Some right ways to analyze (psycho)linguistic data

Shravan Vasishth[1]

[1]**Department of Linguistics, University of Potsdam, Potsdam 14476, Brandenburg, Germany; email: vasishth@uni-potsdam.de; orcid: 0000-0003-2027-1994**

Annual Reviews of Linguistics 2022. AA:1–20

https://doi.org/10.1146/((please add article doi))

Keywords

Statistics, simulation, experimental science, data analysis, significance testing, Bayesian statistics, Bayes factors, estimation, uncertainty quantification, transparency, open science

Abstract

Much has been written on the abuse and misuse of statistical methods, including p-values, statistical significance, etc. 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 vs. 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, the most important ones being using simulation to understand the design properties of one's experiment before running it, visualizing data before carrying out a formal analysis, and simulating data from the fitted model to understand the model's behavior.

Contents

1. INTRODUCTION

> *If you worked in areas inhabited by demons you would be in trouble regardless of the perfection of your experimental designs.*
> (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 to find that two statisticians analyze the same data very differently and even come to different conclusions/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 sub-field has its own particular sets of commonly used statistical models; in linguistics, the modern standard is the linear mixed model, also referred to as the hierarchical model (Pinheiro & Bates 2000). Accordingly, in this review, I will focus on this modeling framework. I will 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 linear mixed model (Bates et al. 2015). If the reader lacks this background, introductory articles like Baayen et al. (2008); Vasishth & Nicenboim (2016); Vasishth et al. (2018b) would be a good starting point. Other, more comprehensive textbook references are provided at the end of this article.

Experimental science is more than careful experiment designs and the use of sophisticated methods like ERPs and eye-tracking. There are six components in an experiment: (i) setting up a research hypothesis, (ii) designing the experiment, (iii) implementing it in software and running the experiment, (iv) pre-processing the data, (v) statistical analysis, and (vi) interpreting the results of the analysis. In linguistics, we have mastered the first four steps, but we have faltered when it comes to the last two, so these are the issues I will focus on.

2. AN EXAMPLE: RELATIVE CLAUSE PROCESSING

In order to make the discussion concrete, I will focus 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 an easy prospect for a clear answer; but I show below that there are important issues to consider before making any decisive claim.

We will consider published data from English and Chinese. English relative clauses are shown in 1a and 1b. 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., Just & Carpenter 1992; Grodner & Gibson 2005; Gibson et al. 2005; Gordon et al. 2001; Fedorenko et al. 2006).[1]

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

In contrast to English, Chinese relatives (see 2a and 2b below; the examples are from Gibson & Wu (2013)) 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 object relatives. Hsiao & Gibson (2003) and Gibson & Wu (2013) argue that this increased gap distance in subject vs. object relatives leads to longer reading times at the head noun in subject relatives. Compare this with English (1a,1b 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 vs. 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.

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

Is there evidence for these predicted patterns in English vs. 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 get a definitive answer to this question. This is in fact what researchers did do for English (Grodner & Gibson 2005) and Chinese (Hsiao & Gibson 2003; Gibson & Wu 2013). It would be fantastic if answering such questions were so easy. As I will show below, obtaining a decisive answer to our research question is rather more involved. If we want clear answers, we need to invest much more time and money than we normally do in psycholinguistics.

[1]There are some important design problems in such an experiment: the relative clause verb is not in the same position in the two conditions, and the pre-critical region is different. However, we ignore these confounds in the design here, noting that these can be mitigated by, for example, comparing the entire relative clause region (Fedorenko et al. 2006).

2.1. The published analyses for English and Chinese

I first summarize the published statistics in the original articles (Grodner & Gibson 2005; Gibson & Wu 2013), and then turn to how one can carry out informative studies that can actually answer the research question.

2.1.1. The original analyses the English data. A critical region of interest in Grodner & Gibson (2005) for which there are clear theoretical predictions is the embedded verb inside the relative clause. The reported statistical analyses show strong evidence against no difference between the two conditions. The estimates are in the predicted direction (object relatives are harder to process than subject relatives: 422 vs. 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, ... p < .001; t2(1, 15) = 14.3, ..., p < .01."

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

2.1.2. The original analyses of the Chinese data. For Chinese, the critical region was the head noun. Gibson & Wu (2013) write the following:

> "... the head noun for the RC ... was read more slowly in the SRC condition [F1(1, 36)=6.92, p<.01; F2(1, 14)=4.62, 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: MinF′(1,33)= 2.77, p=0.11.

In the remainder of the paper, I will revisit the relative clause question, illustrating some of the best practices for planning and conducting experimental studies.

3. PLANNING FUTURE STUDIES ON ENGLIH 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 is easy to obtain (Grodner & Gibson 2005; Gibson & Wu 2013, generously made their data publicly available), one can use estimates from these existing data for planning a future set of studies. If such data is not available, one can either derive estimates of parameters from previously published work through meta-analysis (Vasishth et al. 2013; Jäger et al. 2017; Jäger et al. 2020; Bürki et al. 2020; Nicenboim et al. 2018a, 2021), or carry out a preliminary study to plan for a future study (Nicenboim et al. 2018b). To conserve space, I don't show

[2]The authors carried out an automated data deletion procedure: "Reading times that differed from the mean of a condition and region by more than 3 SDs were omitted from analyses. This adjustment discarded 1.6% of the data." However, I could not reproduce their estimates after following this procedure. In any case, as I show below, it was not necessary to truncate data in this manner in the present case.

the R code used in this paper, 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). Later on, I will discuss alternatives to NHST, specifically Bayes factors and estimation.

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

Type M error: The expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size, given that the result is significant.

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

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 in Vasishth & Gelman (2021), this 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., Pankratz et al. 2021; Logacev & Bozkurt 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 signifciant effect and tries to replicate the study, the effects will generally not be replicable. This issue has been discussed repeatedly in the statistics literature (Lane & Dunlap 1978; Hedges 1984) 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.

3.2. Design and power analysis

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?

Post-hoc or "observed" power: Assuming that the observed effect size and variability (standard error) are equal to the true parameter values, the probability of rejecting the null hypothesis is called post-hoc power or "observed power." It is a one-to-one function of the observed p-value and therefore not informative (Hoenig & Heisey 2001).

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

1. Fit a linear mixed model 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.

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

Table 1 Parameter estimates (with standard errors for the fixed effects) from the linear mixed models fit to the English (Grodner and Gibson, 2005, Experiment 1) and Chinese (Gibson and Wu 2013) relative clause data. 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; and Cor to the correlation between random intercepts and random slopes. Blank cells imply that the parameter in question was not estimated because of convergence problems.

	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: subj (Intercept)	0.318	0.245
sd: subj cond	0.221	0.112
Cor: subj (Intercept) cond	0.58	
sd: item (Intercept)	0.036	0.181
sd: item cond	0.081	
sd: Residual	0.361	0.515
Num. obs.	672	547
Num. groups: subj	42	37
Num. groups: item	16	15

power. Researchers often mistakenly draw inferences about the power properties of their already-conducted study; i.e., they compute "observed power." As discussed in Hoenig & Heisey (2001), this is a pointless exercise: post-hoc power is simply a 1:1 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 will often report such meaningless statistics, usually in order to argue that their null results are meaningful. Some examples of papers that report "observed power" in order to argue that their null results are interpretable are 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 a linear mixed model is the fixed effect 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.

Turning now to the power analyses of the English and Chinese data-sets, as shown in Table 1, in the Grodner & Gibson (2005) data, the intercept and slope on the log ms 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 ms to 89 ms.

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 ranging from -72 ms to 10 ms. These are tentative estimates; one could do a proper meta-analysis and come up with better estimates (e.g., Vasishth et al. 2013).

Back-transforming from the log scale: Suppose that the model for reading time (rt) data is $rt \sim Lognormal(\alpha + \beta \times x, \sigma)$, where α is the intercept, β the slope, `x` the ± 0.5 sum-coded predictor, and σ the residual standard deviation. This is equivalent to $\log(rt) \sim Normal(\alpha + \beta \times x, \sigma)$. One can obtain the predicted median reading times for each condition on the millisecond scale by computing $\exp(\alpha + \beta \times x) - \exp(\alpha - \beta \times x)$ (Nicenboim et al. 2021; Vasishth et al. 2022b).

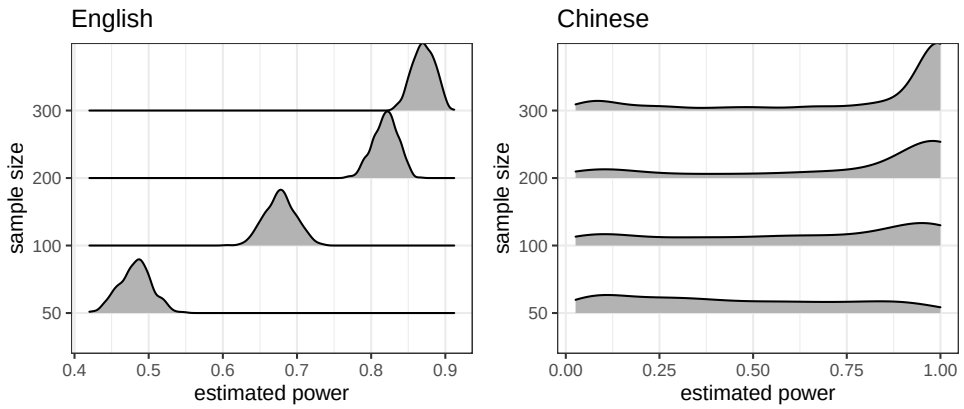

Figure 1

Estimated statistical power using simulation for the English and Chinese relative clause data. Each power distribution is generated by simulating data repeatedly from an assumed effect size of 0.12 (SE : 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, and 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, and 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 (for example, in Table 1, compare the residual standard deviations in Chinese vs. English: 0.52 vs. 0.36).

3.2.2. Results of the prospective power analyses. Figure 1 shows the distribution of power for sample sizes 50, 100, 200, and 300 participants and 16 items given the parameter estimates from the English Experiment 1 of Grodner & Gibson (2005), and the Chinese experiment from Gibson & Wu (2013). The bigger spread in power estimates for Chinese compared to English come from the fact that the Chinese data are much noisier (e.g., the estimated residual standard deviation in 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 42 and 37 (Grodner & Gibson 2005; Gibson & Wu 2013, respectively); these are severely underpowered studies.

Thus, if planning future studies on English and Chinese, and even if one optimistically assumes that the true effect sizes are as 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 effect 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, with similar results; see Vasishth et al. (2022b) for detailed discussion and example code.

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 frequentist null hypothesis testing paradigm. Even if one ends up running a small-sample study due to 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 (I illustrate later what such a tempered claim would look like).

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 boxplot or the like will reveal any extreme or potentially influential values. The mean can be extremely sensitive to extreme values, making a non-significant difference come out significant. An example is the Gibson & Wu (2013) data: if we remove just two extreme data points from the data-set consisting of 547 data points, the effect becomes non-significant. The Gibson & Wu (2013) paper reported this one effect as significant; just plotting the data before analyzing can prevent such erroneous reporting.[3]

Second, individual differences in the effect of interest should be visualized to get a sense of whether random slopes should included in the model. Often, such a visualization already makes it clear what the random effects structure of the linear mixed model 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 it

Figure 2 shows a boxplot 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, it is the subject relative condition. 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).

[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, this automated procedure is not applied consistently even by the same research group. For example, Grodner & Gibson (2005) deleted extreme values, but Gibson & Wu (2013) kept the extreme values that were the sole reason for the significant effect.

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 easier to process than subject relatives) that was reported in Gibson & Wu (2013) is determined by only two extreme data points in subject relatives, out of a total of 547 data points.

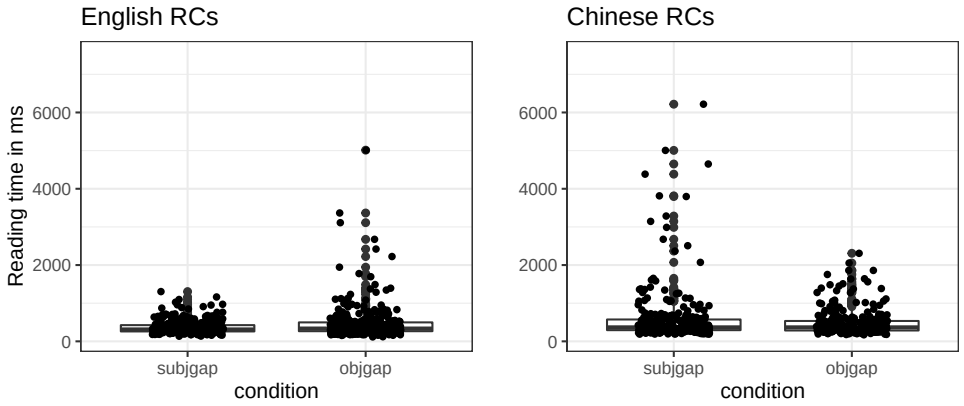

Figure 2

Boxplots showing the distribution of the Grodner and Gibson (2005) Experiment 1 data on English subject and object relative clauses (left) and the Gibson and Wu (2013) data on Chinese relative clauses. Shown are reading times (in ms) by condition at the critical region (the relative clause verb).

The by-participant individual differences in the relative clause effect in English and Chinese are shown in Figure 3. These individual-level plots are not the shrunken estimates from the linear mixed models (Bates et al. 2015), but rather use the data from each subject independently (from the so-called no-pooling model, Gelman & Hill 2007; Vasishth et al. 2022a; Nicenboim et al. 2021). What do we learn from this plot?

1. In the English data, participants 28, 36, and 37 show the largest relative clause effects (0.9, 0.77, 0.71 log ms) compared to the other participants and compared to the mean of 0.12 log ms. On the millisecond scale, these estimates amount to a relative clause effect of 413, 319, and 543 ms, respectively. These values are 10 times larger than the average effect estimated from the linear mixed model (48 ms).
2. In the Chinese data, participant 27 has only two observations for subject relatives (7 or 8 data points are expected)! There is no explanation in Gibson & Wu (2013) as to what causes this data loss. Moreover, participant 11 has an unusually large relative clause effect, which seems to be driven by one extreme value (5006 ms reading time). This extreme value suggests that the effect of a few data points could have a dramatic impact on the statistical inference; this turns out to be the case. As I show below, if we log-transform the data, thereby reducing the impact of these extreme values, the conclusions from the data change radically.

Apart from the above descriptive observations, these plots show considerable variation between participants, suggesting that by-particiant intercepts and slopes will probably be

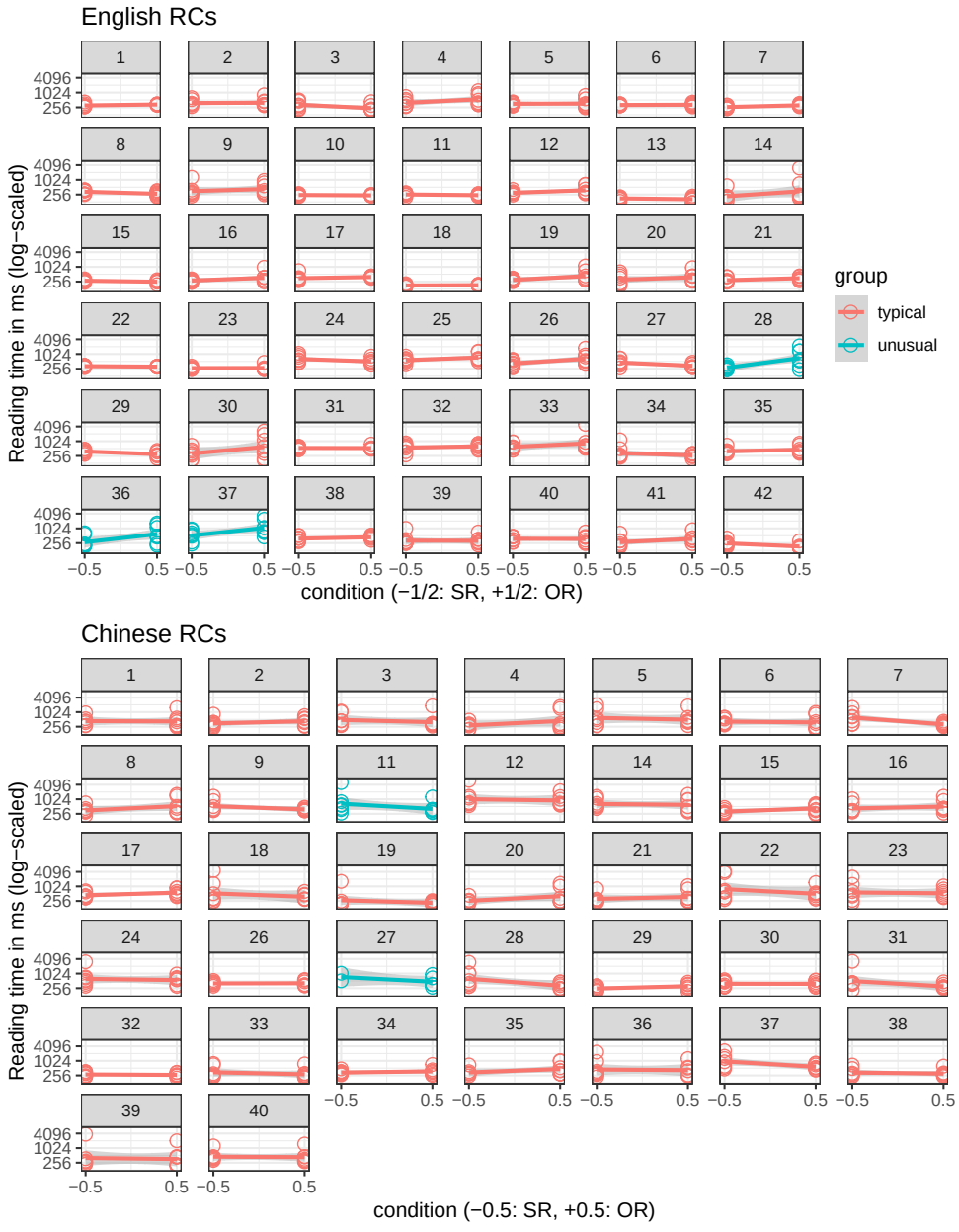

Figure 3

A so-called xy-plot showing the distribution of the Grodner and Gibson (2005) Experiment 1 data on English subject and object relative clauses. Shown are log reading times by condition and by subject at the critical region (the relative clause verb). The participants with unusual responses are marked.

needed in the models. One could draw similar by item plots (omitted here to conserve space).

These figures are only relevant for the reading time data discussed here; due to space limitations, visualizations for different types of data can't be shown here. Examples of good-quality data visualizations are discussed in Wilke (2019).

4.2. Statistical inference

Drawing inferences from the data requires that we specify a statistical model; deciding what 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 will 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 0's and 1's; the appropriate distributional assumption would be a Bernoulli. Statistical software generally assumes that the researcher knows what they are doing and return 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 linear mixed models (LMMs) fit 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 non-significant (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 saw in Table 1 what the estimates from a frequentist linear mixed model 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 RC effect but not for Chinese. As we saw earlier, 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 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".

4.2.1. 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 (Stan Development Team 2016; Carpenter et al. 2017). Standard linear mixed models that linguists are used to fitting with the package `lme4` can now easily be fit 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.

4.2.1.1. 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 *didn't* collect, and the statistical test (the t-test, the Chi-squared test, the 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 null hypothesis significance testing doesn't 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 Bayes, 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's 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 Bayes we talk about the relative clause effect as a probability distribution. As a hypothetical example, we might believe (based on prior data or theory or computational modeling; see O'Hagan et al. 2006; Nicenboim et al. 2021, for how such prior information can be derived) that, in self-paced reading data, the RC 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 RC 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 can then be included in the Bayesian linear mixed model to compute something called the posterior distribution of the RC effect, which gives the updated probability distribution of the RC effect after seeing the data. In other words, a critical advantage of the Bayesian paradigm we have the opportunity to formally build on prior knowledge.

Users of frequentist methods are not accustomed to thinking about and utilizing 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 (Vasishth et al. 2013; Mahowald et al. 2016; Jäger et al. 2017; Nicenboim et al. 2018a, 2020; Bürki et al. 2020; Cox et al. 2022). These kinds of meta-analyses can be very helpful in deriving prior distributions for future studies (Nicenboim et al. 2021; 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

[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. See Schad et al. (2020a, 2022a); Vasishth et al. (2022b) for detailed discussion.

will help.

4.2.1.2. A Bayesian analysis of the relative clause data. Suppose that we re-run the linear mixed models discussed above; this time, we use the Bayesian framework. Here is the formal statement of the linear mixed model for both English and Chinese:

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

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:

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

There are nine parameters in the model $(\alpha, \beta, \sigma_{u0}, \sigma_{u1}, \rho_u, \sigma_{w0}, \sigma_{w1}, \rho_w, \sigma)$, and each parameter gets a prior distribution defined for it. Below, I define so-called regularizing priors for the parameters.

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

The prior on the intercept α implies that the mean reading time can range from 122 to 1339 ms with probability 0.95. The prior on the β parameter implies that the RC effect can range from -81 to 81 ms with probability 0.95. This is a mildly informative prior; what this prior expresses is agnosticism about the sign of the RC effect, but also assumes that the effect is not likely to be very large. For an empirically based justification for such a mildly informative prior, see chapter 6 of Nicenboim et al. (2021).

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

4.2.1.3. 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

Regularizing priors: Regularization here means that implausible values of parameters are a priori ruled out. For example, in the frequentist model fit with `lme4`, one often sees a $+1$ or -1 correlation between varying intercepts and varying slopes, and/or estimates of 0 for variance components; these estimates signal a convergence failure in the `lmer` function because it implies a degenerate variance-covariance matrix (Pinheiro & Bates 2000). This kind of convergence problem cannot occur in a Bayesian model when a regularizing prior (the LKJ prior shown below) is used on the correlation parameters (of course, other convergence problems can occur but these can generally be resolved easily through methods like reparameterization or tuning the sampling algorithm; see Nicenboim et al. 2021, for discussion).

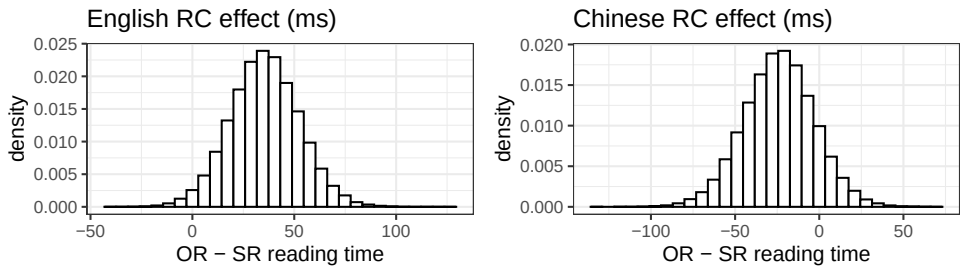

Figure 4

The posterior distributions of the relative clause effect (object relative minus subject relative) in milliseconds at the critical region (the relative clause verb in English, 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.

consistent with the qualitative claim that object relatives will be harder in English and easier in Chinese compared to the respective baseline condition; however, the 95% credible intervals shows that there is quite a lot of variability possible in the estimates. This high variability, or low precision of the estimate, is highly informative because it is an indication that we have relatively sparse data (a detail that we have already independently established with the power analysis earlier). Due to this low precision, no strong conclusions can be drawn about these effects from these data.

Now, if we want to go further and find out whether there is evidence for an RC 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).

4.2.1.4. 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 RC types, is 0 log ms, and the alternative could be that β is $Normal(\mu = 0, \sigma = 0.1)$ on the log ms 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 ms scale (this corresponds to the 95% credible interval [-8,8] on the ms scale) with an alternative model 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 1939/1998). 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 about 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 ms 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 in 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 ms to 16 ms), and for Chinese a small but negative effect, Normal(-0.02,0.01) (this assumes a 95% credible interval from -16 ms to 0 ms). I will use two alternative null hypotheses:
   (a) The null is that $\beta = 0$.
   (b) As an illustration of a null hypothesis that doesn't have a point value, I use Normal(0,0.01); this null hypothesis asserts that the null hypothesis is that the effect is near zero ms (ranging from -8 ms to 8 ms), but not necessarily exactly 0 ms.

The priors for β are summarized below:

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

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 RC effect is at most 4.5 in English and not at all convincing for Chinese. So, the evidence for the RC effect is not particularly strong for either language. 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 towards the mean for this prior distribution; a similar pattern is seen in the analysis of the Chinese data.

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

	Null	Alternative	BF_{01}	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]

Sensitivity analysis: In Bayesian data analysis, and Bayes factor analyses in particular, fitting a model with a range of alternative priors (particularly on the target parameter, here β) is very helpful in understand the impact of the prior on the posterior. This kind of analysis also makes it possible to investigate the data from different a priori theoretical positions.

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 interval does not tell us whether we have evidence for the RC effect; only the Bayes factor can tell us whether we have evidence for an effect (and we don't). 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 why one should never use only an uninformative prior in a Bayes factor analysis (cf. the advice in articles like 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 (for example, adversarial priors that express a competing theoretical prediction than the ones discussed here) if there is good reason to do this. 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 not the Chinese data furnish decisive evidence for a relative clause effect. Contrast this with the published conclusions in Grodner & Gibson (2005) and Gibson & Wu (2013): in both papers, overly strong claims are made for the relative clause effect in the two languages. Given the low power that 40 participants yield for these designs, the Bayes factor analysis is much more nuanced and realistic than the simplistic analysis based on p-values.

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. THE ELEPHANT IN THE ROOM: HOW TO EXPRESS UNCERTAINTY AND STILL GET PUBLISHED?

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: (i) Often, due to logistical or financial reasons, it may be impossible to run a properly powered study; how can one proceed in this situation? (ii) Journals generally tend to reject papers

that do not make a decisive claim; wouldn't expressing uncertainty about the result lead to non-publishable results?

Regarding the first point, it is true that most studies will be similar to the Grodner & Gibson (2005) and Gibson & Wu (2013) studies in being underpowered. But, as I tried to show in this article, such underpowered studies are useful and informative 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, but if one has limited time and/or money, some data is still better than 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 Science Collaboration 2015). In psycholinguistics, we are seeing the consequences of these artificial constraints imposed by journals: Type M errors will be published preferentially (e.g., Levy & Keller 2013), non-significant results will be presented as significant through misleading analyses using aggregated data or ignoring model assumptions (e.g., Hsiao & Gibson 2003; Gibson & Wu 2013), ignoring the MinF′ value and declaring significance anyway (e.g., Van Dyke & McElree 2006; Fedorenko et al. 2006), or repeated null results will be incorrectly argued for using severely underpowered designs (e.g., Vasishth & Lewis 2006). The alternative is to openly accept the uncertainty inherent in data (Vasishth & Gelman 2021). My own experience has has been that it is usually possible to publish underpowered studies in mainstream journals without overstating the claims (e.g., Nicenboim et al. 2018b; Vasishth et al. 2018a; Jäger et al. 2020; Nicenboim et al. 2020; Avetisyan et al. 2020; Lissón et al. 2021).

SUMMARY POINTS

1. Frequentist null hypothesis significance testing is only meaningful 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.

6. OPEN DATA AND CODE STATEMENT

Reproducible code associated with this paper is available from https://vasishth.github.io/ANRVasishth/.

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.

ACKNOWLEDGMENTS

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

LITERATURE CITED

Avetisyan S, Lago S, Vasishth S. 2020. Does case marking affect agreement attraction in comprehension? *Journal of Memory and Language* 112

Baayen RH, Davidson DJ, Bates DM. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:390–412

Bates DM, Maechler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1–48

Berman M, Jonides J, Lewis RL. 2009. In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35:317

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

Bürkner PC. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80:1–28

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, et al. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14:365–376

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, et al. 2017. Stan: A probabilistic programming language. *Journal of statistical software* 76

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

Cohen J. 1962. The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology* 65:145

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. *Journal of memory and language* 54:541–553

Gelman A, Carlin JB. 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9:641–651

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. Bayesian data analysis. Boca Raton, FL: Chapman and Hall/CRC Press, third edition ed.

Gelman A, Hill J. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press

Gibson E, Desmet T, Grodner D, Watson D, Ko K. 2005. Reading Relative Clauses in English. *Cognitive Linguistics* 16:313–353

Gibson E, Wu HHI. 2013. Processing Chinese relative clauses in context. *Language and Cognitive Processes* 28:125–155

Gordon PC, Hendrick R, Johnson M. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition* 27(6):1411–1423

Gordon PC, Hendrick R, Johnson M. 2004. Effects of noun phrase type on sentence complexity. *Journal of Memory and Language* 51:97–104

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

Hedges LV. 1984. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* 9:61–85

Higgins J, Green S. 2008. Cochrane handbook for systematics 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. *The American Statistician* 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. *Ecological monographs* 54:187–211

Ioannidis JP. 2008. Why most discovered true associations are inflated. *Epidemiology* 19:640–648

Jäger LA, 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 LA, Mertzen D, Van Dyke JA, 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 MA, Carpenter PA. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 99(1):122–149

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the american statistical association* 90:773–795

Kruschke J, Liddell TM. 2018. The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review* 25:178–206

Kruschke JK. 2010. What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences* 14:293–300

Lane DM, Dunlap WP. 1978. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology* 31:107–112

Lee MD, Wagenmakers EJ. 2014. Bayesian cognitive modeling: A practical course. Cambridge University Press

Levy RP, Keller F. 2013. Expectation and locality effects in German verb-final structures. *Journal of Memory and Language* 68:199–222

Lissón P, Pregla D, Nicenboim B, Paape D, van het Nederend M, et al. 2021. A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science* 45

Logacev P, Bozkurt Mİ. 2021. Statistical power in response signal paradigm experiments, In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43

Mahowald K, James A, Futrell R, Gibson E. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language* 91:5–27

Moerbeek M, Teerenstra S. 2015. Power analysis of trials with multilevel data. CRC Press

Nicenboim B, Roettger TB, Vasishth S. 2018a. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70:39–55

Nicenboim B, Schad DJ, 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

Nicenboim B, Vasishth S, Engelmann F, Suckow K. 2018b. 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

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, et al. 2006. Uncertain judgements: Eliciting experts' probabilities. John Wiley & Sons

Open Science Collaboration. 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, In *Proceedings of CogSci 2021*, pp. 2176–2182

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 difference in cognition? *Journal of Cognition* 4

Schad DJ, Betancourt M, Vasishth S. 2020a. Toward a principled Bayesian workflow in cognitive science. *Psychological Methods* 26:103–126

Schad DJ, Nicenboim B, Bürkner PC, Betancourt M, Vasishth S. 2022a. Workflow techniques for the robust use of Bayes factors. *Psychological Methods*

Schad DJ, Nicenboim B, Vasishth S. 2022b. Data aggregation can lead to biased inferences in Bayesian linear mixed models

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

Spiegelhalter DJ, Freedman LS, Parmar MK. 1994. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 157:357–416

Stan Development Team. 2016. RStan: the R interface to Stan. R package version 2.14.1

Van Dyke J, McElree B. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language* 55:157–166

Vasishth S, Chen Z, Li Q, Guo G. 2013. Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE* 8:1–14

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 Cognitive Science Conference*. London, UK

Vasishth S, Engelmann F. 2022. Sentence comprehension as a cognitive process: A computational approach. Cambridge, UK: Cambridge University Press

Vasishth S, Gelman A. 2021. How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* 59:1311–1342

Vasishth S, Lewis RL. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language* 82:767–794

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

Vasishth S, Nicenboim B. 2016. Statistical methods for linguistic research: Foundational ideas – Part I. *Language and Linguistics Compass* 10:349–369

Vasishth S, Nicenboim B, Beckman ME, Li F, Kong EJ. 2018b. Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics* 71:141–161

Vasishth S, Schad DJ, Bürki A, Kliegl R. 2022a. 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

Vasishth S, Yadav H, Schad D, Nicenboim B. 2022b. Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics

Wagenmakers EJ, Love J, Marsman M, Jamil T, Ly A, et al. 2018. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review* 25:58–76

Wasserstein RL, Lazar NA. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician* 70:129–133

Wilke C. 2019. Fundamentals of data visualization. O'Reilly