainty in the estimates of subject-specific effects is justified. Nathoo et al. (2018) derived their within-subject credible interval by conditioning on these subject-specific effects. There is nothing wrong in principle with conditioning on parameters that are known. However, as the authors mentioned themselves, the subject-specific effects are in fact not known but need to be estimated, and this is why they replaced them by their ML estimates. Nathoo et al. (2018) described this use of plug-in estimates as giving their approach "an empirical Bayes flavour" (p. 2). Empirical Bayes approaches are not uncommon and often very useful in situations where other approaches such as integrating over nuisance parameters Basu (e.g., 1977), Hecht, Gische, Vogel, and Zitzmann (e.g., 2020) is impractical, for example, when there is no distribution for these parameters to integrate over (Liang & Tsou, 1992). However, this approach results in an interval that tends to be too narrow (Cox & Reid, 1987). Given this issue, it would be prudent to implement a correction to ensure a more accurate representation of uncertainty. We will show that there is a simple way to do this. Key to our approach is a clear understanding of the concept of degrees of freedom, which plays a crucial role in statistics. The within-subject credible interval suggested by Nathoo et al. (2018) is accurate when subject-specific effects are either known or estimated virtually without error, a situation that is rare. In general, these effects require estimation, introducing inherent uncertainty. Because the within-subject credible interval is based on them, this adds an additional amount of uncertainty to the interval. The concept of degrees of freedom was elucidated by Student (also known by his real name Gosset) in his famous article *The probable error of a mean*, which was published in 1908. In a nutshell, the degrees of freedom of an estimate reflect the amount of information underpinning it. This is calculated as the number of data points minus the number of additional estimates used in the estimate's computation. For example, when estimating the variance from a sample, the degrees of freedom equal the number of subjects sampled minus one, because computing the estimate of the variance involves one more estimate—the sample mean.

In a similar vein, we contend that the degrees of freedom in the within-subject credible interval needs to be adjusted. The reason is that to compute this interval, one must use estimates of subject-specific effects. Nathoo et al. (2018) used $(N-1)C$ degrees of freedom in their calculation of interaction mean squares and in the t-distribution for interval computation. This number is valid when subject-specific effects are known: with NC data points and C independently estimated parameters, namely the level means (M_j), the difference between the number of data points and the number of parameters simplifies to:

$$\mathrm{df} = NC - C = (N-1)C \tag{10}$$

which is indeed the degrees of freedom in Nathoo et al. (2018). Note that the grand mean (M) can be computed directly from the level means, which is why this computation does not necessitate subtracting yet another degree of freedom. However, when subject-specific effects are unknown and require estimation, an adjustment in degrees of freedom becomes essential. Without this adjustment, there would be a bias in the interaction mean squares, preventing uncertainty associated with estimating the subject-specific effects from being incorporated into the interval. To properly adjust the degrees of freedom, it is crucial to subtract the number of independently estimated subject-specific effects. Notice that there are only $N-1$ such parameters. One is determined due to the constraint that subject-specific effects sum up to zero. Applying the proposed adjustment, the degrees of freedom become:

$$\mathrm{df}_{\mathrm{adj}} = \mathrm{df} - (N-1) \overset{\text{Eq. (9)}}{=} (N-1)C - (N-1) = (N-1)(C-1) \tag{11}$$

Note that the adjustment is not made in the absence of any assumption about the subject-specific effects but conditional on assuming fixed effects. The adjusted degrees of freedom are simply the default for the interaction effect of two fixed effects. Our correction to the within-subject credible interval essentially involves replacing the degrees of freedom used by Nathoo et al. (2018) by these adjusted ones so that the corrected within-subject interval is:

$$M_j \pm \sqrt{\frac{\mathrm{SS}_{\mathrm{int}}}{(N-1)(C-1)} \Big/ N} \cdot t_{(N-1)(C-1)} \tag{12}$$

This formula shows that pooling the degrees of freedom is also necessary, a detail not addressed in earlier work but included in a recent work by Cousineau, Goulet, and Harding (2021).

Whereas our correction is particularly critical in a within-subject design featuring one factor with two levels (the bias is less pronounced when designs have more factors/levels), researchers might still consider applying the correction in more complex designs. In a design with P within-subject factors, each having C_p levels, the adjustment becomes more nuanced. Instead of using $(N-1)\prod_{k=1}^{P} C_p$ degrees of freedom as Nathoo et al. (2018), we suggest that $(N-1)\left(\prod_{k=1}^{P} C_p - 1\right)$ degrees of freedom be used. Consequently, the corrected within-subject interval for these designs reads $M_j \pm \sqrt{\frac{\mathrm{SS}_{\mathrm{int}}}{(N-1)\left(\prod_{k=1}^{P} C_p - 1\right)} \Big/ N} \cdot t_{(N-1)\left(\prod_{k=1}^{P} C_p - 1\right)}$, where here $\mathrm{SS}_{\mathrm{int}}$ is the sum of squares due to the interaction of the subject factor with all within-subject factors.

4. Summary

Within-subject designs enjoy considerable popularity in experimental psychology due to their enhanced power compared to between-subject designs. These designs typically convey uncertainty in level means through a within-subject interval. There has been considerable scholarly debate on the computation of this type of interval, with Loftus and Masson's (1994) within-subject confidence interval being notably prominent. More recently, Nathoo et al. (2018) proposed a Bayesian within-subject credible interval, which conditions on ML estimates of subject-specific effects, a procedure sometimes called empirical Bayes.

While recognizing the interpretational benefit of Nathoo et al.'s within-subject credible interval, Heck (2019) raised criticism, particularly regarding the treatment of ML estimates as known values. This approach fails to incorporate the uncertainty in these estimates into the interval, a point we concur with. However, we questioned the necessity of incorporating shrunken estimates, which typically occurs when technically (i.e., at the implementation level), subject-specific effects are assumed to stem from a common distribution and thus to be random. We showed that this technical assumption is not necessary and derived a simple, yet effective correction to the within-subject credible interval, mainly involving an adjustment in degrees of freedom and leading to several implications. Firstly, Nathoo et al.'s method is indeed valid, except that it involves too many degrees of freedom, thereby understating uncertainty. Secondly, contrary to common practice, technically assuming random effects is not necessary, which can be viewed as a strength as less assumptions are involved. This leads to the third point: Heck' suggested stepwise procedure as well as any other fully Bayesian approach is not the only possible remedy. Nathoo et al.'s formula and our corrected version have a simple closed form, offering ease of application and avoiding the complexities of fully Bayesian approaches. For example, while fully Bayes via Markov chain Monte Carlo techniques is generally promising, particularly when the model includes many random effects (e.g., Zitzmann, Lüdtke, Robitzsch, & Hecht, 2021), it can be computationally demanding (Hecht et al., 2020), and it requires expert knowledge, for example, in order to diagnose whether chains have converged (see also Zitzmann & Hecht, 2019).

It is noteworthy that by employing the adjusted degrees of freedom, our within-subject credible interval aligns with the within-subject confidence interval as described by Loftus and Masson (1994). It also mirrors Morey's (2008) corrected within-subject confidence interval, the primary distinctions being that Morey's interval uses separate mean squares for each factor level. Our correction is thus similar to Morey's proposed correction to Cousineau's (2005) within-subject confidence interval. Cousineau interval differs from Loftus and Masson's (1994) by a factor of $C/(C-1)$, leading Morey (2008) to recommend multiplying by this factor rectify the bias. Given these prior works, one might question the novelty of our contribution. However, our starting point was a newly proposed Bayesian credible interval (Nathoo et al., 2018). This Bayesian interval had been critically noted for not accounting for uncertainty due to estimating subject-specific effects, prompting Heck (2019) to propose a fully Bayes approach to address this flaw. The present article was intended to critically evaluate the necessity of the assumptions made in this approach. We came to the conclusion that what we have called the random effects assumption is not necessary. Based on this finding, we proposed a simpler, yet valid alternative to Heck' (2019) approach, capitalizing on the theory of degrees of freedom. Somewhat surprisingly, the resulting calculation rule for the corrected Bayesian credible interval turned out to be the same as that for the frequentist counterpart. However, it is important to emphasize that this formal equivalence does not imply that these intervals (and corresponding interpretations!) are the same. Nathoo et al. (2018) invested considerable effort into developing their within-subject credible interval, adhering to fundamental Bayesian principles, such as specifying priors and deriving posteriors using Bayes' theorem. Consequently, the interval we aimed to modify is undeniably Bayesian. Correcting this Bayesian interval by adjusting degrees of freedom does not alter its Bayesian nature; therefore, the corrected interval remains a Bayesian interval. The phenomenon that the formula for calculating it coincides with that used for the frequentist confidence interval is purely coincidental and does not warrant conflating the two types of intervals. To illustrate this point, consider two identical twins. Despite their apparent similarities, we would not claim these different individuals to be the same person. Moreover, neither Nathoo et al. (2018) nor Heck (2019) previously suggested the correction we propose. Indeed, Nathoo et al. (2018) mentioned that a procedure similar to ours, specifically Morey' (2008) correction, might be applicable, but they would not pursue this further (p. 5). Heck (2019) implemented a fully Bayesian approach using MCMC techniques and noted the similarities between the results of his simulation studies and those obtained from the frequentist interval. However, he did not advocate for an adjustment of the degrees of freedom to amend the approach by Nathoo et al. (2018), nor did he provide reasoning for why such an adjustment might be advantageous.

As Nathoo et al. (2018) noted, a Bayesian within-subject credible interval conveys information about "the plausibility of the observed separation between means" (p. 7 f.). Therefore, such an interval is better than none and is certainly more useful than an interval that depends on between-subject variability, which is irrelevant to the hypothesis at hand. However, if the aim is to compare levels, an additional adjustment of the within-subject credible interval is in order. For the frequentist counterpart, the adjustment is multiplying by $\sqrt{2}$. Two level means are significantly different from one another if the second mean lies outside of the first mean's adjusted interval. However, whether the Bayesian interval can be adjusted in a similar manner is unclear. Significance testing is first and foremost a frequentist concept. The Bayesian concept that corresponds most to testing for significance is evaluating the Bayes factor. In contrast to the frequentist within-subject confidence interval, which has the aforementioned relationship with significance testing, the Bayesian within-subject credible interval has not yet been shown to have a similar relationship with the Bayes factor. This has also been critically noted by Nathoo et al. (2018). However, as showing such a relationship reaches beyond the scope of the present article, it must be left for future research.

In conclusion, we presented a correction to Nathoo et al.'s within-subject credible interval that accounts for the uncertainty in estimates of subject-specific effects. Our correction is essentially an adjustment in degrees of freedom in the interaction mean squares and in the t-distribution for interval computation. Our hope is that this corrected within-subject credible interval will promote the adoption of Bayesian intervals, which offer an interpretational benefit over confidence intervals. The fact that our corrected within-subject credible interval aligns with those proposed by Loftus and Masson (1994) and Morey (2008) suggests that our interval could similarly be calculated using the procedures proposed by Baguley (2012) and standard non-Bayesian software (see also Cousineau et al., 2021; Cousineau & O'Brien, 2014; O'Brien & Cousineau, 2014). This alignment may support the pragmatic view that in many applications, the practical differences between frequentist and Bayesian results are minimal when uninformative priors are used, allowing results to be interpreted in either framework (see, e.g., Albers, Kiers, & van Ravenzwaaij, 2018, for this argument). However, in line with Nalborczyk, Bürkner, and Williams (2019), we wish to emphasize that even if results are numerically identical, their interpretations do differ and should consistently align with the methodology employed (see also Zitzmann & Loreth, 2021). If the methodology is Bayesian, interpretations must adhere to the Bayesian perspective, even when results are computed using methods from the frequentist framework that yield similar numerical outcomes.

CRediT authorship contribution statement

Steffen Zitzmann: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Christoph Lindner:** Writing – original draft, Validation, Conceptualization. **Martin Hecht:** Writing – review & editing, Supervision.

References

Albers, C. J., Kiers, H. A. L., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*, 1–8. http://dx.doi.org/10.1525/collabra.149.

American Psychological Association (2019). *Publication manual of the American Psychological Association* (7th ed.). Washington, D.C.: American Psychological Association.

Amrhein, V., & Greenland, S. (2022). Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. *Journal of Information Technology*, *37*, 316–320. http://dx.doi.org/10.1177/02683962221105904.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*, 158–175. http://dx.doi.org/10.3758/s13428-011-0123-7.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association*, *72*, 355–366. http://dx.doi.org/10.1080/01621459.1977.10481002.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45. http://dx.doi.org/10.20982/tqmp.01.1.p042.

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology*, *15*, 226–241.

Cousineau, D., Goulet, M.-A., & Harding, B. (2021). Summary plots with adjusted error bars: The superb framework with an implementation in R. *Advances in Methods and Practices in Psychological Science*, *4*, 1–18. http://dx.doi.org/10.1177/25152459211035109.

Cousineau, D., & O'Brien, F. (2014). Error bars in within-subject designs: A comment on Baguley (2012). *Behavior Research Methods*, *46*, 1149–1151. http://dx.doi.org/10.3758/s13428-013-0441-z.

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society*, *49*, 1–39.

Hecht, M., Gische, C., Vogel, D., & Zitzmann, S. (2020). Integrating out nuisance parameters for computationally more efficient Bayesian estimation – An illustration and tutorial. *Structural Equation Modeling*, *27*, 483–493. http://dx.doi.org/10.1080/10705511.2019.1647432.

Heck, D. (2019). Accounting for estimation uncertainty and shrinkage in Bayesian within-subject intervals: A comment on Nathoo, Kilshaw, and Masson (2018). *Journal of Mathematical Psychology*, *88*, 27–31. http://dx.doi.org/10.1016/j.jmp.2018.11.002.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. http://dx.doi.org/10.3758/s13423-013-0572-3.

Hollands, J. G., & Jarmasz, J. (2010). Revisiting confidence intervals for repeated-measures designs. *Psychonomic Bulletin & Review*, *17*, 135–138. http://dx.doi.org/10.3758/PBR.17.1.135.

Liang, K.-Y., & Tsou, D. (1992). Empirical Bayes and conditional inference with many nuisance parameters. *Biometrika*, *79*, 261–270. http://dx.doi.org/10.2307/2336837.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490. http://dx.doi.org/10.3758/BF03210951.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64. http://dx.doi.org/10.20982/tqmp.04.2.p061.

Nalborczyk, L., Bürkner, P.-C., & Williams, D. R. (2019). Pragmatism should not be a substitute for statistical literacy, a commentary on Albers, Kiers, and van Ravenzwaaij (2018). *Collabra: Psychology*, *5*, 1–5. http://dx.doi.org/10.1525/collabra.197.

Nathoo, F. S., Kilshaw, R. E., & Masson, M. E. J. (2018). A better (Bayesian) interval estimate for within-subject designs. *Journal of Mathematical Psychology*, *86*, 1–9. http://dx.doi.org/10.1016/j.jmp.2018.07.005.

O'Brien, F., & Cousineau, D. (2014). Representing error bars in within-subject designs in typical software packages. *The Quantitative Methods for Psychology*, *10*, 56–67. http://dx.doi.org/10.20982/tqmp.10.1.p056.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.

Student (1908). The probable error of a mean. *Biometrika*, *1*, 1–25.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386. http://dx.doi.org/10.1037/1082-989X.6.4.371.

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling*, *26*, 646–661. http://dx.doi.org/10.1080/10705511.2018.1545232.

Zitzmann, S., & Loreth, L. (2021). Regarding an "almost anything goes" attitude toward methods in psychology. *Frontiers in Psychology*, *12*, 1–4. http://dx.doi.org/10.3389/fpsyg.2021.612570.

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of Bayesian approaches in small samples: A comment on Smid, McNeish, Miočević, and van de Schoot (2020). *Structural Equation Modeling*, *28*, 40–50. http://dx.doi.org/10.1080/10705511.2020.1752216.

Zitzmann, S., Nagengast, B., Hübner, N., & Hecht, M. (2024). A simple solution to heteroscedasticity in multilevel nonlinear structural equation modeling. Manuscript (submitted for publication).