ANNUAL REVIEWS

Annual Review of Linguistics

Some Right Ways to Analyze (Psycho)Linguistic Data

Shravan Vasishth

Department of Linguistics, University of Potsdam, Potsdam, Germany;
email: vasishth@uni-potsdam.de

Annu. Rev. Linguist. 2023. 9:273–91

First published as a Review in Advance on November 1, 2022

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-031220-010345

Keywords

experimental linguistics, data analysis, Bayesian methods, hypothesis testing, Bayes factors, uncertainty quantification

Abstract

Much has been written on the abuse and misuse of statistical methods, including *p* values, statistical significance, and so forth. I present some of the best practices in statistics using a running example data analysis. Focusing primarily on frequentist and Bayesian linear mixed models, I illustrate some defensible ways in which statistical inference—specifically, hypothesis testing using Bayes factors versus estimation or uncertainty quantification—can be carried out. The key is to not overstate the evidence and to not expect too much from statistics. Along the way, I demonstrate some powerful ideas, including the use of simulation to understand the design properties of one's experiment before running it, visualization of data before carrying out a formal analysis, and simulation of data from the fitted model to understand the model's behavior.

1. INTRODUCTION

> If you worked in areas inhabited by demons you would be in trouble regardless of the perfection of your experimental designs.
>
> —Stuart H. Hurlbert (1984, p. 192)

Despite the title of this review, there are no clearly "right" ways to analyze data. Statistical data analysis is an inherently subjective process, and it would not be unusual for two statisticians to analyze the same data very differently and even come to different conclusions and decisions. Yet, both approaches could, at least technically, be correct. Nevertheless, there are some basic principles that come from best practice in statistics that can improve the quality of our statistical inferences. Every subfield has its own particular sets of commonly used statistical models; in linguistics, the modern standard is the linear mixed model (LMM), which is also referred to as the hierarchical model (Pinheiro & Bates 2000). Accordingly, in this review, I focus on this modeling framework and discuss both frequentist and Bayesian versions of the hierarchical model.

In what follows, I assume that the reader has a basic knowledge of the *t* test and type I and II errors and has some experience with the LMM (Bates et al. 2015). If the reader lacks this background, introductory articles such as those by Baayen et al. (2008), Vasishth & Nicenboim (2016), and Vasishth et al. (2018) would be good starting points. Other, more comprehensive textbook references are provided in the Literature Cited.

Experimental science is more than careful experiment designs and the use of sophisticated methods like event-related potentials and eye tracking. There are six components in an experiment: (*a*) setting up a research hypothesis and developing clear, testable predictions; (*b*) designing the experiment; (*c*) implementing it in software and running it; (*d*) preprocessing the data; (*e*) statistical analysis; and (*f*) interpreting the results of the analysis. It is far from trivial to execute steps *a*–*d*; but even if one is perfectly able to carry these steps out, if one does not draw valid statistical inferences from the data analysis, there is the potential to seriously mislead oneself. Because I was asked to write an article on data analysis, I focus here on these last two steps.

2. AN EXAMPLE: RELATIVE CLAUSE PROCESSING

To make the discussion concrete, I focus here on a simple example of a research question: Are object relative clauses harder to process than subject relatives? This seems like a simple question with an easy prospect for a clear answer, but I show below that there are important issues to consider before making any decisive claim.

In this discussion, I consider published data from English and Chinese. The data are from Grodner & Gibson (2005) and Gibson & Wu (2013). I use these data as they are two of the relatively few published data sets that happen to be available and because they have the simplest possible design (comparison of two conditions' means). My goal is not to single out the data from these particular authors. I could have illustrated the problems I discuss here using my own data, but my own experiments are usually designs with more than two conditions, which make the presentation unnecessarily complex.

English relative clauses are shown in examples 1a and b. The vertical bars in the example sentences show the partitioning of the regions of interest when a method like self-paced reading is used. Work on English relatives has consistently shown that, at the relative clause verb, subject relative clauses are read faster than object relatives (e.g., Fedorenko et al. 2006, Gibson et al. 2005, Gordon et al. 2001, Grodner & Gibson 2005, Just & Carpenter 1992):[1]

[1]There are some important design problems in the experiment that Just & Carpenter (1992) carried out; this design was used in subsequent studies. Some important problems here are that the relative clause verb is not

(1a) The senator | who | **interviewed** | the journalist | resigned.

(1b) The senator | who | the journalist | **interviewed** | resigned.

In contrast to English, Chinese relatives (see examples 2a and b below) have prenominal relative clauses; in English, relative clauses appear postnominally. This difference in the position of the relative clause has the interesting consequence that the distance between the gap in the relative clause and the head noun modified by the relative clause is longer in subject relatives than in object relatives. Hsiao & Gibson (2003) and Gibson & Wu (2013) argue that this increased gap distance in subject versus object relatives leads to longer reading times at the head noun in subject relatives. Compare this with English (examples 1a and b above), in which the distance between the head noun and the gap in the relative clause is longer in object relatives, leading to longer reading times at the relative clause verb in object versus subject relatives. Thus, English and Chinese are expected to show opposite patterns: a subject-relative advantage in English, and an object-relative advantage in Chinese:

(2a) yaoqing | fuhao | de | **guanyuan** | xinhuaibugui
invite | tycoon | REL | official | have bad intentions
'The official who invited the tycoon had bad intentions.' (Subject relative clause)

(2b) fuhao | yaoqing | de | **guanyuan** | xinhuaibugui
tycoon | invite | REL | official | have bad intentions
'The official who the tycoon invited had bad intentions.' (Object relative clause)
(Gibson & Wu 2013, p. 134)

Is there evidence for these predicted patterns in English versus Chinese?

In psycholinguistics, it is commonly assumed that one can just run a self-paced reading study with some 40 or so participants and multiple items, and if the statistical test is "significant," one has a definitive answer to the research question. This assumption is false.

To summarize an important takeaway from this review, a single experiment is almost never going to give decisive evidence regardless of whether the effect is considered significant; running higher-powered studies and demonstrating replicability are critical to making discovery claims.

As I show below, obtaining a decisive answer to our research question is rather more involved than running relatively small sample studies and then carrying out a *t* test or analysis of variance (ANOVA) on them. If we want clear answers, we need to invest much more time and money than we normally do in psycholinguistics. This seems like an unpleasant message, but it is the reality, and the sooner researchers confront it, the sooner their statistical inferences will become defensible. In the subsections below, I first summarize the published statistics in the original articles (Gibson & Wu 2013, Grodner & Gibson 2005) and then turn to how one can carry out informative studies that can actually answer the research question.

2.1. The Original Analyses of the English Data

A critical region of interest in Grodner & Gibson's (2005) article for which there are clear theoretical predictions is the embedded verb inside the relative clause. The reported statistical analyses show strong evidence against there being no difference between the two conditions.

in the same position in the two conditions, and the precritical region is different. However, I ignore these confounds in the design here, noting that these can be mitigated by, for example, comparing the entire relative clause region, as done by Fedorenko et al. (2006).

The estimates are in the predicted direction (object relatives are harder to process than subject relatives: 422 versus 355 ms, a 67-ms difference):[2]

> Planned comparisons between the two conditions revealed significant differences at the embedded verb, $t1(1, 41) = 11.9, \ldots p < 0.001$; $t2(1, 15) = 14.3, \ldots, p < 0.01$. (Grodner & Gibson 2005, p. 269)

Although the authors did not carry out a MinF′ test (Clark 1973), the published statistics allow us to compute the MinF′ statistic, which is $\text{MinF}'(1,51) = 83.67$, $p = 2.46 \times 10^{-12}$; this seems to be strong evidence against the null hypothesis of no difference.

Type M error: the extent to which the true effect size is overestimated, given a significant result (expressed as a ratio)

Type S error: the probability of observing an effect with the incorrect sign, given a significant result

2.2. The Original Analyses of the Chinese Data

For Chinese, the critical region was the head noun. Gibson & Wu (2013, p. 142) write the following:

> [T]he head noun for the RC [relative clause]. . .was read more slowly in the SRC [subject-extracted relative clause] condition [$F_1(1, 36) = 6.92, \ldots p = .01$; $F_2(1, 14) = 4.62, \ldots p < .05$].

The authors did not carry out the MinF′ analysis, but computing the MinF′ statistic shows that the published claim is not statistically significant: $\text{MinF}'(1,33) = 2.77$, $p = 0.11$. Thus, the MinF′ computation seems to furnish no evidence against the null hypothesis.

In the remainder of this review, I revisit the relative clause question and illustrate some of the best practices for planning and conducting experimental studies.

3. PLANNING FUTURE STUDIES ON ENGLISH AND CHINESE RELATIVE CLAUSES

Suppose now that we are planning a future set of studies to investigate the claims for English and Chinese. Because the published data on English and Chinese relative clauses are easy to obtain [Gibson & Wu (2013) and Grodner & Gibson (2005) generously made their data publicly available], one can use estimates from these existing data to plan a future set of studies. If such data are not available, one can either derive estimates of parameters from previously published work through meta-analysis (Bürki et al. 2020, 2022; Cox et al. 2022; Jäger et al. 2017, 2020; Mahowald et al. 2016; Nicenboim et al. 2018a, 2022; Vasishth et al. 2013) or carry out a preliminary study to plan for a future study (Nicenboim et al. 2018b). To conserve space, I do not show the R code used in this review, but all the examples shown here can be reproduced using the accompanying code and data.

I begin by assuming that the researcher is working within the framework of frequentist null hypothesis significance testing (NHST). Below, I discuss alternatives to NHST, specifically Bayes factors and estimation. I assume here that the reader is familiar with the definitions of type I and II errors.

When planning a study, it is important to set a sample size that gives one reasonably high statistical power. Why is it so crucial to aim for high power? The short answer is that type M error and type S error make even significant effects uninformative (Gelman & Carlin 2014). I explain this point next.

[2]The difference I report below (see Section 4.3.3) will not match the published estimates because I did not follow the analysis methodology used by the original authors.

3.1. Why Prospective Power Analysis Is So Important

When statistical power is low, the most obvious problem is a high probability of failing to reject the null hypothesis (Hoenig & Heisey 2001). As discussed by Vasishth & Gelman (2021), this problem has real, practical consequences for linguistics; if power is low, even if one repeatedly gets null results across multiple experiments, this does not imply that one has found evidence in favor of the null. The field is full of incorrect statistical inferences based on such null results from underpowered studies (e.g., Logacev & Bozkurt 2021, Pankratz et al. 2021).

There is another, more insidious effect that low power has: Statistically significant effects will tend to come from exaggerated estimates of the effect of interest. If one obtains such a significant effect and tries to replicate the study, the effects will generally not be replicable. This issue has been discussed repeatedly in the statistics literature (Hedges 1984, Lane & Dunlap 1978) but has not reached linguistics or psychology. Other names for type M error are the "winner's curse" and the "vibration of effects" (Button et al. 2013) and the "vibration ratio" (Ioannidis 2008).

I turn next to an exemplary design and power analysis for a hypothetical future study on relative clauses in English and Chinese. An important point to stress is that I am not using power analysis to draw inferences about the existing studies; that kind of analysis is called a post hoc power analysis and would be pointless to carry out because once the experiment is done and the *p* value has been computed, power is a function of the *p* value (Hoenig & Heisey 2001). When I talk about prospective power, I am talking about power as it relates to a future study. I use existing data for this purpose, but the goal is to understand the properties of the design given what we know so far.

3.2. Design and Power Analysis for Planning a Future Study Given Existing Data

Psychologists and statisticians have repeatedly pointed out (Cohen 1962, 1988; Gelman & Carlin 2014; Moerbeek & Teerenstra 2015) that this kind of design analysis can and should be done when planning a future experiment. However, power analysis has largely been ignored in linguistics. What would such a power analysis look like?

3.2.1. Power estimation using simulation. Power can be estimated by carrying out the following steps.

1. Fit an LMM to the existing data and extract all parameter estimates (see **Table 1** for the estimates from the English and Chinese data).
2. Use the parameter estimates to generate simulated data repeatedly.
3. Test for significance in each simulation run; the proportion of significant results is the estimated power, under the assumption that the estimates from the current study reflect the reality (this is an unrealistic assumption given that a low-powered study can give overestimates; accordingly, the power estimates should be treated as an optimistic upper bound).

Below, I show what these steps yield. But first, a cautionary note. The above steps rely on existing data, but it is crucial to understand that the intention here is not to draw inferences about the power properties of the existing data—this is called post hoc or observed power—but rather to plan a future study. That is, the goal is prospective power. Researchers often mistakenly draw inferences about the power properties of their already-conducted study; that is, they compute observed power. As discussed by Hoenig & Heisey (2001), this is a pointless exercise: Post hoc power is simply a one-to-one function of the observed *p* value. Observed power furnishes no new information about the already-conducted study. Despite this well-known problem with observed power, psychologists often report such meaningless statistics, usually to argue that their null results

Table 1 Parameter estimates (with standard errors for the fixed effects) from the linear mixed models fitted to the English and Chinese relative clause data

	English	Chinese
Fixed effects		
(Intercept)	5.883 (0.05)	6.062 (0.07)
Cond	0.124 (0.05)	−0.07161 (0.05)
Random effects		
SD: subject (intercept)	0.318	0.245
SD: subject cond	0.221	0.112
Cor: subject (intercept) cond	0.58	NE
SD: item (intercept)	0.036	0.181
SD: item cond	0.081	NE
SD: residual	0.361	0.515
Number observed	672	547
Number of groups: subject	42	37
Number of groups: item	16	15

In the table, cond refers to the sum-coded predictor, relative clause type, with subject relatives coded −0.5 and object relatives +0.5; SD refers to the standard deviation; Cor refers to the correlation between random intercepts and random slopes; and NE implies that the parameter in question was not estimated because of convergence problems. English data from Grodner & Gibson (2005, Experiment 1). Chinese data from Gibson & Wu (2013).

are meaningful. Some examples of papers that report observed power to argue that their null results are interpretable are those by Gordon et al. (2004) and Berman et al. (2009).

Once we are clear about the intention behind using previous data for a power analysis (planning the sample size for a future study), we can safely proceed to compute power.

Usually, the primary parameter of interest in an LMM is the fixed effects slope. In the relative clause example, the slope would represent (under an appropriate sum-contrast coding; Schad et al. 2020b) the difference in means between the two conditions.

In the power analyses of the English and Chinese data sets (**Table 1**), in Grodner & Gibson's (2005) data, the intercept and slope on the log-millisecond scale are approximately 6 and 0.12 log ms, respectively, and the standard error of the slope is 0.05 log ms. If we take these estimates as an initial guess at the range of plausible effect sizes, the effect can be assumed a priori to approximately range from 8 to 89 ms (for details on how to obtain these estimates in milliseconds, see Nicenboim et al. 2022).

Similarly, in the Chinese data (Gibson & Wu 2013), the intercept and slope are approximately 6 and −0.07 log ms, and the standard error of the slope is approximately 0.05 log ms. This implies that, as theory predicts, in Chinese there is a pattern consistent with an object relative advantage. The effect size in milliseconds is −31 ms with a 95% confidence interval of −72 to 10 ms. These are tentative estimates as they are based only on one study; one could (and should!) do a proper meta-analysis and come up with better estimates for Chinese (for an initial attempt, see Vasishth et al. 2013).

3.2.2. Results of the prospective power analyses. **Figure 1** shows the distribution of power for sample sizes of 50, 100, 200, and 300 participants and 16 items given the parameter estimates from Grodner & Gibson's (2005) English Experiment 1 and from Gibson & Wu's (2013) Chinese experiment. The bigger spread in power estimates for Chinese compared with English comes from the fact that the Chinese data are much noisier (e.g., the estimated residual standard deviation in

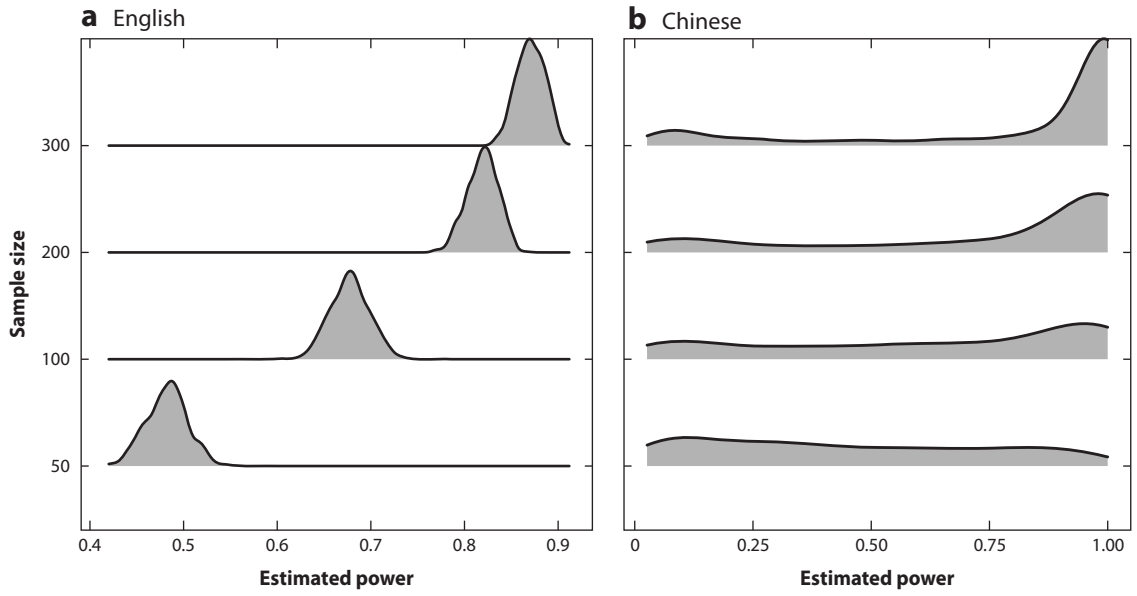

Figure 1

Estimated statistical power using simulation for the (*a*) English and (*b*) Chinese relative clause data. Each power distribution is generated by simulating data repeatedly from an assumed effect size of 0.12 (standard error: 0.05) log ms for English and an assumed effect size of −0.07 (0.05) log ms for Chinese. All other parameters (the variance components, correlation) are assumed to be point values. The uncertainty in the power calculation stems from the uncertainty about the assumed effect size (the fixed effects slope), which represents the mean difference in reading time between the two relative clause types as well as the uncertainty due to the variance components (the random effects). The bigger spread in the power estimates in Chinese comes from the fact that the data that the power analysis is based on were much noisier than in English (e.g., in **Table 1**, compare the residual standard deviations in Chinese versus English: 0.52 versus 0.36).

Chinese is 1.5 times larger than in English). It is clear from this plot that if we want to be reasonably sure that we have at least 80% power, we will need at least 300 participants for this design. The original studies had sample sizes of 42 (Grodner & Gibson 2005) and 37 (Gibson & Wu 2013); such sample sizes would lead to severely underpowered studies.

Thus, if planning future studies on English and Chinese, and even if one optimistically assumes that the true effect sizes are those observed in the above two studies, the sample sizes needed to detect the effects with statistical power at approximately 80% would be much larger than the sample sizes commonly used in such experiments. A crucial point to keep in mind is that even with the larger sample size, the uncertainty about the power achieved—which comes from the uncertainty about the effect size—will remain. Despite this uncertainty, a larger-sample study would be a huge improvement over these two small-sample experiments.

Notice that in the above power analyses, only the uncertainty of the fixed effects predictor was taken into account, not the uncertainty associated with the variance components and correlation. If one were to take all that uncertainty into account, the power distribution would become even wider (even more uncertain). Incidentally, one can carry out a Bayesian version of a power analysis using Bayes factors and observe similar results (for detailed discussion and example code, see Vasishth et al. 2022b).

In summary, despite the uncertainties inherent in power analysis, it is nevertheless a useful tool for planning sample sizes when one is committed to working within the NHST paradigm. Even if one ends up running a small-sample study because of time or resource limitations, such power

analyses can be useful for understanding how strong one's conclusions can be once the data come in. If one has no choice but to report a low-powered study's findings in a paper, then the claims have to be tempered accordingly (see, e.g., Vasishth & Gelman 2021).

4. AFTER THE DATA ARE COLLECTED

Once the data are in, the first step should be to visualize the data and only then to carry out the statistical analysis. The visualization serves two important purposes.

First, a box plot or the like will reveal any extreme or potentially influential values. The mean can be extremely sensitive to extreme values, making a nonsignificant difference come out significant. An example is Gibson & Wu's (2013) data: If we remove just two extreme data points from the data set consisting of 547 data points, the effect becomes nonsignificant. Gibson & Wu's (2013) paper reported this one effect as significant; just plotting the data before analyzing can stop us from such overenthusiastic reporting.[3]

Second, individual differences in the effect of interest should be visualized to get a sense of whether random slopes should be included in the model. Often, such a visualization already makes it clear what the random effects structure of the LMM should look like. Formal model comparison methods exist, but these are all completely focused on statistical significance testing. As discussed above, NHST makes no sense at all unless statistical power is high, and high statistical power is a luxury we rarely enjoy in linguistics (see Section 3).

4.1. Visualize Data Before Analyzing Them

Figure 2 shows a box plot for the English and Chinese data. It is quite striking that the variability in one condition is larger than in the other: In English, the object relative condition has larger variance, and in Chinese, the subject relative condition does. What useful information do these plots deliver? Here are some insights from **Figure 2**.

1. The difference between relative clause types in English and Chinese might have to do with differences in the variance between the two conditions rather than (just) the difference in means. This heterogeneity in variance can have important consequences for statistical inference, especially when—as Grodner & Gibson (2005) and Gibson & Wu (2013) did—*t* tests or repeated-measures ANOVA are carried out (Schad et al. 2022b). Another possibility that the figure raises is that both the English and Chinese data might be generated not from a single distribution but from a hierarchical finite mixture distribution, such as a mixture of lognormals (Vasishth et al. 2017).
2. Even if one ignores the difference in variance between the two conditions, the extreme values could unduly influence the mean difference. As I show below, in Chinese the statistically significant effect (object relatives being easier to process than subject relatives) reported by Gibson & Wu (2013) is determined by only two extreme data points in subject relatives, out of a total of 547 data points.

The by-participant individual differences in the relative clause effect (OR − SR difference in milliseconds) in English and Chinese are shown in **Figure 3**. These individual-level plots are not the shrunken estimates from the LMMs (Bates et al. 2015) but rather use the data from each

[3] Researchers often use automated trimming procedures to remove potentially influential data points; this kind of automated data deletion is not something any statistician would do. Moreover, in some cases, research groups do not apply this automated procedure consistently within a given analysis.

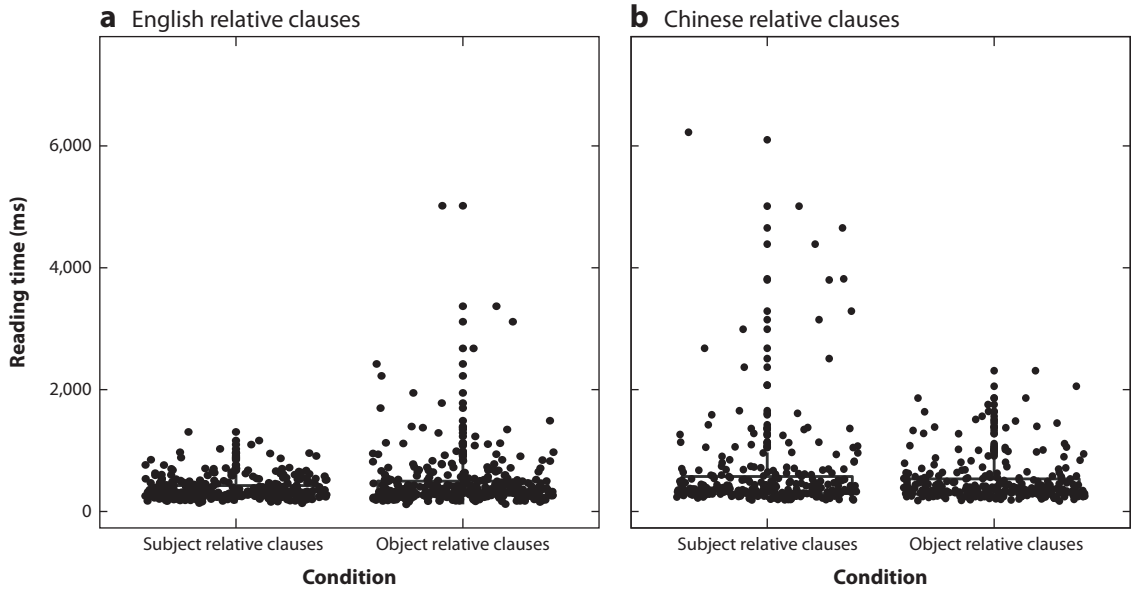

Figure 2

Box plots showing the distribution of (*a*) Grodner & Gibson's (2005) Experiment 1 data on English subject and object relative clauses and (*b*) Gibson & Wu's (2013) data on Chinese relative clauses. Shown are reading times (in milliseconds) by condition at the critical region (the relative clause verb).

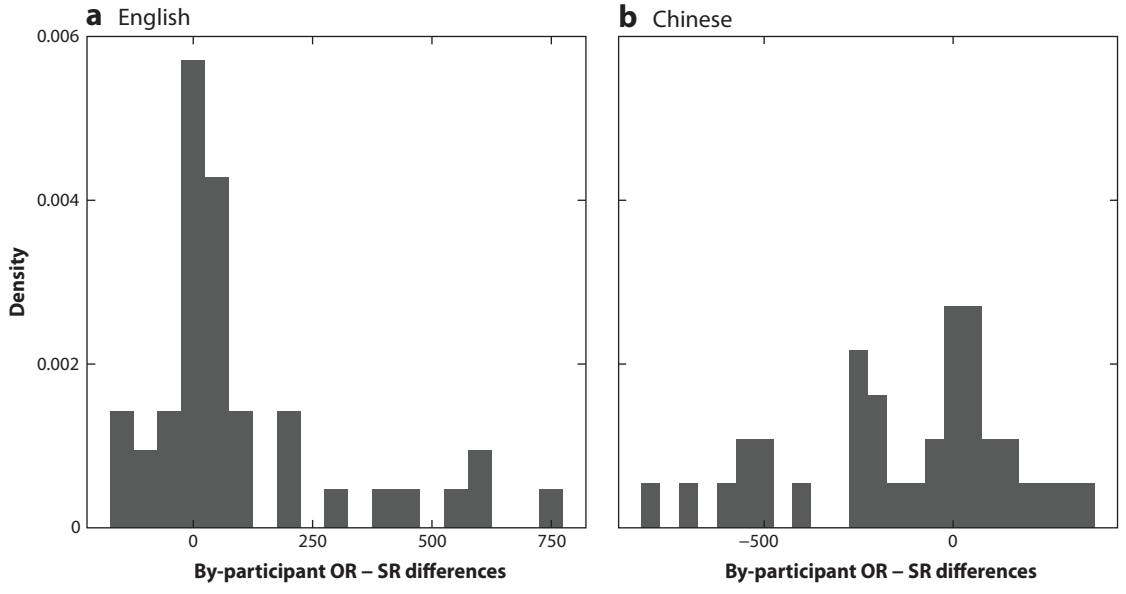

Figure 3

Histograms showing the distribution of differences in OR − SR reading times in milliseconds between participants in the (*a*) English and (*b*) Chinese data. It is important to notice that a few participants (in both the English and Chinese data) have extreme values, which will heavily bias (overestimate) the overall estimated effects in the two languages. This is type M error (Gelman & Carlin 2014) in action. Abbreviations: OR, object relative; SR, subject relative.

subject considered independently (from the so-called no-pooling model; Gelman & Hill 2007, Nicenboim et al. 2022, Vasishth et al. 2022a). What do we learn from this plot?

A major takeaway from the histograms in **Figure 3** is that a few participants with extreme effect estimates are heavily skewing the mean OR − SR effect. These extreme estimates will lead to biased estimates.

Apart from the above descriptive observations, these plots show considerable variation between participants, suggesting that by-participant intercepts and slopes will probably be needed in the models. One could draw similar by-item plots (omitted here to conserve space).

These figures are relevant only for the reading time data discussed here; because of space limitations, visualizations for different types of data cannot be shown in this review. Wilke (2019) discusses examples of good-quality data visualizations.

4.2. Statistical Inference

Drawing inferences from the data requires that we specify a statistical model; deciding on an appropriate model for one's data is a subjective step. Even with seemingly simple statistical tests like the *t* test, much can go wrong if the model assumptions are not met. For example, in the one-sample *t* test, violating the normality and independence assumptions can lead to invalid inferences from hypothesis tests. It is not unheard of for researchers to fit a *t* test to binary data; this amounts to assuming that a normal distribution is generating zeros and ones (the appropriate distributional assumption would be a Bernoulli). Statistical software generally assumes that the researcher knows what they are doing and returns no warning if the model assumptions are not met, so it is easy to go wrong if one treats statistical tests like automated procedures (for examples of incorrect uses of the *t* test, see Nicenboim et al. 2018a, Vasishth et al. 2022a). With LMMs fitted to reading time data, violations of the normality assumption can dramatically change the inferences we draw from the data and model. For example, in the Chinese data, if we fit the LMM to raw reading times (using the normal likelihood), then the effect comes out significant ($t = -2.15$), but if we remove the two extreme values in the subject relative conditions (see **Figure 2**), the *t* value suddenly becomes nonsignificant ($t = -1.76$). The log-transformed analysis shown in **Table 1** is unaffected by the extreme values because the log transform downweights the two influential values.

We already have seen in **Table 1** what the estimates from a frequentist LMM were for the English and Chinese data. If one were doing a hypothesis test using these frequentist model estimates, the standard conclusion would be that we have clear evidence for the English relative clause effect but not for Chinese. As indicated above, both conclusions would be misleading because of the danger of type M error arising from underpowered studies.

In this section, I discuss a more nuanced way to work with such data sets. I focus on statistical inference using Bayesian hierarchical (linear mixed) models and on two different ways of thinking about inference: estimation and hypothesis testing. The Bayesian approach is chosen here because—as I demonstrate below—it is more conservative and more flexible than standard NHST, and it directly answers the research question itself (instead of rejecting a straw-man null hypothesis). Bayesian modeling also allows us to focus on quantifying the uncertainty regarding the effect of interest instead of talking about hard binary distinctions like "effect present/absent."

4.3. Bayesian Hierarchical Models

Over the last decade or so, it has become relatively easy to fit Bayesian hierarchical (aka linear mixed) models using the programming language Stan (Carpenter et al. 2017, Stan Dev. Team 2022). Standard LMMs that linguists are used to fitting with the package `lme4` can now easily be fitted using the front end to Stan, `brms` (Bürkner 2017), which uses a very similar syntax.

The real barrier to using Bayesian models in research is not the mathematical or computational complexity but rather the change in perspective that is needed.

4.3.1. Some important ideas in Bayesian methodology. In frequentist modeling, the data are random and the parameters are fixed, unknown point values. This means that the statistical inferences are based on data that we did not collect, and the statistical test (e.g., *t* test, chi-square test, *F* score) quantifies evidence in terms of what could have happened hypothetically in the data assuming that some null hypothesis is true; the focus is not on the research hypothesis but on how improbable the test statistic is in some imaginary, counterfactual world of infinite replications, given the null hypothesis. Frequentist NHST does not tell us anything directly about the research hypothesis of interest (Wasserstein & Lazar 2016); it only tells us what the evidence against the null is. In this sense, although NHST answers a question, it answers the wrong one.

By contrast, in the Bayesian framework, the data are considered to be fixed—you get what you get.[4] In Bayesian statistics, it is the parameters that are random variables; parameters have probability distributions associated with them. Thinking about parameters as random variables has far-reaching implications: Now we no longer talk about "the" relative clause effect (object minus subject relative clause processing difference) as if it were some invariant, unknown point value like 50 ms "out there in nature" (the reader will probably agree that it would be absurd to think about an effect as an invariant point value, but that is in fact the assumption in frequentist modeling). In Bayesian statistics we talk about the relative clause effect as a probability distribution. As a hypothetical example, we might believe (based on prior data, on theory, or on computational modeling; for how such prior information can be derived, see O'Hagan et al. 2006, Nicenboim et al. 2022) that, in self-paced reading data, the relative clause effect might be Normal($\mu = 50, \sigma = 10$) on the millisecond scale. This kind of statement asserts that we believe a priori (before the data from our experiment come in) that we are 95% certain that the true value of the relative clause effect lies between 30 and 70 ms; the range [30,70] ms is often called a 95% credible interval. This kind of prior knowledge/belief (which, as mentioned above, can be derived using expert elicitation, computational modeling, meta-analysis, etc.) can then be included in the Bayesian LMM to compute something called the posterior distribution of the relative clause effect, which gives the updated probability distribution of the relative clause effect after seeing the data. In other words, a critical advantage of the Bayesian paradigm is that we have the opportunity to formally build on prior knowledge.

Users of frequentist methods are not accustomed to thinking about and using prior knowledge in data analysis, but it is standard practice in areas like medicine (Higgins & Green 2008) to derive a quantitative summary of what is known so far and to use that knowledge in future analyses (Spiegelhalter et al. 1994). Such evidence synthesis has examples in psycholinguistics as well (Bürki et al. 2020, 2022; Cox et al. 2022; Jäger et al. 2017; Mahowald et al. 2016; Nicenboim et al. 2018a, 2020; Vasishth et al. 2013). These kinds of meta-analyses are very helpful in deriving prior distributions for future studies (Nicenboim et al. 2022, Vasishth & Engelmann 2022).

The end product of a Bayesian analysis is a probability distribution on the parameter (more precisely, the joint distribution of the parameters in the model); all the inferences about the research problem are made based directly on this information, not via the properties of imaginary replications of the data as in the frequentist approach. A concrete example will help.

[4]One can of course think about the consequences of what would happen under hypothetical repeated sampling even in the Bayesian context; in other words, we can ask ourselves what would happen if the data were random as well (for detailed discussion, see Schad et al. 2020a, 2022a; Vasishth et al. 2022b).

The next section presupposes that the reader is familiar with LMMs. For a detailed exposition of such models in the context of psycholinguistic studies, readers are referred to Vasishth et al. (2022a) and Nicenboim et al. (2022).

4.3.2. A Bayesian analysis of the relative clause data. Suppose that we rerun the LMMs discussed above; this time, we use the Bayesian framework. Here is the formal statement of the LMM for both English and Chinese:

$$rt_{ij} \sim \text{LogNormal}[\alpha + u_{0i} + w_{0j} + (\beta + u_{1i} + w_{1j}) \times so_{ij}, \sigma],$$

where

$$\begin{pmatrix} u_0 \\ u_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u\right), \quad \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_w\right)$$

$$\Sigma_u = \begin{pmatrix} \sigma_{u0}^2 & \rho_u \sigma_{u0} \sigma_{u1} \\ \rho_u \sigma_{u0} \sigma_{u1} & \sigma_{u1}^2 \end{pmatrix} \quad \Sigma_w = \begin{pmatrix} \sigma_{w0}^2 & \rho_w \sigma_{w0} \sigma_{w1} \\ \rho_w \sigma_{w0} \sigma_{w1} & \sigma_{w1}^2 \end{pmatrix}.$$

The `lme4` syntax for this model is as follows:

```
lmer(log(rt) ~ so + (1+so|subj) + (1+so|item),dat).
```

There are nine parameters in the model (α, β, σ_{u0}, σ_{u1}, ρ_u, σ_{w0}, σ_{w1}, ρ_w, σ), and each parameter gets a prior distribution defined for it. Below, I define so-called regularizing priors for the parameters:

$$\begin{aligned} \alpha &\sim \text{Normal}(6, 0.6) \\ \beta &\sim \text{Normal}(0, 0.1) \\ \sigma_u, \sigma_w, \sigma &\sim \text{Normal}(0, 0.5) \text{ where } \sigma_u > 0 \\ \rho_u, \rho_w &\sim \text{LKJ}(2) \end{aligned}$$

The prior on the intercept α implies that the mean reading time can range from 122 to 1,339 ms with a probability of 0.95. The prior on the β parameter implies that the relative clause effect can range from -81 to 81 ms with a probability of 0.95. This is a mildly informative prior; what this prior expresses is agnosticism about the sign of the relative clause effect, but it also assumes that the effect is not likely to be very large. For an empirically based justification for such a mildly informative prior, readers are referred to Nicenboim et al. (2022, chapter 6).

The standard deviations have truncated standard normal distributions as priors (truncated at 0 because standard deviations cannot be negative), and the correlations have a so-called LKJ prior whose parameter, 2, downweights extreme correlations like ± 1.

4.3.3. Results of the Bayesian analysis: using estimation. The posterior distributions of the relative clause effect for English and Chinese are shown in **Figure 4**. These posteriors directly answer our research questions for English and Chinese. The estimates of the English relative clause effect are 35 ms, 95% credible interval [2, 69] ms, and for Chinese, -24 ms, [-66, 17] ms.

It is possible to draw our conclusions using just these estimates and their uncertainties (Kruschke 2010, Kruschke & Liddell 2018). It is clear that the posterior distributions are consistent with the qualitative claim that object relatives will be harder in English and easier in Chinese compared with the respective baseline condition; however, the 95% credible intervals show that there is quite a lot of variability possible in the estimates. This high variability, or low precision of the estimate, is very informative because it is an indication that we have relatively sparse data.

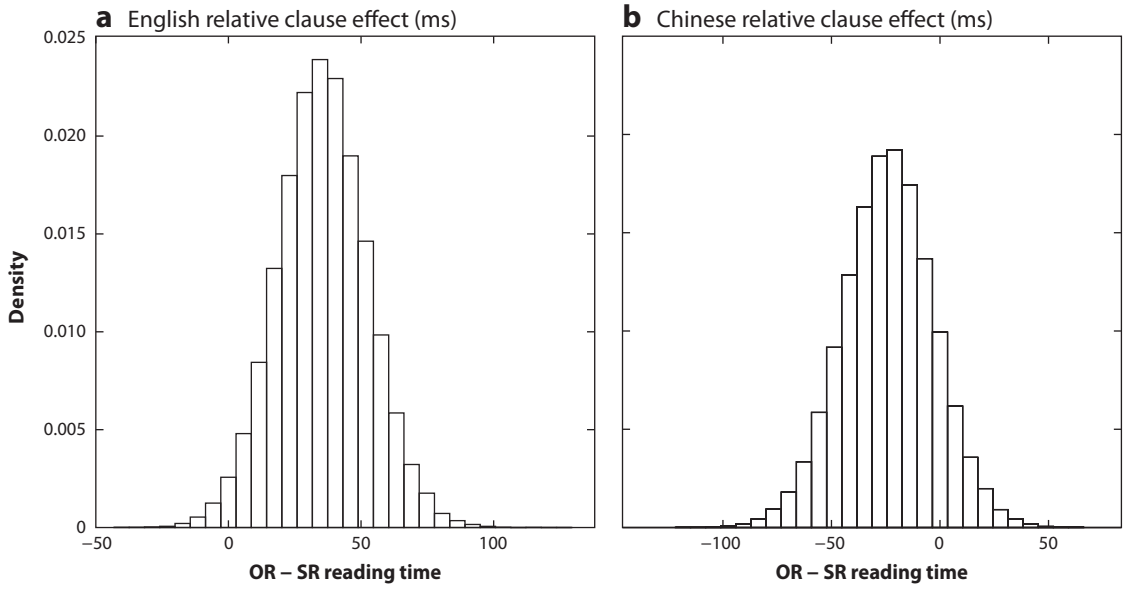

Figure 4

The posterior distributions of the relative clause effect (OR − SR) in milliseconds at the critical region: (*a*) the relative clause verb in English and (*b*) the head noun in Chinese. These posterior distributions give us estimates of plausible values of this effect, given the Bayesian linear mixed models and the data at hand. Abbreviations: OR, object relative; SR, subject relative.

Given this low precision, no strong conclusions can be drawn about these effects from these data (as discussed above, this is regardless of whether the effect is considered significant or not under a frequentist analysis).

Now, if we want to go further and find out whether there is evidence for a relative clause effect in English and Chinese (i.e., if we want to make a discovery claim), we will have to do a formal hypothesis test; we will need to compute the Bayes factor (Kass & Raftery 1995).

4.3.4. Results of the Bayesian analysis: using Bayes factors for hypothesis testing. In essence, the Bayes factor compares the likelihood (more precisely, the marginal likelihood) of the baseline model (the so-called null model) against the likelihood based on some alternative model. The null model could be that the parameter β, which represents the difference between the two relative clause types, is 0 log ms, and the alternative could be that β is Normal($\mu = 0$, $\sigma = 0.1$) on the log-millisecond scale. A powerful property of the Bayes factor is that the null and alternative models can be any competing models (e.g., Rouder & Haaf 2021); one is not restricted to assuming a simple point-value null hypothesis. For example, for English, one could compare a null model that assumes that the effect is a priori Normal(0, 0.01) on the log-millisecond scale (this corresponds to the 95% credible interval [−8,8] on the ms scale) with an alternative model that assumes that the effect is, say, Normal(0.02, 0.01) in the English data (I illustrate the use of such a null hypothesis below).

The end result of a Bayes factor analysis is the relative likelihood of the two models being compared, presented as a ratio. For example, when comparing a null model with the alternative, if the ratio is 3, this means that the null model is three times more likely than the alternative, given the prior on the parameter of interest. The order in which the model comparison is done determines how the Bayes factor is interpreted; for example, if we were comparing the alternative

with the null, then the Bayes factor mentioned above would be $\frac{1}{3}$. For this reason, when reporting Bayes factors, one usually signals the order in which the comparison was done: With the null model marked as 0, and the alternative as 1, we would write $BF_{01} = 3$ or $BF_{10} = \frac{1}{3}$. Generally, strong evidence in favor of the null or alternative is considered to be a Bayes factor larger than 10 [this follows from a suggested scale in Jeffreys 1998 (1939)]. Thus, a Bayes factor analysis either gives us evidence for the alternative, evidence for the null, or an inconclusive result.

Here, it is extremely important to understand that it makes little sense to report a single Bayes factor for a particular analysis; a so-called sensitivity analysis should be done using a range of priors on the target parameter to compute the Bayes factor (Schad et al. 2022a). Such a sensitivity analysis is necessary because the Bayes factor can change depending on the prior specification (Lee & Wagenmakers 2014); accordingly, to interpret the Bayes factor, one needs to understand what the prior distribution implies about our belief regarding that parameter. An example of a sensitivity analysis will help here.

I will compute Bayes factors with three different priors on the β parameter. The names used for the priors below are adapted from Spiegelhalter et al. (1994) and Gelman et al. (2014).

1. The mildly informative Normal(0, 0.1) prior mentioned above; here the null hypothesis is that $\beta = 0$.
2. An agnostic or uninformative prior, Normal(0, 1), that allows a wide range of possible values ranging from −948 to 948 ms; here, too, the null hypothesis is that $\beta = 0$.
3. An enthusiastic prior (one for English and another for Chinese) that represents a prior belief that is consistent with the theoretical claims discussed by Grodner & Gibson (2005) and Gibson & Wu (2013). For English, the prior assumes a small but positive effect, Normal(0.02, 0.01) (this assumes a 95% credible interval from 0 to 16 ms), and for Chinese a small but negative effect, Normal(−0.02, 0.01) (this assumes a 95% credible interval from −16 to 0 ms). I will use two different null hypotheses:
   (a) The null is that $\beta = 0$.
   (b) To illustrate a null hypothesis that does not have a point value, I use Normal(0, 0.01); this null hypothesis asserts that the effect is near 0 ms (ranging from −8 to 8 ms) but not necessarily exactly 0 ms.

The priors for β are summarized below:

$$\beta \sim \begin{cases} \text{Normal}(0, 0.1) & \text{Mildly informative prior} \\ \text{Normal}(0, 1) & \text{Agnostic/uninformative prior} \\ \text{Normal}(0.02, 0.01) & \text{Informative (enthusiastic) prior (English)} \\ \text{Normal}(-0.02, 0.01) & \text{Informative (enthusiastic) prior (Chinese)} \end{cases}$$

The results of the Bayes factor analysis are shown in **Table 2**. What does this Bayes factor analysis show? First, notice that regardless of which set of priors we choose, the evidence for the relative clause effect is at most 4.5 in English and is not at all convincing for Chinese. So, the evidence for the relative clause effect is not particularly strong for either language. This is a very important takeaway from the analysis presented here.

Second, notice that the more informative prior Normal(0, 0.1) pushes the posterior closer to zero, and the informative prior in English, Normal(0.02, 0.01), pushes the posterior toward the mean for this prior distribution; a similar pattern is seen in the analysis of the Chinese data. This is a general characteristic of Bayesian analysis: The posterior is a compromise between the prior and the likelihood. The more informative the prior, the more influence it has in determining the posterior. Third, notice that the mere fact that zero is or is not included in the 95% credible

Table 2 The Bayes factor analysis under four different sets of priors, and posterior estimates of the relative clause effect in English and Chinese

	Null	Alternative	BF_{10}	Posterior mean and 95% CrI
English	$\beta = 0$	$\beta \sim$ Normal(0, 0.1)	4.55	35 [2, 69]
	$\beta = 0$	$\beta \sim$ Normal(0, 1)	0.95	45 [8, 85]
	$\beta = 0$	$\beta \sim$ Normal(0.02, 0.01)	2.44	8 [2, 15]
	$\beta \sim$ Normal(0, 0.01)	$\beta \sim$ Normal(0.02, 0.01)	2.19	8 [2, 15]
Chinese	$\beta = 0$	$\beta \sim$ Normal(0, 0.1)	0.95	−24 [−66, 17]
	$\beta = 0$	$\beta \sim$ Normal(0, 1)	0.13	−31 [−80, 16]
	$\beta = 0$	$\beta \sim$ Normal(−0.02, 0.01)	1.53	−9 [−18, −1]
	$\beta \sim$ Normal(0, 0.01)	$\beta \sim$ Normal(−0.02, 0.01)	1.5	−9 [−18, −1]

Notice that this analysis does not furnish convincing evidence for the relative clause effect in either language; this is in stark contrast to the frequentist analyses discussed above, which were based on whether the *p* value was below 0.05 or not. Abbreviations: BF, Bayes factor; CrI, credible interval.

interval does not tell us whether we have evidence for the relative clause effect; only the Bayes factor can tell us whether we have evidence for an effect (and we do not). Fourth, notice that whenever the prior is uninformative [here, Normal(0, 1)], the Bayes factor is unduly biased in favor of the null hypothesis; this is one important reason to avoid using only an uninformative prior in a Bayes factor analysis (cf. the misleading advice in, e.g., Wagenmakers et al. 2018, to compute Bayes factors using so-called default priors that are uninformative). Finally, one could imagine computing Bayes factors under other priors (e.g., adversarial priors that express a competing theoretical prediction other than the ones discussed here) if there is good reason to do so. The great advantage of the Bayes factor lies in its flexibility in allowing us to investigate the evidence for our hypothesis of interest (expressed as the prior on β) relative to some appropriate null hypothesis (we are no longer restricted to a point null like $\beta = 0$ as in frequentist NHST).

Thus, the overall conclusion from the Bayes factor analysis would be that neither the English nor the Chinese data furnish decisive evidence for a relative clause effect. By contrast, a frequentist analysis delivers misleading conclusions, as discussed above.

It is generally the case that the Bayes factor will furnish a more realistic picture than frequentist NHST of what we learned from the data, regardless of whether we use estimation to draw inferences or carry out explicit hypothesis testing.

5. LEARNING TO ACCEPT UNCERTAINTY

Analyzing data as suggested in this article means that we need to be willing to express uncertainty about the conclusions. Two practical problems that arise are the following.

1. Often, for logistical or financial reasons, it may be impossible to run a properly powered study; how can one proceed in this situation?
2. Journals generally tend to reject papers that do not make a decisive claim; would expressing uncertainty about the result lead to nonpublishable results?

Regarding the first point, it is true that most studies in linguistics will end up being underpowered. But, as I have tried to show in this article, such underpowered studies are useful and informative if seen as preliminary studies that future researchers can build on, either in a meta-analysis or for planning follow-up studies. Of course, when possible, one should try to run as high-powered a study as one can and to always attempt a replication, but if one has limited time and/or money, having some data is still better than having no data at all.

Regarding the standards that journals impose on papers, reviewers and editors will have to reflect on the fact that statistical analysis will usually only get us so far; those looking for certainty in statistics will be disappointed. The replication crisis should have made this clear to everyone (Open Sci. Collab. 2015). In psycholinguistics, we are seeing the consequences of these artificial constraints imposed by journals: Type M errors will be published preferentially; nonsignificant results will be presented as significant through misleading analyses using aggregated data or ignoring model assumptions, ignoring the MinF′ value, and declaring significance anyway; and repeated null results will be incorrectly argued for using severely underpowered designs. The alternative—which journal editors and reviewers will need to learn to accept—is to openly acknowledge the uncertainty inherent in data (Vasishth & Gelman 2021).

SUMMARY POINTS

1. Frequentist null hypothesis significance testing is meaningful only when power is high.
2. Simulating data before conducting an experiment is a very important component of the analytical principle; simulation tells us what we can in principle learn from our experiment design.
3. Knowledge will advance better in the field if the focus is on reporting estimates and their uncertainties without necessarily carrying out the usual, largely artificial hypothesis tests. This is likely to result in a much better quantitative understanding of the phenomenon being studied, and evidence synthesis (meta-analysis) can then be used to build on previous work.
4. To establish whether an effect exists, a formal model comparison with a baseline model is necessary. The Bayes factor is the most conservative, informative, and flexible way to carry out such a hypothesis test.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

OPEN DATA AND CODE STATEMENT

Reproducible code associated with this review is available at **https://vasishth.github.io/ARLVasishth/**.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 317633480, SFB 1287 (2021–2025).

LITERATURE CITED

Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59:390–412

Bates DM, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1). **https://doi.org/10.18637/jss.v067.i01**

Berman M, Jonides J, Lewis RL. 2009. In search of decay in verbal short-term memory. *J. Exp. Psychol.: Learn. Mem. Cogn.* 35:317–33

Bürki A, Alario FX, Vasishth S. 2022. When words collide: Bayesian meta-analyses of distractor and target properties in the picture-word interference paradigm. *Q. J. Exp. Psychol.* In press. **https://doi.org/10.1177/17470218221114644**

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. *J. Mem. Lang.* 114:104125

Bürkner PC. 2017. brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80(1). **https://doi.org/10.18637/jss.v080.i01**

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14:365–76

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, et al. 2017. Stan: a probabilistic programming language. *J. Stat. Softw.* 76(1). **https://doi.org/10.18637/jss.v076.i01**

Clark H. 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *J. Verbal Learn. Verbal Behav.* 12:335–59

Cohen J. 1962. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* 65:145–53

Cohen J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum. 2nd ed.

Cox CMM, Keren-Portnoy T, Roepstorff A, Fusaroli R. 2022. A Bayesian meta-analysis of infants' ability to perceive audio-visual congruence for speech. *Infancy* 27:67–96

Fedorenko E, Gibson E, Rohde D. 2006. The nature of working memory capacity in sentence comprehension: evidence against domain-specific working memory resources. *J. Mem. Lang.* 54:541–53

Gelman A, Carlin JB. 2014. Beyond power calculations: assessing Type S (sign) and Type M (magnitude) errors. *Perspect. Psychol. Sci.* 9:641–51

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press. 3rd ed.

Gelman A, Hill J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge Univ. Press

Gibson E, Desmet T, Grodner D, Watson D, Ko K. 2005. Reading relative clauses in English. *Cogn. Linguist.* 16:313–53

Gibson E, Wu HHI. 2013. Processing Chinese relative clauses in context. *Lang. Cogn. Proc.* 28:125–55

Gordon PC, Hendrick R, Johnson M. 2001. Memory interference during language processing. *J. Exp. Psychol.: Learn. Mem. Cogn.* 27(6):1411–23

Gordon PC, Hendrick R, Johnson M. 2004. Effects of noun phrase type on sentence complexity. *J. Mem. Lang.* 51:97–104

Grodner D, Gibson E. 2005. Consequences of the serial nature of linguistic input. *Cogn. Sci.* 29:261–90

Hedges LV. 1984. Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *J. Educ. Stat.* 9:61–85

Higgins J, Green S. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. New York: Wiley-Blackwell

Hoenig JM, Heisey DM. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.* 55:19–24

Hsiao FPF, Gibson E. 2003. Processing relative clauses in Chinese. *Cognition* 90:3–27

Hurlbert SH. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* 54:187–211

Ioannidis JP. 2008. Why most discovered true associations are inflated. *Epidemiology* 19:640–48

Jäger LA, Engelmann F, Vasishth S. 2017. Similarity-based interference in sentence comprehension: literature review and Bayesian meta-analysis. *J. Mem. Lang.* 94:316–39

Jäger LA, Mertzen D, Van Dyke JA, Vasishth S. 2020. Interference patterns in subject-verb agreement and reflexives revisited: a large-sample study. *J. Mem. Lang.* 111:104063

Jeffreys H. 1998 (1939). *The Theory of Probability*. Oxford, UK: Oxford Univ. Press

Just MA, Carpenter PA. 1992. A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99(1):122–49

Kass RE, Raftery AE. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–95

Kruschke JK. 2010. What to believe: Bayesian methods for data analysis. *Trends Cogn. Sci.* 14:293–300
Kruschke JK, Liddell TM. 2018. The Bayesian New Statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. Bull. Rev.* 25:178–206
Lane DM, Dunlap WP. 1978. Estimating effect size: bias resulting from the significance criterion in editorial decisions. *Br. J. Math. Stat. Psychol.* 31:107–12
Lee MD, Wagenmakers EJ. 2014. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge Univ. Press
Logacev P, Bozkurt Mİ. 2021. Statistical power in response signal paradigm experiments. *Proc. Annu. Meeting Cogn. Sci. Soc.* 43:2211–17
Mahowald K, James A, Futrell R, Gibson E. 2016. A meta-analysis of syntactic priming in language production. *J. Mem. Lang.* 91:5–27
Moerbeek M, Teerenstra S. 2015. *Power Analysis of Trials with Multilevel Data*. Boca Raton, FL: CRC Press
Nicenboim B, Roettger TB, Vasishth S. 2018a. Using meta-analysis for evidence synthesis: the case of incomplete neutralization in German. *J. Phonet.* 70:39–55
Nicenboim B, Schad D, Vasishth S. 2022. *An Introduction to Bayesian Data Analysis for Cognitive Science*. Forthcoming. **https://vasishth.github.io/bayescogsci/book/**
Nicenboim B, Vasishth S, Engelmann F, Suckow K. 2018b. Exploratory and confirmatory analyses in sentence processing: a case study of number interference in German. *Cogn. Sci.* 42(Suppl. 4):1075–100
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* 142:107427
O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, et al. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities*. London: Wiley
Open Sci. Collab. 2015. Estimating the reproducibility of psychological science. *Science* 349:aac4716
Pankratz E, Yadav H, Smith G, Vasishth S. 2021. Statistical properties of the speed-accuracy trade-off (SAT) paradigm in sentence processing. *Proc. Annu. Meeting Cogn. Sci. Soc.* 43:2176–82
Pinheiro JC, Bates DM. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag
Rouder JN, Haaf JM. 2021. Are there reliable qualitative individual differences in cognition? *J. Cogn.* 4(1):46
Schad DJ, Betancourt M, Vasishth S. 2020a. Toward a principled Bayesian workflow in cognitive science. *Psychol. Methods* 26:103–26
Schad DJ, Nicenboim B, Bürkner PC, Betancourt M, Vasishth S. 2022a. Workflow techniques for the robust use of Bayes factors. *Psychol. Methods*. **https://doi.org/10.1037/met0000472**
Schad DJ, Nicenboim B, Vasishth S. 2022b. Data aggregation can lead to biased inferences in Bayesian linear mixed models. arXiv:2203.02361 [stat.ME]
Schad DJ, Vasishth S, Hohenstein S, Kliegl R. 2020b. How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Mem. Lang.* 110:104038
Spiegelhalter DJ, Freedman LS, Parmar MK. 1994. Bayesian approaches to randomized trials. *J. R. Stat. Soc. A* 157:357–416
Stan Dev. Team. 2022. RStan: the R interface to Stan. *Statistical Software*. **https://cran.r-project.org/web/packages/rstan/vignettes/rstan.html**
Vasishth S, Chen Z, Li Q, Guo G. 2013. Processing Chinese relative clauses: evidence for the subject-relative advantage. *PLOS ONE* 8(10):e77006
Vasishth S, Chopin N, Ryder R, Nicenboim B. 2017. Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture process: a case study involving Chinese relative clauses. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pp. 1278–83. London: Comput. Found. Cogn.
Vasishth S, Engelmann F. 2022. *Sentence Comprehension as a Cognitive Process: A Computational Approach*. Cambridge, UK: Cambridge Univ. Press
Vasishth S, Gelman A. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* 59:1311–42
Vasishth S, Nicenboim B. 2016. Statistical methods for linguistic research: foundational ideas – Part I. *Lang. Linguist. Compass* 10:349–69

Vasishth S, Nicenboim B, Beckman ME, Li F, Kong EJ. 2018. Bayesian data analysis in the phonetic sciences: a tutorial introduction. *J. Phonet.* 71:141–61

Vasishth S, Schad D, Bürki A, Kliegl R. 2022a. *Linear Mixed Models in Linguistics and Psychology: A Comprehensive Introduction*. Forthcoming. **https://vasishth.github.io/Freq_CogSci/**

Vasishth S, Yadav H, Schad D, Nicenboim B. 2022b. Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics. *Comput. Brain Behav.* **https://doi.org/10.1007/s42113-021-00125-y**

Wagenmakers EJ, Love J, Marsman M, Jamil T, Ly A, et al. 2018. Bayesian inference for psychology. Part II: example applications with JASP. *Psychonom. Bull. Rev.* 25:58–76

Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* 70:129–33

Wilke C. 2019. *Fundamentals of Data Visualization*. Sebastopol, CA: O'Reilly