Computational Brain & Behavior (2023) 6:102–126
https://doi.org/10.1007/s42113-021-00125-y

ORIGINAL PAPER

Sample Size Determination for Bayesian Hierarchical Models Commonly Used in Psycholinguistics

Shravan Vasishth[1] · **Himanshu Yadav**[1] · **Daniel J. Schad**[2] · **Bruno Nicenboim**[3]

Accepted: 13 December 2021 / Published online: 4 March 2022

Abstract

We discuss an important issue that is not directly related to the main theses of the van Doorn et al. (*Computational Brain and Behavior*, 2021) paper, but which frequently comes up when using Bayesian linear mixed models: how to determine sample size in advance of running a study when planning a Bayes factor analysis. We adapt a simulation-based method proposed by Wang and Gelfand (*Statistical Science* 193–208, 2002) for a Bayes factor-based design analysis, and demonstrate how relatively complex hierarchical models can be used to determine approximate sample sizes for planning experiments.

Keywords Sample size determination · Bayesian data analysis · Hierarchical models · Power analysis

Introduction

The papers that appear in this special issue are intended to be a response to van Doorn et al. (2021). Our main goal here is to address an important issue that is not directly related to the van Doorn et al. (2021) paper but is very relevant for researchers planning to use hierarchical models and Bayes factors in their research. This issue is sample size calculation when planning a Bayes factor-based study.

Before we turn to the main topic of our paper, we briefly comment on the van Doorn et al. (2021) paper.

A Brief Comment on van Doorn et al. 2021

In van Doorn et al. (2021), the authors discuss how Bayes factors can be computed using the `BayesFactor` package (Morey et al., 2015). The authors address the important question of what the appropriate models are that should be compared in a Bayes factor-based analysis. Among the different types of models that they consider, two models are of particular interest in psycholinguistics and psychology. Taking a two-condition repeated-measures design as an example, they compare (a) a model that includes varying intercepts and varying slopes as well as a two-level predictor as a fixed effect (their Model 6) with (b) a model with the same random effects structure as in Model 6, but no fixed effect for the predictor (their Model 5). This is a nested model comparison: Model 5 is nested within Model 6. Nested model comparisons of this type are quite close to the typical hypothesis testing approach needed in fields like linguistics, psycholinguistics, and psychology.

There are several potentially important limitations in the van Doorn et al. (2021) paper.

Issue 1: Ignoring item-level variability through aggregation can be dangerous First, in psycholinguistics and related areas, the typical repeated-measures design has not only random effects for subjects but also crossed random effects for items—this common situation is not considered in the van Doorn et al. (2021) paper. There is a very important reason for including items as a random effect: in a now-classic article, Clark (1973) showed that items are an important source of variance in the data (this point is reiterated in Westfall et al., 2017), and must be modeled if we want generalizability (Singmann et al., 2021; Yarkoni, 2020). The approach that van Doorn et al. (2021) take in their paper is to aggregate the data over items; this is potentially very dangerous because it will tend to hide this crucial source of variance. In fact, one of the greatest advantages of a hierarchical model is that it allows one to

✉ Shravan Vasishth
vasishth@uni-potsdam.de

[1] Department of Linguistics, University of Potsdam, Potsdam, Germany

[2] Department of Psychology, Health and Medical University, Potsdam, Germany

[3] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, The Netherlands

model different sources of variance simultaneously. Thus, one important comment we have on the van Doorn et al. (2021) paper is that aggregation should in general *never* be done.

Issue 2: Effect sizes should generally not be standardized Second, van Doorn et al. (2021) raise the following important question: "How can we construct an effect size that is meaningfully standardized? In other words, what variance should we standardize by?" This question is important because the prior specification for the effect of interest is affected by the standardization. However, this question presupposes that effect sizes should be standardized at all—we concur with another paper in this special issue (Singmann et al., 2021) that standardizing may not be a good idea, except in very specific situations (Baguley, 2009). Effect sizes will generally have more meaningful and more interpretable priors if they are defined on the scale of theoretical interest. For example, in reading studies, meta-analyses yield a range of effect size estimates on the millisecond scale (Jäger et al., 2017; Vasishth et al., 2013). These estimates—and crucially, the uncertainty of the estimates—can form the basis for a prior distribution for a future study. Another example is event related potentials; there, the dependent measure is in microvolts. For specific experiment designs, meta-analysis (Nicenboim et al., 2020) allows us to define informative priors that can be used in planning future experiments. An example is the registered report by Stone et al. (2021) that uses informative priors developed in (Nicenboim et al., 2020) for a planned ERP study. Focusing on a standardized effect size as a point value will be misleading, because one always has some uncertainty on one's prior beliefs (Baguley, 2009). At least for experiment designs that one has some prior experience with (or quantitative theoretical predictions for), we see no compelling reason to standardize the effect size.

Issue 3: Modeling varying intercept and varying slope correlations can be important Third, although the authors refer to their Model 6 as the "full model", this is technically not a full model (Barr et al., 2013, call it a "maximal" model): it does not include correlations between the subject random intercepts and random slopes. As they explain in their paper, the reason that van Doorn et al. (2021) drop the correlation term is that "The BayesFactor package…does not explicitly model correlations between random slopes and intercepts." Although these two problems may not be serious for many fields, for psycholinguistics they render the van Doorn et al. (2021) paper less useful. This is because the correlations can be of central interest when studying individual differences (e.g., Yadav et al., 2021; Pregla et al., 2021). It is therefore necessary to have the capability to model these correlations. Unlike the `BayesFactor` package, the probabilistic programming language `Stan` (Carpenter et al., 2017) and the associated front-end packages `rstanarm` (Goodrich et al., 2020) and `brms` (Bürkner, 2017) make it easy to include these correlations using the standard `lme4` syntax.

Some of the limitations of the van Doorn et al. (2021) paper stem from the limitations built into the `BayesFactor` package. Today, there exist much more flexible probabilistic programming environments such as Stan (Carpenter et al., 2017), which can easily deal with the issues we raise above. In previous work (Schad et al., 2020), we have discussed models with a full variance-covariance structure for subjects and for items; this work elaborates on our three observations above about the van Doorn et al. (2021) paper. In the Schad et al. (2020) paper, a detailed workflow is presented that begins with prior predictive checks and ends with posterior predictive checks for relatively complex hierarchical models. In a subsequent paper (Schad et al., 2021), we expand on the sometimes extraordinary difficulties associated with Bayes factors calculations in hierarchical models of the type that are used in psycholinguistics. Perhaps the most important issue discussed in that paper—which seems to be underemphasized in discussions of Bayes factors—is that Bayes factors require careful sensitivity analyses (a range of increasingly informative priors on the target parameter). Such sensitivity analyses are typically not carried out. Instead, researchers routinely report a single Bayes factor, usually under some "default prior" like the Cauchy prior (for some recent examples, see Hammerly et al., 2019; Montero-Melis et al., 2019). These kinds of Bayes factors analyses using "default priors" (which, if we understand it correctly, are baked into the `BayesFactor` package) have the potential to be extremely misleading. This kind of oversimplified reporting of Bayes factors should be strongly discouraged.

The two papers by Schad and colleagues can be seen as complementing the van Doorn et al. (2021) paper that appears in this special issue, because they provide details on (i) fitting complex hierarchical models that the authors do not consider, and (ii) carrying out Bayes factor based hypothesis testing with such complex models.

Although van Doorn et al. (2021) do not discuss this issue, it is not obvious how one can plan sample sizes when intending to use Bayesian hierarchical models. This is the main topic of our paper, and we discuss this issue next.

Approaches to Sample Size Planning in Bayesian Analyses

It may sound surprising to Bayesian modelers that sample size planning is even something to plan for: One of the many advantages of Bayesian modeling is that it is straightforward to plan an experiment without necessarily specifying the sample size in advance (e.g., Spiegelhalter et al., 2004; Schönbrodt and Wagenmakers, 2018). Indeed, in our own research, running an experiment until some precision criterion in the posterior distribution is reached (Freedman et al., 1984; Spiegelhalter et al., 1994; Kruschke, 2014; Kruschke & Liddell, 2018) is our method of choice (Jäger et al., 2020; Vasishth et al., 2018; Stone et al., 2021). This approach is possible to implement if one has sufficient financial resources (and time) to keep running an experiment till a particular precision criterion is reached.

However, even when planning a Bayesian analysis, there can be situations where one needs to determine sample size in advance. One important situation where this becomes necessary is when one applies for research funding. In a funding proposal, one obviously has to specify the sample size in advance in order to ask for the necessary funds for conducting the study. Other situations where sample size planning is needed is in the design of clinical trials, the design of replication trials, and when pre-registering experiments and/or preparing registered reports.

There already exist good proposals on how to work out sample sizes in advance, specifically in the case of Bayesian analyses (e.g., Schönbrodt & Wagenmakers, 2018; Weiss, 1997). The proposal by Schönbrodt and Wagenmakers (2018), which seems to be a simpler version of a much earlier paper by Wang and Gelfand (2002), aims to ensure that the researcher obtains strong evidence for the effect being estimated. They consider three scenarios: a fixed-n design, where the sample size is fixed in advance; an open-ended sequential design, where the experiment is run until strong evidence emerges either for or against an effect; and a sequential design with an upper bound for the sample size.

Although the Schönbrodt and Wagenmakers (2018) proposals are appropriate for many settings, their example case study is again a relatively simple design: a two-sample t-test with a directional hypothesis. Although this is pedagogically useful, this kind of design is almost never used in areas like psycholinguistics. As a consequence, for the newcomer from such a field, it is not obvious how one can compute sample sizes for complex hierarchical models that are the norm. A second problem with their proposal is that they use a Cauchy prior for the target parameter when carrying out Bayes factor calculations. Their approach could be appropriate in the application areas that the authors considered. However, in many psycholinguistic studies, the effect size is rather small relative to the sources of variance in the data (Jäger et al., 2017; Jäger et al., 2020; Nicenboim et al., 2020; Nicenboim et al., 2018; Bürki et al., 2020; Bürki et al., 2020), and in such studies a more informative prior is almost always necessary when doing Bayes factors analyses (Nicenboim & Vasishth, 2016; Schad et al., 2020; Schad et al., 2021). Although Stefan et al. (2019) address this issue using informative priors, they again use relatively simple statistical models, so it is again not at all obvious to the beginning Bayesian how these recommendations can be scaled up for complex hierarchical models.

The present paper focuses on the question: how to plan a sample size using Bayes factors for relatively complex hierarchical models? We provide a simulation-based approach for planning sample sizes for such situations. We focus exclusively on experiment designs that require the use of relatively complex hierarchical models that capture multiple variance components simultaneously.

It is very surprising that the simulation-based approach of Wang and Gelfand (2002) seems to have largely escaped the attention of researchers (this work is mentioned in passing in Kruschke, 2014, but it has largely been ignored in other work in psychology).

In the present paper, we unpack the approach taken in this important paper; the approach is important because it provides an easy-to-implement workflow for doing sample size calculations using complex hierarchical models of the type we discuss here. Because the application of the Wang and Gelfand (2002) simulation-based method will not be completely obvious to beginning researchers, we provide a complete example, with fully reproducible code.

Because there is no fixed convention in Bayesian statistics that gives specific names to different types of priors, in this paper, we follow the conventions that we use in Nicenboim et al. (2021). We use four broad categories of prior, which are on a continuum and not hard categorical distinctions: (a) flat, uninformative priors: these are priors that are as uninformative as possible (examples: Cauchy or Uniform priors); (b) regularizing priors: these priors downweight a priori unlikely values (example: reading times can only be positive); (c) principled priors: these encode all (or most of) the theory-neutral information that we have for a research problem (example: a Normal(6,0.6) prior for average reading time at a word, assuming a log-normal likelihood); and (d) informative priors: these incorporate prior knowledge (example: the Normal(0.12,0.4) that we use below in our example).

The Wang and Gelfand (2002) approach is as follows. We have adapted the procedure outlined below slightly for our purposes, but the essential ideas are due to these authors.

1. Decide on a distribution of effect sizes you wish to detect.
2. Choose a criterion that counts as a threshold for a decision. This can be a Bayes factor of, say, 10 (Jeffreys, 1939/1998).[1]
3. Then do the following for increasing sample sizes n:

[1] The Bayes factor is just one of many possible performance criteria; see Wang and Gelfand (2002) for some other alternatives.

(a) Simulate prior predictive data *niter* times (say, $niter = 100$) for sample size n; use informative priors (these are referred to as sampling priors in Wang and Gelfand (2002)).

(b) Fit the model to the simulated data using uninformative priors (these are called fitting priors in Wang and Gelfand (2002)), and derive the posterior distribution each time, and compute the Bayes factor using a null model that assumes a zero effect for the parameter of interest.

(c) Display, in one plot, the *niter* posterior distributions and the Bayes factors. If the chosen decision criterion is met reasonably well under repeated sampling for a given sample size, choose that sample size.

Figure 1 shows a schematic summary of the Wang and Gelfand procedure. For psychology and psycholinguistics, this procedure is in principle easy to implement. However, a crucial modification is necessary. Because the Bayes factor is so sensitive to the prior specification (see Schad et al., 2021, and the references cited there), it makes little sense to follow the Wang and Gelfand suggestion to use uninformative priors as fitting priors. Instead, we propose in this paper that the sampling priors and the fitting priors *both* be informative. In the example we show below, we keep the two types of prior identical because the informative priors represent what we currently know about the research question, and because we know from previous investigations that they are not unduly influential in determining the posterior distributions (Schad et al., 2021). Informative sampling priors make sense for efficiency reasons: the model will converge faster if the priors do not allow a wide range of (implausible) values a priori. For Bayes factors calculations, informative fitting priors are necessary anyway because effects are a priori likely to be small, and because using “default priors” such as a Cauchy prior will heavily bias the Bayes factor in favor of the null (Kruschke & Liddell, 2018; Schad et al., 2021; Nicenboim et al., 2020). If only estimation is the goal, not Bayes factors, then uninformative fitting priors can of course be used, because the posterior distribution tends to not be sensitive to the prior specification unless the data are very sparse (Kruschke & Liddell, 2018). Indeed, it is because of this prior sensitivity that Bayes factors analyses often include a sensitivity analysis, which amounts to reporting Bayes factor calculations under increasingly informative priors.

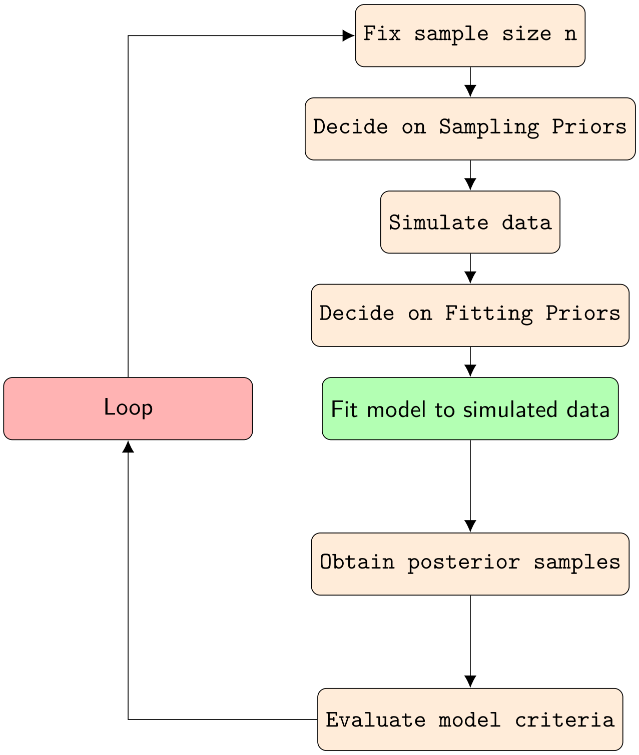


Fig. 1 A modified version of the workflow suggested by Wang and Gelfand (2002). The box colored green (labeled “Fit model to simulated data”) can be computationally very intensive. The box labeled “Loop” indicates that the procedure has to be repeated for each sample size chosen; this step will also be computationally intensive

With the above discussion as background, in the rest of this paper we provide a practical example of how the Wang and Gelfand approach can be adapted for research in psychology and psycholinguistics. We assume here that the reader is familiar with the van Doorn et al. (2021), and more generally with linear mixed modeling theory and practice, in both the frequentist and Bayesian frameworks (Bates et al., 2015; Vasishth et al., 2021; Nicenboim et al., 2021).

Example: a Two-Condition Repeated-Measures Design

In this case study, we consider a classic question in psycholinguistics: the subject vs. object relative clause (RC) processing difference in sentence processing. RCs are perhaps the single most studied syntactic construction in psycholinguistics. Examples of subject and object RCs are shown in (1).

(1) (a) Subject relative
The senator who met the journalist resigned.
(b) Object relative
The senator who the journalist met resigned.

The signature property of such relative clauses is that a clause (starting with the word *who*) interrupts the main clause of the sentence by modifying the subject of the sentence (here, *senator*). This interruption of the main clause leads to an increase in processing difficulty; in the above examples, it is generally more difficult to work out who met whom in object relatives vs. subject relatives. Thus, reading times at the word *met* are of theoretical interest. As an aside, notice that the experiment design in example (1) is confounded: the critical word *met* is not only not in the same word position in the two sentences, but the pre-critical region is different (*who* vs. *journalist*). It is well-known that differences in the pre-critical region (or even earlier regions) can cause differential amounts of spillover onto the critical region (e.g., Vasishth and Lewis, 2003; Mitchell, 1984). These potential confounds make it difficult to interpret reading time differences in the critical region. Unfortunately, this issue has largely been ignored in the psycholinguistics literature. Because our main point in this paper does not depend on these potential confounds in the design, we will ignore this issue, but we do acknowledge that such designs are potentially fatally flawed.

Despite the above issues with the design, a robust and uncontroversial finding for English is that, at the word *met*, object relatives have longer reading times than subject relatives. Various theoretical explanations have been proposed for this processing difference (see Grodner & Gibson, 2005, for a review of the theoretical proposals).

In this case study, we assume that we are planning a future reading study on English RCs. We will base our sample size planning on the original Grodner and Gibson (2005) data; that study had 42 participants and 16 items in a standard Latin-square repeated-measures design. We know from previous meta-analyses and power analyses of reading studies in psycholinguistics (Jäger et al., 2017) that a sample size of 42 subjects for the Grodner and Gibson (2005) design will lead to a hopelessly underpowered design. A future experiment should therefore have a larger sample size; the question is how much larger. We present our adaptation of the Wang and Gelfand procedure for answering this question.

Because the Grodner and Gibson (2005) data are available, we can use the parameter estimates from these data to obtain some initial ballpark estimates for the different variance components and fixed effects. The standard frequentist linear mixed models analysis using the `lme4` package in R (Bates et al., 2015) is the following. The variable n indexes the nth row of the data frame; the dependent variable is reading time (rt) in milliseconds at the critical word of interest; and the variables subj and item refer to the subjects and items, arranged in a Latin-square design. The variables u and w are the adjustments by subjects and items to the intercept α and the slope β. The predictor is a sum-coded variable (so); object relatives are coded $+1/2$ and subject relatives $-1/2$ (Schad et al., 2020). The slope therefore gives us the effect size, in log ms, of the difference in means between the object and subject relative. σ is the residual standard deviation.

$$\begin{aligned} rt_n \sim\ & LogNormal(\alpha + u_{subj[n],1} + w_{item[n],1} + so_n \cdot \\ & (\beta + u_{subj[n],2} + w_{item[n],2}), \sigma) \end{aligned} \qquad (1)$$

The varying intercepts and varying slopes are represented by the variables u and w, assuming that i indexes subjects and j indexes items (in the Grodner and Gibson (2005) data, $i = 1, \ldots, 42$, $j = 1, \ldots, 16$). The corresponding variance-covariance matrices are shown below. The variables τ_{u_1}, τ_{u_2} are the standard deviations of the subject random effects, and τ_{w_1}, τ_{w_2} the standard deviations of the item random effects.

$$\begin{pmatrix} u_{i,1} \\ u_{i,2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma}_{\boldsymbol{u}}\right) \text{ where } i = 1, \ldots, 42 \qquad (2)$$

$$\mathbf{\Sigma}_{\boldsymbol{u}} = \begin{pmatrix} \tau_{u_1}^2 & \rho_u \tau_{u_1} \tau_{u_2} \\ \rho_u \tau_{u_1} \tau_{u_2} & \tau_{u_2}^2 \end{pmatrix} \qquad (3)$$

$$\begin{pmatrix} w_{j,1} \\ w_{j,2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma}_{\boldsymbol{w}}\right) \text{ where } j = 1, \ldots, 16 \qquad (4)$$

$$\mathbf{\Sigma}_{\boldsymbol{w}} = \begin{pmatrix} \tau_{w_1}^2 & \rho_w \tau_{w_1} \tau_{w_2} \\ \rho_u \tau_{w_1} \tau_{w_2} & \tau_{w_2}^2 \end{pmatrix} \qquad (5)$$

Although this is a relatively simple two-condition design, the Bayesian hierarchical model is already quite complex, with 9 parameters for the fixed effects and the variance components, and an additional 42 × 2 parameters for the by-subject adjustments, $u_{i,1}$ and $u_{i,2}$, and 16 × 2 parameters for the by-item adjustments, $w_{j,1}$ and $w_{j,2}$. That's a total of 125 parameters.[2]

Fitting the model using the `lme4` package assuming a log-normal likelihood yields the following estimates:

```
Random effects:
Groups   Name        Variance Std.Dev. Corr
subj    (Intercept) 0.101050 0.31788
         cond        0.049104 0.22160 0.58
item    (Intercept) 0.001719 0.04145
         cond        0.007850 0.08860 1.00
Residual             0.129841 0.36033
Number of obs: 672, groups: subj, 42; item,
 16
Fixed effects:
            Estimate Std. Error t value
(Intercept) 5.88306   0.05202 113.082
so          0.12403   0.04932   2.515
```

[2] Unlike in Bayesian hierarchical models, in frequentist linear mixed models the varying intercepts and varying slopes, u and w, are integrated out and are referred to as conditional modes (Bates et al., 2015). That is, in frequentist linear mixed models, the individual u and w values are not parameters; rather, only the standard deviations and correlations associated with u and w are parameters.

An immediate problem that we notice here is that the correlation between the varying intercepts and varying slopes by items cannot be estimated: the model returns a correlation of 1, which means that we have an ill-conditioned variance-covariance matrix. This suggests that the model may be overparameterized (Pinheiro & Bates, 2000, p. 156). This failure to estimate the correlation is due to the relatively low number of items (16); this kind of overparameterization is a common issue in linear mixed models (Bates et al., 2015).

Usually, the Bayesian hierarchical model will not experience this failure to produce a sensible posterior distribution for a parameter; this is because of the regularizing priors that are generally used in Bayesian models (Nicenboim et al., 2021). Of course, it can happen that Bayesian hierarchical models can also end up being overparameterized; such problems usually lead to convergence warnings.

The first step in fitting a Bayesian model is deciding on the priors; we address this point next and demonstrate how we came up with informative priors for the target parameter and for the variance components. The priors that we develop here are not intended to be generally applicable, but have been developed for the broad class of psycholinguistic reading studies exemplified by the Grodner and Gibson (2005) design.

Eliciting Priors for the OR-SR Difference

First, we explain how one can come up with an informative prior specification for the effect of relative clause type on reading time; this is the slope β in the model in Eq. 1. Theory suggests (see Grodner & Gibson, 2005) that subject relatives in English should be easier to process than object relatives, at the relative clause verb. This means that a priori, we expect the difference between object and subject relatives to be positive in sign. What would be a reasonable mean (and a plausible range of variation) for this effect? We can look at previous research to obtain some ballpark estimates. For example, Just and Carpenter (1992) carried out a self-paced reading study on English subject and object relatives, and their Figure 2 (p. 130) shows that the difference between the two relative clause types at the relative clause verb ranges from about 10 ms to 100 ms (depending on working memory capacity differences in different groups of subjects). This is already a good starting point, but we can look at some other published data to gain more confidence about the approximate difference between the conditions. For example, Reali and Christiansen (2007) investigated subject and object relatives in four self-paced reading studies; in their design, the noun phrase inside the relative clause was always a pronoun, and they carried out analyses on the verb plus pronoun, not just the verb as in Grodner and Gibson (2005). We can take into account the estimates from this study for developing out prior because including a pronoun like “I”, “you”, or “they” in a verb region is not going to increase reading times dramatically (short words are usually read quickly). The hypothesis for Reali and Christiansen (2007) was that because object relatives containing a pronoun occur more frequently in corpora than subject relatives containing a pronoun, the relative clause verb should be processed *faster* in object relatives than subject relatives. This is the opposite of the prediction for the reading times at the relative clause verb discussed in Grodner and Gibson (2005). The authors report comparisons for the pronoun and relative clause verb taken together (i.e., pronoun+verb in object relatives and verb+pronoun in subject relatives). In experiment 1, they report a −57 ms difference between object and subject relatives, with a 95% confidence interval ranging from −104 to −10 ms. In a second experiment, they report a difference of −53.5 ms with a 95% confidence interval ranging from −79 to −28 ms; in a third experiment, the difference was −32 ms [−48, −16]; and in a fourth experiment, −43 ms [−84, −2]. Thus, given these data from English, if we were investigating the effect of pronouns in relative clauses, we would want to allow the prior values to range from −100 ms to approximately 0 ms. Although the sign of the effect is the opposite to the one we expect, the absolute range of variation is still within 10 and 100 ms.

Another study involved English relative clauses is by Fedorenko et al. (2006). In this self-paced reading study, Fedorenko and colleagues compared reading times within the entire relative clause phrase (the relative pronoun and the determiner+noun+verb sequence inside the relative clause—four words). Their data show that object relatives are harder to process than subject relatives; the difference in means is 460 ms, with a confidence interval [299, 621] ms. This difference is much larger than in the other studies mentioned above, but this is because of the long region of interest considered—it is well-known that the longer the reading/reaction time, the larger the standard deviation and therefore the larger the potential difference between means (Wagenmakers & Brown, 2007). Obviously, we cannot take this larger range into account in developing our prior, but it is still useful to know the summed up reading time over four words will be approximately 460 ms, which is 115 ms per word.

This previous data from English relative clause studies gives us some empirical basis for assuming that the object minus subject relative clause difference in the Grodner and Gibson (2005) design on English could range from 10 to 100 ms or so. If the reading time were recorded for the entire relative clause region, as in Fedorenko et al. (2006), obviously the prior would have to be different.

There is further supporting evidence from the literature that designs such as the current one will have a relatively small effect size on the millisecond scale. In a recent investigation (Jäger et al., 2017), we established this in a meta-analysis of one broad class of phenomena: similarity-based interference in sentence comprehension. Interference here refers to the difficulty experienced by the comprehender during sentence comprehension (e.g., in reading studies) when they need to retrieve a particular word from their working memory but other words with similar features hinder retrieval. The meta-analysis reported in Jäger et al. (2017), which is based on published data from more than 80 studies, suggests that the effect sizes for interference effects range from at most −50 to 50 ms, depending on the phenomenon (some kinds of interference cause speed-ups, others cause slow-downs; see the discussion in Engelmann et al., 2020, 12). In reading studies, whenever papers report unusually large effects for interference or related phenomena, these are usually what Gelman and Carlin (2014) call Type M errors (for real-life examples, see Jäger et al., 2020; Nicenboim et al., 2018; Vasishth et al., 2018). Given that the Grodner and Gibson (2005) design falls within the broader class of interference effects (Lewis & Vasishth, 2005; Vasishth et al., 2019; Vasishth & Engelmann, 2022), it is reasonable to choose informative priors that reflect this observed range of interference effects in the literature. Of course, when analyzing actual data, one must investigate the effect using a range of priors on the target parameter to interpret the Bayes factor analysis (Schad et al., 2021).

Now, because the Grodner and Gibson (2005) data is being analyzed with a log-normal likelihood, the prior for the slope parameter has to be on the log scale. Therefore, for the present purposes, we will define an informative prior on the log scale for the slope parameter: Normal(0.12, 0.04). Assuming a mean reading time of 6 log ms, this prior roughly corresponds to an effect size on the millisecond scale that has a 95 credible interval ranging from 16 ms to 81 ms.

Deciding on Principled Priors for the Other Parameters in the Model

In a repeated-measures two-condition design with reading time in milliseconds as a dependent measure, and crossed subjects and items (this is usually a Latin-square design), the likelihood usually chosen is a log-normal. In the log-normal likelihood, the parameters are on the log scale, and so if we set truncated $Normal(\mu = 0, \sigma = 1)$ priors (truncated so that the values can be only positive) on the variance components, these will generate unrealistic prior predictive data (Schad et al., 2021). Simulations show that a principled prior on the random effects variance components (the components τ) would be $Normal(0, 0.1)$ and for the residual standard deviation, $Normal(0, 0.5)$; see Schad et al. (2020) and Schad et al. (2021) for details. Instead of principled priors, one could certainly use informative priors for the variance components if one has previous data on the research topic. Similarly, for the intercept, we choose a principled prior of Normal(6,0.6) on the log ms scale; our investigations show that this choice of prior for the intercept generates realistic mean reading time data (Schad et al., 2021). For the correlation matrices in the random effects, we use a regularizing LKJ(2) prior; this prior downweights ± 1 as possible values for the correlations and therefore prevents the ill-conditioned variance-covariance matrix we saw earlier with the `lme4` model fit. The LKJ prior (Lewandowski et al., 2009) is available in `Stan` (Carpenter et al., 2017) and `brms` (Bürkner, 2017).

Graphically Investigating the Prior Predictive Data to Evaluate the Priors

Figure 2 shows the impact of the principled priors on the intercept and variance components, and the informative prior on the slope parameter, both on the log millisecond scale and the millisecond scale. This figure shows that the priors for the intercept (Normal(6,0.6)) and slope (Normal(0.12,0.04)) parameters are reasonable, as they generate realistic distributions on the ms scale. Notice that one impact of the variance components is to create a little extra uncertainty on the effect size of the relative clause effect.

With the above discussion as background, we present an implementation of the modification to the Wang and Gelfand (2002) approach.

Bayes Factor Analysis for Sample Size Calculation: a Modification of the Wang and Gelfand 2002 Approach

The Bayes factor-based approach to sample size calculation has the property that it takes into account all sources of variance in the parameters. This includes the uncertainty on the parameter of interest, the slope parameter in the linear mixed model shown in Eq. 1 above. The priors we derived above are shown below.

$$
\begin{aligned}
\alpha &\sim Normal(6, 0.6)\\
\beta &\sim Normal(0.12, 0.04)\\
u_1 &\sim Normal(0, \tau_{u_1})\\
u_2 &\sim Normal(0, \tau_{u_2})\\
w_1 &\sim Normal(0, \tau_{w_1})\\
w_2 &\sim Normal(0, \tau_{w_2})\\
\tau_{u_1} &\sim Normal_+(0, 0.1)\\
\tau_{u_2} &\sim Normal_+(0, 0.1)\\
\tau_{w_1} &\sim Normal_+(0, 0.1)\\
\tau_{w_2} &\sim Normal_+(0, 0.1)
\end{aligned}
$$

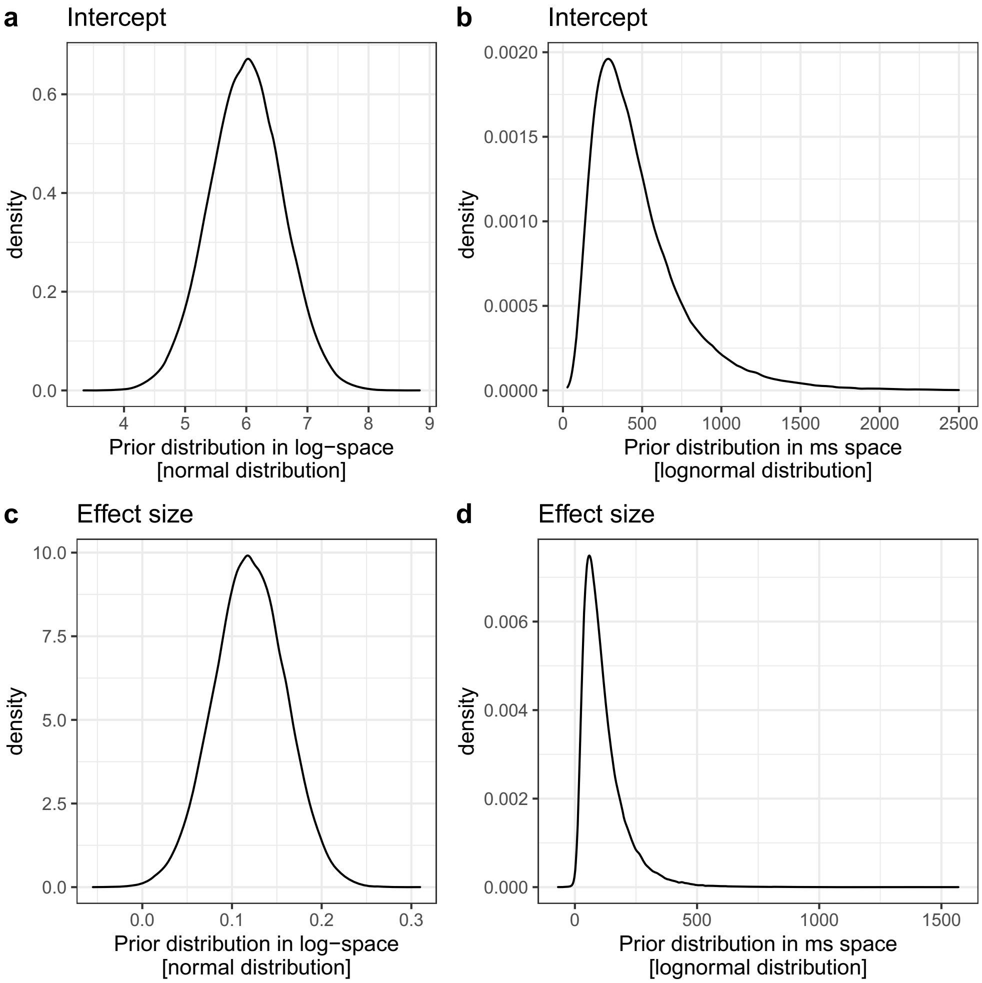


Fig. 2 Prior distribution in log-space and in ms space for a toy example of a linear regression. Panel **a** displays the prior distribution of the intercept in log-space. Panel **b** displays the prior distribution of the intercept in ms space. Panel **c** displays the prior distribution of the effect size in log-space, marginalizing over the intercept. Panel **d** displays the prior distribution of the effect size in ms space

$$\begin{aligned} \rho_u, \rho_w &\sim LKJ(2) \\ \sigma &\sim Normal_{+}(0, 0.5) \end{aligned} \tag{6}$$

To keep the computation tractable for this paper, we will simulate 100 datasets for each of the sample sizes 42, $42 \times 5 = 210$, and $42 \times 9 = 378$ subjects. Then, we fit the data using 10,000 iterations per chain (the warmup is 2000 iterations). In the Appendix, we show that even larger subject sample sizes, $42 \times 13 = 546$, $42 \times 17 = 714$, $42 \times 21 = 882$, and $42 \times 25 = 1050$ lead to computational problems in estimating Bayes factors using the bridge sampling procedure that we use (Gronau et al., 2017). For such large sample sizes, 50,000 iterations would be needed per chain (Bürkner, 2017), which is computationally costly.

The approach we take is the following:

1. Define informative priors (Schad et al., 2020) for this particular design and method (as discussed above).
2. Generate prior predictive data using the informative priors (the sampling priors).
3. Then fit the model repeatedly, again using the informative priors as fitting priors (this deviates from the Wang and Gelfand recommendation).

In the simulations below, one could have varied the items as well; but for simplicity we keep the number of items constant in this example.

Computing Hardware and Approximate Computing Times

We used a server with 40 physical cores and 80 logical cores. (There were 2 sockets, 20 physical cores per socket, and 2 threads per core, which make a total of 80 logical cores).

A typical Stan model with four Markov chains requires four cores to parallelize the sampling. We want to fit models on 100 simulated datasets for each sample. We first fit Bayesian models with 4 chains (consuming 4 cores) on 20 simulated datasets at a time (as this would use all 20×4 cores). This process is repeated five times to get Bayes factors on all 100 simulated datasets for each sample size.

The recorded completion times for each sample are as follows:

- 42 subjects: approximately 1 h.
- 210 subjects: approximately 8 h.
- 378 subjects: approximately 31 h.

Results

Figure 3 shows the posterior distributions and Bayes factors for each of the three sample sizes (42, 210, 378); the Bayes factors are truncated at 1×10^{20} to make it easier to view the credible intervals (only six data points are elided due to the truncation, for sample size 378). The figure shows that, as expected, the proportion of Bayes factors greater than 10 increases with increasing sample size. The figure also reveals that the 95% credible intervals of the posterior distributions in all three sample sizes either land within or overlap with the prior 95% credible interval of the effect (the horizontal broken lines). It is only in the small sample size (42 subjects) that we get one posterior distribution that lands entirely outside the range specified a priori. Because our prior range of effect sizes is relatively large (Normal(0.12,0.04)), using the region of practical equivalence (ROPE) approach for hypothesis testing (Freedman et al., 1984; Spiegelhalter et al., 1994; Kruschke, 2014) will be of only limited use; the ROPE approach will be more useful if the prior predicted range of effects is relatively narrow.

The second interesting observation is that the amount of variability in the Bayes factor under repeated sampling increases with sample size. This becomes clearer visually in Fig. 4, which shows the distributions of the log Bayes factors for the three sample sizes; the ridge plots have to be displayed on the log scale in order to be able to show the spread.

In sum, given the simulations shown above, we would plan for approximately 300 subjects when planning a study.

Discussion

We have presented an adapted version of the Wang and Gelfand (2002) approach to planning sample size when using linear mixed models and Bayes factors. We used an

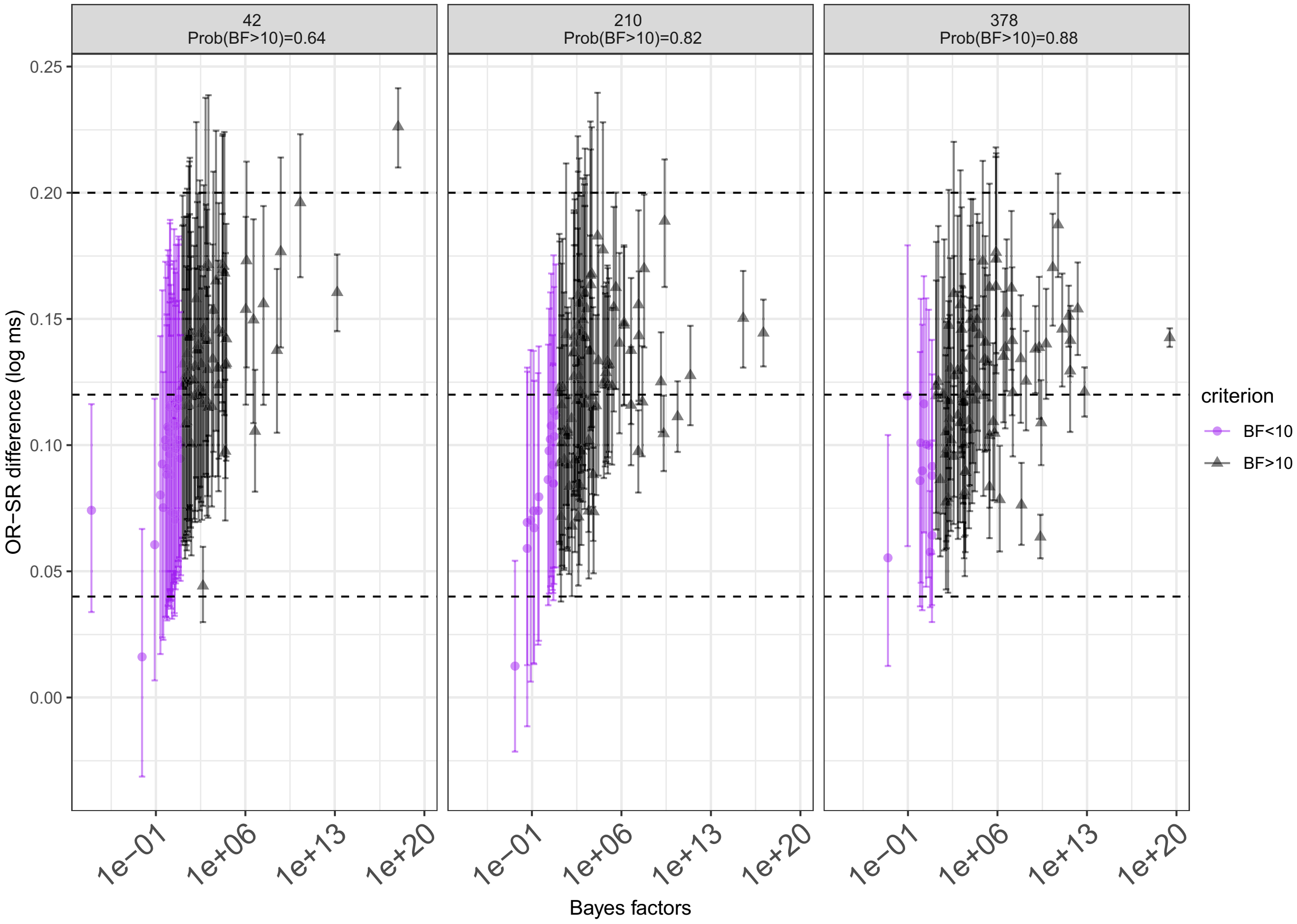


Fig. 3 The y-axis has posterior means and 95% credible intervals of the effect of interest (object minus subject relative clause reading times), and the x-axis has Bayes factors truncated at 1e+20 to make it easier to see the distribution of Bayes factors. The three plots show the distributions for the three sample sizes. Also shown for each sample size is the probability of Bayes factor being greater than 10 under hypothetical repeated sampling

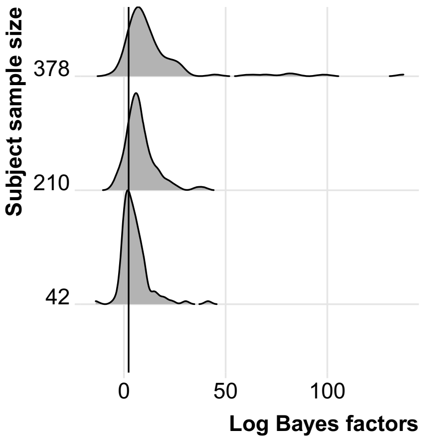


Fig. 4 The distributions of log Bayes factors under repeated sampling, for the three subject sample sizes. The vertical line show the log Bayes factor corresponding to log(10) in favor of the alternative

example from a reading experiment design that has been widely used in psycholinguistics, the subject vs. object relative clause difference.

In the Bayes factor-based analysis, the main deviation we made from the Wang and Gelfand method was to use the same informative priors in both the sampling and fitting stages of the modeling. Wang and Gelfand had proposed using informative priors as fitting priors, and uninformative priors as fitting priors.

The motivation for this deviation in our approach is that, at least in psycholinguistics and related areas, it rarely makes sense to compute Bayes factors using (only) uninformative priors, especially on the target parameter. If one does use uninformative priors, this tends to heavily bias the Bayes factor in favor of the null (Schad et al., 2021). This bias has the consequence that the original Wang and Gelfand approach would lead to misleading conclusions (a tendency to evidence for a null effect, even when the null is very likely to be false).

In contrast to the Bayesian analysis, a conventional power analysis would be a lot less informative for the range of effect sizes we assume here. We demonstrate this point in the Appendix. The key issue with the frequentist power analysis is that once one takes the uncertainty of the estimate into account, the estimate of the power becomes so uncertain that, for planning purposes, it is all but useless.

A frequent objection that we encounter to the type of Bayesian analyses we have presented here and in other papers is that the workflow is very time-consuming compared to a frequentist analysis. This complaint about the speed with which one can complete an analysis is very important to address because this kind of attitude can completely derail a scientific research program.

Bayesian analyses using hierarchical models are generally always more time-consuming than frequentist ones. In the current context, our simulations took some 40 h, and this was with a computer with 40 physical cores and 80 logical cores. By contrast, the frequentist power analysis in the Appendix takes a mere 20 minutes on an M1 chip Macbook Pro without any special parallelization procedure—the Bayesian analysis is 120 times slower. Reacting to such radical differences in timing, many psychologists and psycholinguists have become increasingly unhappy. A remarkable expression of this sentiment was a tweet from a prominent and influential psycholinguist. The author of the tweet is a former editor-in-chief of one of the most important journals in psycholinguistics, the *Journal of Memory and Language*, and is therefore in a very influential position in the field. The tweet stated that statistical data analysis should be like going to the toilet. The essential point was that data analysis should be quick, and one should not become obsessed with it. A related point that one often hears is that one should not need to acquire much statistical knowledge either.[3]

The general sentiment expressed in comments like these is very common; the first author of the present paper has encountered many psycholinguists, psychologists, and linguists who think that (for example) hierarchical modeling is over-rated because of its excessive complexity; their proposed alternative is simpler paired t-tests and repeated-measures ANOVA (which, ironically, are also hierarchical models).

These kinds of demands for speed and easy analyses come with a cost. If one truly believes that statistics data analysis should be like going to the toilet, one should not be surprised if the end result of the analysis turns out to be crap. What actually lies behind the demand for speed and simplicity is the mistaken understanding that all one needs to look at in a data analysis is the p-value. Such a single-minded focus on the p-value is driven by a semantic drift away in psychology and other areas from the real goals of statistical data analysis. As statisticians have repeatedly pointed out (e.g., Wasserstein & Lazar, 2016; McShane et al., 2019), the goal certainly should not be quick binary conclusions based on oversimplified models. Uncertainty quantification is key to understanding data (Vasishth & Gelman, 2021), and as the van Doorn et al. (2021) article also suggests, hierarchical models are a very important tool for achieving this goal.

We turn next to the sentiment that, because time is limited, a scientist should not have the responsibility to master statistics sufficiently to be able to interpret and analyze their data. This kind of attitude has been encouraged through the cursory education in statistics that is the norm

[3] For more details, see the blog post by Phillip Alday, which responds to this tweet: https://phillipalday.com/blog/2020/12/22/statistics-is-not-shit/.

in cognitive science disciplines, and through the creation of apparently easy-to-use software, which conveniently makes some default assumptions that are kept hidden from the user. If a scientist is using a particular technical tool (here, statistics) to study a research question, they do in fact have to invest time into understanding the tool they are relying on to make statistical inferences. As Singmann et al. (2021) eloquently put it: “For statistical modeling to serve the goals of science, models cannot be based on default assumptions, but should instead be based on an understanding of their coordination function and on how they represent causal mechanisms that may be expected to generalize to other related scenarios.” To go beyond defaults, one cannot avoid taking the time to engage deeply with statistical theory.[4]

In closing, we hope that this paper, along with our two other companion papers (Schad et al., 2020; Schad et al., 2021), provides a useful starting point for researchers who wish to use complex Bayesian hierarchical models and plan sample sizes in advance using such models.

Appendix

Frequentist power analysis

For comparison with the Bayes factors analysis in the main text, we provide a comparable frequentist power analysis. One important conclusion here is that the power analysis will be essentially useless once one takes the uncertainty of the effect size into account.

In psycholinguistics, it is still extremely uncommon to do power analyses before running an experiment. A more common approach is to monitor the experiment until significance is reached; in such sequential testing scenarios, the adjustment to Type I error probability that is necessary (Pocock, 2013) is always ignored in all the psycholinguistic research we are aware of. However, power analyses have started emerging in the psycholinguistic literature (e.g., Stack et al., 2018), and could even become standard practice in the coming years.

As a quick reminder of the terminology for the reader, in the discussion below, we use the following terms: (a) **effect size**: this is defined as the estimate of the difference between the means in the two conditions of interest on the scale of interest (i.e., not standardized); (b) **power**: The probability of correctly detecting an effect δ assuming that the null hypothesis that $\delta = 0$ is false with some particular value $\delta' \neq 0$; (c) **power curve**: since power is a function of the effect size, standard deviation, and sample size, it is standard to plot the power function with respect to one or more of these variables; (d) **power analysis**: this refers to an estimate of the power, either using analytical calculations, or using simulation.

The standard simulation-based approach to computing frequentist power functions is as follows:

1. Obtain estimates for the parameters in the model, e.g., from a previous study.
2. Given a subject sample sample size n (and some sample size for items), generate simulated data based on previously obtained estimates of parameters.
3. For each sample size, repeatedly fit linear mixed models to these simulated datasets, and determine the proportion of cases where a significant effect is found. This gives the estimated power for each sample size n.
4. Plot the power curves: the power estimates against sample size, given specific values of the effect size. We will display not just one power curve but three curves, for the mean and the lower and upper confidence intervals of the effect size.

Step 1 Estimate parameters from the Grodner and Gibson (2005) data:

```
data("df_gg05_rc")
df_gg05_rc$cond <- ifelse(df_gg05_rc$condition == "objgap", 1 / 2, -1 / 2)
n <- length(unique(df_gg05_rc$subj))
## ignoring the singularity warning:
m_gg05 <- lmer(log(RT) ~ cond + (1 + cond | subj) +
  (1 + cond | item), df_gg05_rc)
summary(m_gg05)
```

Step 2 Simulate data for frequentist power analysis:

There are two ways to simulate data.

Method 1: Write a data-generation process from scratch.

[4] Phillip Alday’s keynote lecture at the Fifth Summer School on Statistical Methods in Linguistics and Psychology (SMLP) articulates this point in more detail: https://youtu.be/4E-XTvJuaaY.

```
# assumes that no. of subjects and no. of items is divisible by 2.
gen_fake_lnorm <- function(nitem = 16,
                           nsubj = 42,
                           beta = c(6, 0.12),
                           ranefsd = c(0.32, 0.22, 0.04, 0.09),
                           corr = c(.6, .6),
                           sigma_e = 0.31) {
  ## prepare data frame for two condition latin square:
  g1 <- data.frame(
    item = 1:nitem,
    condition = rep(
      letters[1:2],
      nitem / 2
    )
  )
  g2 <- data.frame(
    item = 1:nitem,
    condition = rep(
      letters[2:1],
    nitem / 2
  )
)

## assemble data frame:
fakedat <- rbind(
  g1[rep(
    seq_len(nrow(g1)),
    nsubj / 2
  ), ],
  g2[rep(
    seq_len(nrow(g2)),
    nsubj / 2
  ), ]
)

## add subjects:
fakedat$subj <- rep(1:nsubj, each = nitem)

## add contrast coding:
fakedat$cond <- ifelse(fakedat$condition == "a", -1 / 2, 1 / 2)

## Define variance covariance matrices:
Sigma_u <- matrix(c(
  ranefsd[1]^2,
  corr[1] * ranefsd[1] * ranefsd[2],
  corr[1] * ranefsd[1] * ranefsd[2],
  ranefsd[2]^2
), nrow = 2)

Sigma_w <- matrix(c(
  ranefsd[3]^2,
  corr[2] * ranefsd[3] * ranefsd[4],
  corr[2] * ranefsd[3] * ranefsd[4],
  ranefsd[4]^2
), nrow = 2)

## subj ranef
u <- MASS::mvrnorm(
  n = length(unique(fakedat$subj)),
  mu = c(0, 0), Sigma = Sigma_u
)
# item ranef
w <- MASS::mvrnorm(
    n = length(unique(fakedat$item)),
    mu = c(0, 0), Sigma = Sigma_w
  )

  ## generate data:
  N <- dim(fakedat)[1]
  rt <- rep(NA, N)
  for (i in 1:N) {
    rt[i] <- rlnorm(1, beta[1] +
      u[fakedat[i, ]$subj, 1] +
      w[fakedat[i, ]$item, 1] +
      (beta[2] +
        u[fakedat[i, ]$subj, 2] +
        w[fakedat[i, ]$item, 2]) * fakedat$cond[i], sigma_e)
  }

  fakedat$rt <- rt
  fakedat
}
```

Method 2: Use a built-in function from the designr package (Rabe et al., 2021). The advantage here is compactness of code, but the disadvantage is that the code conceals the underlying generative process. Below, we show how a single simulated dataset can be generated using designr.

```
# Create latin-square design
design <-
  fixed.factor("X", levels = c("X1", "X2")) +
  random.factor("subj", instances = 21) +
  random.factor("item", instances = 8) +
  random.factor(c("subj", "item"), groups = c("X"))
dat <- design.codes(design)
contrasts(dat$X) <- c(+1 / 2, -1 / 2)
dat$cond <- model.matrix(~X, dat)[, 2]

# simulate data
dat$ysim <- simLMM(
  form = ~ 1 + cond + (1 + cond | subj) + (1 + cond | item),
  data = dat,
  Fixef = c(6, 0.12),
  VC_sd = list(
    sj = c(0.32, 0.22),
    it = c(0.04, 0.09),
    resid = 0.31
  ),
  CP = 0.6,
  empirical = FALSE, verbose = TRUE,
  family = "lognormal"
)

## Data simulation from a linear mixed-effects model (LMM)
## Formula: gaussian ~ 1 + cond + ( 1 + cond | subj ) + ( 1 + cond | item )
## empirical = FALSE
## Random effects:
##  Groups    Name          Std.Dev. Corr
##  subj      (Intercept)   0.32
##            cond          0.22     0.60
##  item      (Intercept)   0.04
##            cond          0.09     0.60
##  Residual                0.31
## Number of obs: 672, groups:  subj, 42; item, 16
## Fixed effects:
## (Intercept)        cond
##        6.00        0.12

head(dat)

## # A tibble: 6 x 5
##   subj   item   X      cond  ysim
##   <fct>  <fct>  <fct> <dbl> <dbl>
## 1 subj01 item01 X1      0.5  914.
## 2 subj01 item02 X2     -0.5  415.
## 3 subj01 item03 X1      0.5  562.
## 4 subj01 item04 X2     -0.5  413.
## 5 subj01 item05 X1      0.5 1008.
## 6 subj01 item06 X2     -0.5  376.
```

Step 3 Compute frequentist power as a function of subject and item sample size:

```
compute_power_freq <- function(nsubj = NULL, nitem = 16,
                               b = 0.12, correlations = c(0.6, 0.6)) {
  tvalsfreq <- rep(NA, 500)
  for (i in 1:500) {
    #  print(paste("iter",i,sep=" "))
    fixefs <- c(6, b)
    dat <- gen_fake_lnorm(
      nsubj = nsubj, nitem = nitem,
      beta = fixefs, corr = correlations
    )
    mtest <- lmer(log(rt) ~ cond + (1 + cond | subj) +
      (1 + cond | item), dat,
    control = lmerControl(
      optimizer = "nloptwrap",
      calc.derivs = FALSE
    )
    )
    tvalsfreq[i] <- summary(mtest)$coefficients[2, 3]
  }
  mean(abs(tvalsfreq) > 2)
}
```

Step 4 Compute power for 42, 42 × 2, and 42 × 3 subjects (and 16 items), with a given effect size and its uncertainty estimates (in log ms):

```
tictoc::tic()
pow42freq12 <- compute_power_freq(nsubj = 42, b = 0.12)
pow42freq04 <- compute_power_freq(nsubj = 42, b = 0.04)
pow42freq20 <- compute_power_freq(nsubj = 42, b = 0.20)
tictoc::toc()
## 218.92 seconds

tictoc::tic()
pow42freq12 <- compute_power_freq(nsubj = 42*2, b = 0.12)
pow42freq04 <- compute_power_freq(nsubj = 42*2, b = 0.04)
pow42freq20 <- compute_power_freq(nsubj = 42*2, b = 0.20)
tictoc::toc()
## 391.253 sec elapsed

tictoc::tic()
pow42freq12 <- compute_power_freq(nsubj = 42*3, b = 0.12)
pow42freq04 <- compute_power_freq(nsubj = 42*3, b = 0.04)
pow42freq20 <- compute_power_freq(nsubj = 42*3, b = 0.20)
tictoc::toc()
# 579.358 sec elapsed
```

Figure 5 shows the result of the frequentist power analysis. From this power analysis, we would conclude that a sample size of 100 to 150 subjects would probably suffice if we want to achieve 80% power. This conclusion is of course conditional on the assumed effect size of 0.12 log ms (48 ms), and a Type I error probability of 0.05. In fact, as our Bayesian analysis will suggest, this power estimate is quite an optimistic one for psycholinguistic phenomena such as the English relative clause construction (see Jäger et al., 2017; Nicenboim et al., 2018; Vasishth et al., 2018; Jäger et al., 2020; Nicenboim et al., 2020, for extensive discussion). This becomes clear if one considers the uncertainty on the effect size: if one looks at the power estimates using the lower and upper bounds of the 95% confidence interval of the effect as possible alternative effect sizes, the power function shows a huge amount of uncertainty, so much so that the power analysis itself is not much use for planning purposes. Thus, it is highly problematic to compute power (as is normally done in psychology and psycholinguistics) using only a point estimate of the effect size, ignoring the uncertainty on that estimate.

Bayes factor-based analysis (an adaptation of the Wang and Gelfand approach)

Step 1 Define informative priors:

```
priors <- c(
  set_prior("normal(6, 0.6)", class = "Intercept"),
  set_prior("normal(0.12, 0.04)",
    class = "b",
    coef = "so"
  ),
  set_prior("normal(0, 0.1)", class = "sd"),
  set_prior("normal(0, 0.5)", class = "sigma"),
  set_prior("lkj(2)", class = "cor")
)

priorsNULL <- c(
  set_prior("normal(6, 0.6)", class = "Intercept"),
  # set_prior("normal(0, 0.05)", class = "b", coef = "so"),
  set_prior("normal(0, 0.1)", class = "sd"),
  set_prior("normal(0, 0.5)", class = "sigma"),
  set_prior("lkj(2)", class = "cor")
)
```

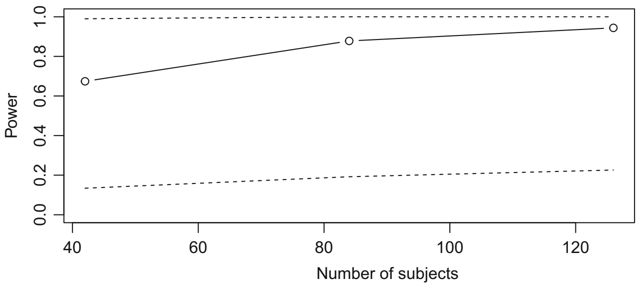


Fig. 5 The results of a frequentist power analysis; shown are different simulation-based power estimates for increasing subject sample sizes in the Grodner and Gibson (2005) experiment design. The solid line shows the power estimates for an effect size of 0.12 on the log ms scale (48 ms), and the broken lines show the power estimates for 0.04 and 0.20 log ms respectively (corresponding to 16 ms and 81 ms, respectively)

Step 2 Generate prior predictive data:

```
nsim <- 1000
Nsj <- 42 # determine no. of subjects
Nit <- 16 # determine no. of items

# prepare simulation
beta0 <- beta1 <- sigma_u0 <- sigma_u1 <-
  sigma_w0 <- sigma_w1 <-
  rho_u <- rho_w <- sigma <- NA
rtfakemat2 <- matrix(NA, nrow(expdesign), nsim)
set.seed(123)
# simulations
for (i in 1:nsim) {
  # simulate parameters
  beta0[i] <- -1
  while (beta0[i] < 0) {
    beta0[i] <- SimFromPrior(priors, class = "Intercept")
  }
  beta1[i] <- SimFromPrior(priors, class = "b", coef = "so")
  sigma_u0[i] <- SimFromPrior(priors, class = "sd")
  sigma_u1[i] <- SimFromPrior(priors, class = "sd")
  sigma_w0[i] <- SimFromPrior(priors, class = "sd")
  sigma_w1[i] <- SimFromPrior(priors, class = "sd")
  rho_u[i] <- SimFromPrior(priors, class = "cor")
  rho_w[i] <- SimFromPrior(priors, class = "cor")
  sigma[i] <- SimFromPrior(priors, class = "sigma")
  # simulate data
  rtfakemat2[, i] <- genfake(
    dat = expdesign, nsubj = Nsj,
    nitem = Nit,
    beta0 = beta0[i],
    beta1 = beta1[i],
    sigma_u0 = sigma_u0[i],
    sigma_u1 = sigma_u1[i],
    sigma_w0 = sigma_w0[i],
    sigma_w1 = sigma_w1[i],
    rho_u = rho_u[i],
    rho_w = rho_w[i],
    sigma = sigma[i]
  )
}
truePars2 <- data.frame(
  beta0, beta1, sigma_u0,
  sigma_u1, sigma_w0, sigma_w1,
  rho_u, rho_w, sigma
)
```

Step 3 Compute Bayes factors and posteriors of the effect for the simulated data (for use with multicore machines):

```
## Fit models with 10000 iterations per chain
nsim <- 100
for (t in 1:5) { # Work division
  fit_models <- function(process_id) {
    # This function will run in parallel
    # for each process id
    i <- process_id + ((t - 1) * 20)
    df.bf_gg05 <- data.frame(matrix(ncol = 5, nrow = 0))
    colnames(df.bf_gg05) <- c(
      "simulation_id",
      "estimate",
      "lower.q", "upper.q",
      "bayes_factor"
    )

    print(paste("iter", i, sep = " "))
    expdesign$fakert <- rtfakemat2[, i]
    m1 <- brm(fakert ~ so + (1 + so | subj) +
      (1 + so | item), expdesign,
    family = lognormal(), prior = priors,
    warmup = 2000,
    iter = 10000,
    cores = 4,
    save_pars = save_pars(all = TRUE)
    )
    save(m1,
      file = paste(paste(paste(paste("Full_model_",
        as.character(Nsj),
        sep = ""
      ),
      "subjects_sample_simulation_",
      sep = ""
      ),
      as.character(i),
      sep = ""
      ),
      "_10K_iter.Rda",
      sep = ""
      )
    )

    m0 <- brm(fakert ~ 1 + (1 + so | subj) + (1 + so | item), expdesign,
      family = lognormal(), prior = priorsNULL,
      warmup = 2000,
      iter = 10000,
      cores = 4,
      save_pars = save_pars(all = TRUE)
    )
    save(m0,
      file = paste(paste(paste(paste("Null_model_",
        as.character(Nsj),
        sep = ""
      ),
      "subjects_sample_simulation_",
      sep = ""
      ),
      as.character(i),
      sep = ""
      ),
      "_10K_iter.Rda",
      sep = ""
      )
    )

    df.bf_gg05[nrow(df.bf_gg05) + 1, ] <- c(
      i,
      unname(posterior_summary(m1)[2, c(1, 3, 4)]), NA
    )
    df.bf_gg05
  }
  fit_parallel <- mclapply(1:20, fit_models,
    mc.cores = 20
  ) %>% bind_rows()
  save(fit_parallel,
    file = paste(paste(paste(paste("Model_fits_",
      as.character(Nsj),

      sep = ""
    ),
    "subjects_sample_part_",
    sep = ""
    ),
    as.character(t),
    sep = ""
    ), "_batch1_10K_iter.Rda",
    sep = ""
    )
  )
}
```

Step 4 Plot the results; see Figs. 3 and 4 in the main text.

Stability issues in computing Bayes factors for larger sample sizes

The following precomputed data has repeated calculations of Bayes factors for 546 subjects, using 10,000 or 50,000 simulations. Also stored in this data frame are the mean and 95% credible intervals for the target parameter (the fixed effect slope in the linear mixed model).

```r
load("precomputed/more_vs_less_iter_expt_546subjects.Rda")
# Removing simulation 4 because of missing results
bfs <- subset(bfs, simulation_id != 4)

# Relabeling simulations
bfs$simulation_id <- ifelse(bfs$simulation_id > 3,
                            bfs$simulation_id - 1, bfs$simulation_id)
```

Figure 6 shows the result of nine simulations; shown are the distribution of log Bayes factors for 10,000 vs. 50,000 iterations when the Bayes factor was repeatedly calculated. It is clear from the figure that the (log) Bayes factor is very unstable for 10,000 iterations per chain, compared to 50,000 iterations. The implication is that for such large sample sizes, many more iterations per chain are needed than 10,000; repeated sampling to compute Bayes factors would be very time-consuming. Figure 7 shows a different visualization of the instability of the (log) Bayes factor for 10,000 vs. 50,000 iterations.

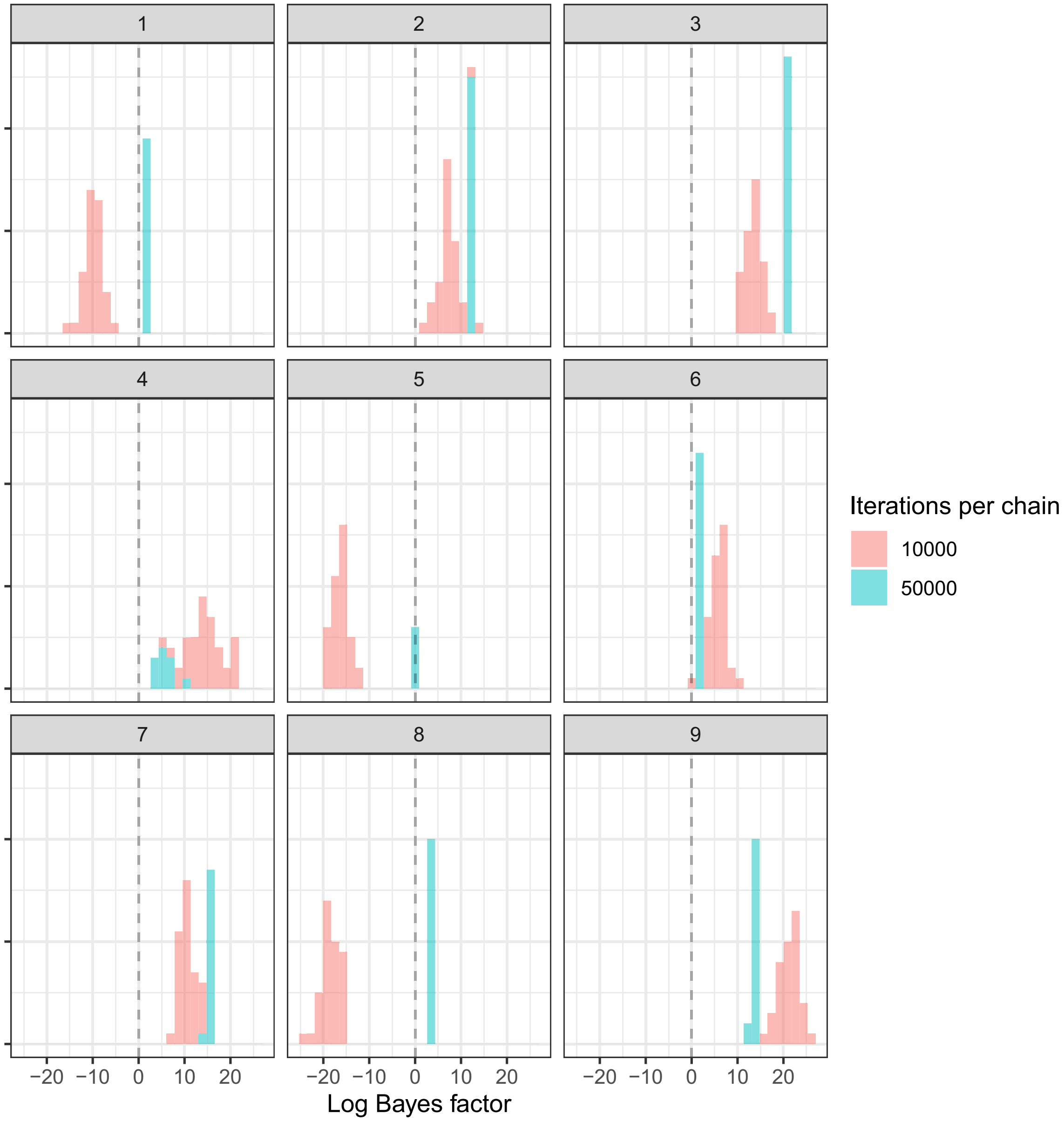
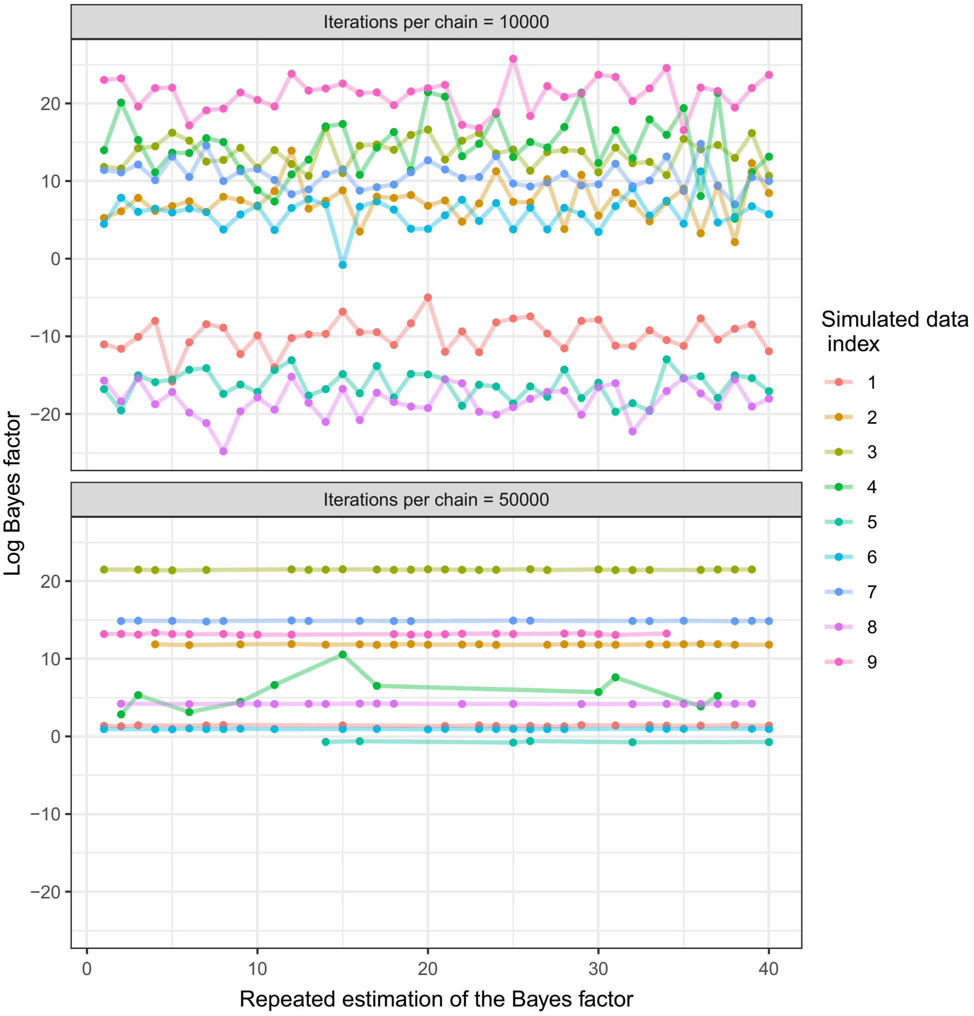


Fig. 6 The distribution of repeatedly computed (log) Bayes factors for nine simulated datasets, using either 10,000 or 50,000 iterations per chain, with subject sample size 546

Fig. 7 The stability of repeatedly computed (log) Bayes factors for nine simulated datasets, using either 10,000 or 50,000 iterations per chain, with subject sample size 546



Author Contribution SV conceived of and wrote the initial draft of the paper and the code. HY parallelized the code. BN, DS, and SV developed the informative priors for the application area discussed in this paper, and all four authors contributed to writing the paper.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287 (2021-2025), Project Q, PIs: Shravan Vasishth and Ralf Engbert.

Code Availability All the code and data associated with this paper can be downloaded from https://osf.io/hjgrm.

Declarations

Ethics Approval Not applicable

Consent to Participate Not applicable

Consent for Publication Not applicable

Conflict of Interest The authors declare no competing interests.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology*, *100*(3), 603–617.

Barr, D.J., Levy, R., Scheepers, C., & Tily, HJ (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D.M., Kliegl, R., Vasishth, S., & Baayen, H (2015). Parsimonious mixed models. Unpublished manuscript.

Bates, D.M., Maechler, M., Bolker, B.M., & Walker, S (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürki, A., Elbuy, S., Madec, S., & Vasishth, S (2020). What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language*, *114*, 104125.

Bürki, A., Alario, F.-X., & Vasishth, S (2020). When words collide: Bayesian meta-analyses of distractor and target properties in the picture-word interference paradigm. submitted.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1).

Clark, H.H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–59.

Engelmann, F., Jäger, L.A., & Vasishth, S (2020). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*, *43*, e12800.

Fedorenko, E., Gibson, E., & Rohde, D (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of memory and language*, *54*(4), 541–553.

Freedman, L.S., Lowe, D., & Macaskill, P (1984). Stopping Rules for Clinical Trials Incorporating Clinical Opinion. *Biometrics*, *40*(3), 575–586.

Gelman, A., & Carlin, J.B. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gronau, Q.F., Singmann, H., & Wagenmakers, E-J (2017). bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors. R package version 0.4-0.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–290.

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1.

Hammerly, C., Staub, A., & Dillon, B (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology*, *110*, 70–104.

Jäger, L.A., Engelmann, F., & Vasishth, S (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Jäger, L.A., Mertzen, D., Van Dyke, J.A., & Vasishth, S (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. Journal of Memory and Language, 111.

Jeffreys, H. (1939/1998). The theory of probability. Oxford University Press.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *PR*, *99*(1), 122–149.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.

Kruschke, J., & Liddell, T.M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206.

Lewandowski, D., Kurowicka, D., & Joe, H (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Lewis, R.L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

McShane, B.B., Gal, D., Gelman, A., Robert, C., & Tackett, JL (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Montero-Melis, G., van Paridon, J., Ostarek, M., & Bylund, E (2019). Does the motor system functionally contribute to keeping words in working memory? A pre-registered replication of Shebani and Pulvermüller (2013, Cortex).

Morey, R.D., Rouder, J.N., Jamil, T., & Morey, RD (2015). Package 'bayesfactor'.

Mitchell, D.C. (1984). An evaluation of subject-paced reading tasks and other methods of investigating immediate processes in reading. In D.E. Kieras, & M.A. Just (Eds.) *New Methods in Reading Comprehension Research*. Hillsdale, N.J. Erlbaum.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistics Compass*, *10*, 591–613.

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. Cognitive Science, 42.

Nicenboim, B., Vasishth, S., & Rösler, F (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. Neuropsychologia, 142.

Nicenboim, B., Schad, D.J., & Vasishth, S (2021). Introduction to Bayesian Data Analysis for Cognitive Science. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series; URL: https://vasishth.github.io/bayescogsci/.

Pregla, D., Lissón, P., Vasishth, S., Burchert, F., & Stadie, N (2021). Variability in sentence comprehension in aphasia in German. *Brain and Language*, *222*, 105008.

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.

Pocock, S.J. (2013). Clinical trials: A practical approach. John Wiley & Sons.

Rabe, M., Kliegl, R., & Schad, DJ (2021). Designr: Balanced Factorial Designs. R package version 0.1.11.

Reali, F., & Christiansen, M.H. (2007). Word chunk frequencies affect the processing of pronominal object-relative clauses. *The Quarterly Journal of Experimental Psychology*, *60*(2), 161–170.

Schad, D.J., Betancourt, M., & Vasishth, S (2020). Toward a principled Bayesian workflow: A tutorial for cognitive science. *Psychological Methods*, *26*, 103–126.

Schad, D.J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S (2021). Workflow Techniques for the Robust Use of Bayes Factors. Psychological Methods. Available from arXiv:2103.08744v2.

Schad, D.J., Vasishth, S., & Kliegl, R (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. Journal of Memory and Language, 110.

Schönbrodt, F.D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142.

Singmann, H., Cox, G.E., Kellen, D., Chandramouli, S., Davis-Stober, C., Dunn, J.C., Gronau, Q.F., Kalish, M., McMullin, S.D., Navarro, D., & et al (2021). Statistics in the Service of Science: Don't let the Tail Wag the Dog.

Spiegelhalter, D.J., Abrams, K.R., & Myles, JP (2004). Bayesian approaches to clinical trials and health-care evaluation. John Wiley & Sons.

Spiegelhalter, D.J., Freedman, L.S., & Parmar, MK (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *157*(3), 357–416.

Stack, C.MH., James, A.N., & Watson, DG (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877.

Stefan, A.M., Gronau, Q.F., Schönbrodt, F.D., & Wagenmakers, E-J (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058.

Stone, K., Nicenboim, B., Vasishth, S., & Roesler, F (2021). Understanding the effects of constraint and predictability in ERP. Neurobiology of Language. In-principle acceptance of Registered Report.

van Doorn, J., Aust, F., Haaf, J.M., Stefan, A., & Wagenmakers, E-J (2021). Bayes factors for mixed models. Computational Brain and Behavior.

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*, 968–982.

Vasishth, S., & Engelmann, F. (2022). Sentence Comprehension as a Cognitive Process: A Computational Approach. Cambridge University Press.

Vasishth, S., Chen, Z., Li, Q., & Guo, G (2013). Processing Chinese Relative Clauses: Evidence for the Subject-Relative Advantage. *PLoS ONE*, *8*(10), 1–14.

Vasishth, S., Mertzen, D., Jäger, L.A., & Gelman, A (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175.

Vasishth, S., Schad, D.J., Bürki, A., & Kliegl, R (2021). Linear Mixed Models for Linguistics and Psychology: A Comprehensive Introduction. Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series, https://vasishth.github.io/Freq_CogSci/.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. Linguistics. Accepted.

Vasishth, S., & Lewis, R. (2003). Decay and interference in human sentence processing: Parsing in a unified theory of cognition. In *Proceedings of the 16th Annual CUNY Sentence Processing Conference*. MIT, Cambridge, MA.

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, *114*(3), 830.

Wang, F., & Gelfand, A.E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Statistical Science, pp 193–208.

Wasserstein, R.L., & Lazar, N.A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *46*(2), 185–191.

Westfall, J., Nichols, T.E., & Yarkoni, T (2017). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. Wellcome Open Research, 1(23).

Yadav, H., Paape, D., Smith, G., Dillon, B.W., & Vasishth, S (2021). Individual differences in cue weighting in sentence comprehension: An evaluation using Approximate Bayesian Computation. Submitted to Open Mind.

Yarkoni, T. (2020). The generalizability crisis. The Behavioral and Brain Sciences, pp 1–37.