Shravan Vasishth* and Andrew Gelman

How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis

https://doi.org/10.1515/ling-2019-0051
Received January 11, 2020; accepted June 13, 2021; published online September 7, 2021

Abstract: The use of statistical inference in linguistics and related areas like psychology typically involves a binary decision: either reject or accept some null hypothesis using statistical significance testing. When statistical power is low, this frequentist data-analytic approach breaks down: null results are uninformative, and effect size estimates associated with significant results are overestimated. Using an example from psycholinguistics, several alternative approaches are demonstrated for reporting inconsistencies between the data and a theoretical prediction. The key here is to focus on committing to a falsifiable prediction, on quantifying uncertainty statistically, and learning to accept the fact that – in almost all practical data analysis situations – we can only draw uncertain conclusions from data, regardless of whether we manage to obtain statistical significance or not. A focus on uncertainty quantification is likely to lead to fewer excessively bold claims that, on closer investigation, may turn out to be not supported by the data.

Keywords: experimental linguistics; statistical data analysis; statistical inference; uncertainty quantification

1 Introduction

Statistical tools are widely employed in linguistics and in related areas like psychology to quantify empirical evidence from planned experiments and corpus analyses. Usually, the goal is to objectively assess the evidence for one or another scientific position. Typically, conclusions from data are framed in decisive language. Examples are statements like: "we found a significant/robust effect of

***Corresponding author: Shravan Vasishth**, Department of Linguistics, University of Potsdam, Potsdam, Germany, E-mail: vasishth@uni-potsdam.de. https://orcid.org/0000-0003-2027-1994
Andrew Gelman, Department of Statistics, Columbia University, New York, NY, USA, E-mail: gelman@stat.columbia.edu. https://orcid.org/0000-0002-6975-2601

(some factor) X on (dependent variable) Y." If researchers fail to find a significant effect, too often they will incorrectly conclude that they have evidence for no effect: phrases like "X had no effect on Y" are often used in published papers: the conclusion is often framed as evidence of absence, rather than absence of evidence. Claims based on data tend to be stated deterministically because established practice in psychology, psycholinguistics, and linguistics generally tells us to place our results into one of two bins: "significant" or "not significant"; Greenland (2017) calls it dichotomania. When a result turns out to be statistically significant, we are taught to believe that we have found the truth. Even a single, small-sample experiment can be treated as big news, worthy of publication in a major journal. This way of thinking is fundamentally incorrect, a distortion of the underlying statistical theory.

A major reason for these misunderstandings stems from the perfunctory education provided in statistics, in both linguistics and psychology programs worldwide. Learning statistical theory and practice are inseparable from scientific reasoning; and contrary to what is an increasingly popular belief in linguistics, experimentally grounded research is no guarantee that research will become more grounded in objective facts, as opposed to the subjective beliefs that are traditionally used in intuition-based linguistic theorizing. What's missing in statistics education in these fields is basic training in what kinds of answers statistics can and cannot provide.

We begin by revisiting the underlying principles and assumptions of null hypothesis significance testing. This review, although very basic in nature, is necessary because in our experience many researchers are not clear on the details of the one-sample t-test. Then, we suggest some alternative ways in which conclusions can be drawn from data. In this paper, we assume that the reader has encountered some of the foundational ideas behind null hypothesis significance testing: the *t*- and *p*-value, Type I, II errors, and statistical power. A recent introductory book that is specifically aimed at linguists is available (Winter 2019); also see Navarro (2013) and the special issue on *Emerging data analysis in phonetic sciences* (Roettger et al. 2019).

We stress that there is nothing fundamentally new in this paper. Many researchers, especially in psychology, have covered the topics we discuss in published work, and much more extensively than we do here; the reader will benefit from reading this literature. Some examples are Cumming (2014), Kruschke (2013, 2014), Kruschke and Liddell (2018), Simmons et al. (2011), Verhagen and Wagenmakers (2014), Wagenmakers et al. (2018), and Yarkoni (2020). The contribution of the present paper is only to demonstrate, through some practical examples, how uncertainty can be communicated in linguistic research, and to explain why

statistical significance testing is not informative unless certain very specific conditions are met.

1.1 The logic of significance testing

The standard logic of significance-based testing can be illustrated by considering a simple example. Suppose we are interested in the difference in reading times between two conditions a and b. To make the discussion concrete, we will consider here a phenomenon called agreement attraction (Wagers et al. 2009). The claim in the psycholinguistics literature is that in sentences like (1), which are both ungrammatical, comprehenders read the auxiliary verb *were* faster in (1a) than in (1b).

(1) a. *The bodybuilder$_{+subject}^{-plural}$ who met the trainers$_{-subject}^{+plural}$ were $\{_{subject}^{plural}\}$ …

b. *The bodybuilder$_{+subject}^{-plural}$ who met the trainer$_{-subject}^{-plural}$ were $\{_{subject}^{plural}\}$ …

Several theoretical explanations have been proposed to account for this observed speedup. One of them (Engelmann et al. 2020; Vasishth et al. 2019; Wagers et al. 2009) is the claim that when the human sentence comprehension system encounters the plural marked auxiliary verb *were*, an attempt is made to access a plural-marked subject from memory in order to determine who the main actor of the sentence is. The search in memory for a plural-marked subject is initiated using a set of so-called retrieval cues (shown in brackets at the auxiliary verb in 1); the nouns are assumed to have a feature-specification marking, among other things, its subject status and number. The correct target for retrieval is the subject noun *bodybuilder* but it does not have the right plural feature specification (this is what makes both the sentences ungrammatical). However, there is a non-subject (*trainers*) in (1a) that is plural-marked, and this noun occasionally is mistaken for the grammatical subject of the sentence.

Thus, based on the quantitative predictions (shown later, in Figure 5) of the model reported in Engelmann et al. (2020), the research hypothesis is that the auxiliary verb in (1a) will be read faster than in (1b). The statistical test of this hypothesis can be carried out in the frequentist paradigm by assuming that the reading times at the auxiliary verb in (1a) and (1b) have some unknown but fixed true mean reading times μ_a and μ_b respectively. A null hypothesis is set up which states that the difference between these two means is 0, i.e., that the two means are identical. Conventionally, we write this null hypothesis as $H_0 : \delta = \mu_a - \mu_b = 0$.

Having set up the null hypothesis, we collect data from n participants for both (1a) and (1b); usually, a Latin square design is employed (Arunachalam 2013). How the sample size n is decided on will be discussed in Section 3. For now, we assume

that we somehow decide to sample data from n participants, and each participant delivers one data point for condition (a) and one for condition (b). If each participant delivers more than one data point for each condition, an average of those multiple points is taken, so that what goes into the statistical test is one data point per participant per condition. In practice, we usually collect multiple data points from each participant for each condition and do not need to take the average as described here; but we can disregard this detail for now. For further discussion of how to analyze such repeated measures data without having to average the data, see Bates et al. (2015) and Winter (2019).

Given these data, we first compute a vector that contains each participant's difference score d, and then compute the mean difference between the two conditions, $\overline{d}$.

The standard procedure is to compute the observed mean difference in reading time:

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

We also compute the sample standard deviation s of the differences scores d_i:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \overline{d})^2}{n-1}}$$

Then, we compute the estimate of the standard error from the estimated standard deviation s and the sample size n:

$$\widehat{SE} = \frac{s}{\sqrt{n}}$$

The standard error gives us an estimate of the standard deviation of the sampling distribution of the difference of sample means under (hypothetical) repeated sampling: if (counterfactually) we were to run the experiment repeatedly with new participants from the same population, for large enough sample sizes, the distribution of sample means we would obtain would have a Normal distribution with estimated standard deviation of $\widehat{SE} = s/\sqrt{n}$; see Draper and Smith (1998) for further details.

In null hypothesis significance testing (NHST), we are interested in quantifying how much some statistic computed from our data deviates from outcomes expected under the null hypothesis. That is, in our case, assuming there is no difference between these conditions, we want to quantify the extent to which the difference we found is at odds with the null-hypothesized value of 0. To this end,

we compute a statistic called the *t*-statistic, which tells us how many standard error units the sample mean is away from the hypothesized mean $\delta = 0$.

$$t \cdot \widehat{SE} = \overline{d} - \delta$$

As shown in Figure 1, if the absolute value of the *t*-statistic is "large enough", i.e., if the sample mean of the differences is far enough away in standard error units in either direction from the hypothesized difference of means, the convention is to reject the null hypothesis. Glossing over some details and simplifying slightly, "large enough" is considered to be an absolute value (in standard error units) equal to or larger than 2. This is a simplification because what constitutes a large enough *t*-value depends on the sample size; but this simplification is good enough if we have 20 or more participants, which is usually the case at least in psycholinguistics. Usually, a so-called *p*-value is computed alongside the *t*-value; the *p*-value gives us the probability of observing the absolute *t*-value we obtained, or some value more extreme, assuming that the null hypothesis is true. The *p*-value cannot represent the probability that the null hypothesis is true – we have already assumed that it is true when we compute the *p*-value.

Once we reject the null hypothesis, the convention is to treat this rejection as evidence *for the specific research hypothesis we had*. In our case, the research hypothesis is that δ has a negative sign, so if we can reject the null hypothesis, we conclude that we have evidence for this claim. This is technically not correct, because all we have evidence against is the null hypothesis. In other words, the NHST approach doesn't tell us how well our research hypothesis fits with the data; it only tells us how improbable the test statistic (the *t*-value or the like) is assuming that the null hypothesis is true.

$$t \cdot \widehat{SE} = \bar{d} - \delta$$

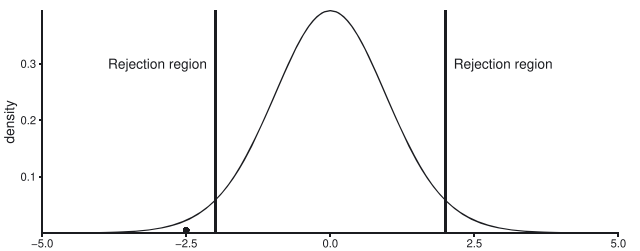

Figure 1: An illustration of the two-sided t-test. If the observed *t*-value (the black dot) falls in either of the rejection regions in the tails, then the null hypothesis is rejected.

The so-called confidence intervals are usually reported alongside the statistical test. For example, it is common to report a 95% confidence interval around the sample mean: $\overline{d} \pm 2 \times SE$. The confidence interval has a rather convoluted meaning that is prone to misinterpretation (Hoekstra et al. 2014). If a *p*-value is not provided, the confidence interval is often used as a proxy for the null hypothesis test: if 0 is not in the interval, then the null hypothesis is rejected. Used in this way, the confidence interval just becomes another equivalent way to conduct null hypothesis tests, raising the same problems that arise with the *t*-value based decision criterion. As we show in this paper, the confidence interval can be used to quantify uncertainty about the effect of interest, without making binary decisions like "accepting" or "rejecting" the null hypothesis. For a related discussion, see Cumming (2014) and Gelman and Greenland (2019).

1.2 Some problems with significance testing

In null hypothesis significance testing, we erroneously go from (i) data, (ii) some assumed statistical model and the assumptions associated with the model, and (iii) a theoretical prediction, to a decisive claim about the phenomenon we are interested in studying (in the above example, for the agreement attraction effect). There are at least two important problems with this approach to data analysis:

– *Low-power studies, when filtered by statistical significance, will lead to mis-estimation.* If the probability of obtaining a difference in means that represents the true effect (statistical power) is low, then one of two things can happen. If we run the study multiple times (i.e., under repeated sampling), either we will obtain null results repeatedly, or we will occasionally get significant or even highly significant effects that are overestimates of the quantity of interest (in our case, the difference in means). The null results will be inconclusive, even if we obtain them repeatedly. What is worse, any significant effects we find, no matter how low the *p*-value, will be overestimates or Type M(agnitude) errors (Gelman and Carlin 2014); they could even have the wrong sign (Type S error). We show below that, at least in one subset of phenomena studied in psycholinguistics, statistical power is often surprisingly low. Thus, low power has two consequences: when researchers repeatedly obtain a non-significant effect, they will often incorrectly conclude that there is evidence for no effect. For an example of such an invalid conclusion, see Phillips et al. (2011). When a significant effect is obtained, this outcome will be based on a mis-estimation of the true value of the parameter of interest. Mis-estimation might not seem like such a bad thing if the estimated effect is in the "right" direction; but it has the

bad consequence that future research will end up overestimating power, perpetuating invalid inferences.

– *Significant effects will often be non-replicable.* When power is low, any significant effect that is found in a particular experiment will tend not to replicate. In other words, in larger-sample direct replication attempts, the effect size will tend to be smaller and a statistically significant effect will tend to be found to be non-significant. Recent papers from psycholinguistics discuss this point in detail (Jäger et al. 2020; Nicenboim et al. 2020; Vasishth et al. 2018). Here, studies that originally showed a significant or near-significant effect were not replicable: the effect sizes in the replication attempts were smaller, and the original significant (or near-significant) effect did not come out significant. This inability to replicate an effect can be due to low power of the original experimental design, but even if power is high, especially in experiments involving human participants, effects can vary from study to study.

Psychologists (Cohen 1962, 1988) have long pointed out the importance of ensuring high statistical power for making discovery claims, but until recently these recommendations have largely been ignored in linguistics, psychology, and psycholinguistics; some recent papers that take power into account are Brehm and Goldrick (2017), Stack et al. (2018), and Zormpa et al. (2019). In response to the replication crisis that (partly) resulted from underpowered studies (Open Science Collaboration 2015), several remedies have been suggested, such as reducing the probability of committing a Type I error to 0.005 (Benjamin et al. 2018), or abolishing statistical significance testing entirely (McShane et al. 2019). But in any experimentally oriented research program, there is no substitute for an adequately powered study, and direct replications, if one's aim is to establish whether one's results are robust. As discussed later in this paper, when high-powered studies are simply impossible to carry out, other approaches, such as evidence synthesis, are needed.

Figure 2 shows power estimates based on the meta-analysis in Jäger et al. (2017) for reading studies on agreement attraction and closely related topics; for typical effect sizes (10–50 ms), and commonly seen standard deviations (150–300 ms) in reading studies (self-paced reading and total reading time in eye-tracking), and routinely used participant sample sizes (30–60), estimates of power are generally well below 80%. These estimates of power should give us pause.

When planning an experiment or research program, it is important to develop a good understanding of what the prospective power is; i.e., what the probability is of detecting, in a future study, an effect with a particular magnitude. If power is low, frequentist null hypothesis significance testing in an individual study will *never* yield meaningful results because, as discussed above, every possible

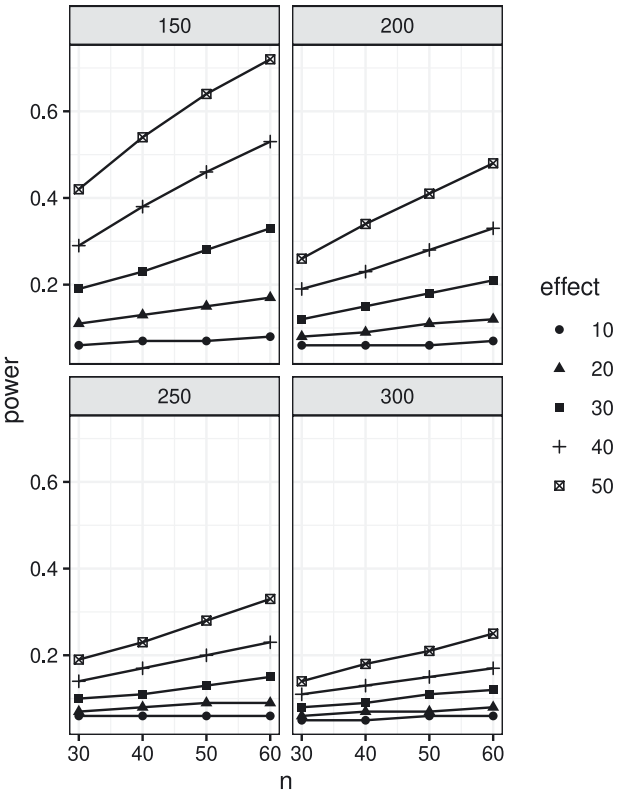

Figure 2: Power estimates for different numbers of participants (30, 40, 50, 60), assuming a standard deviation (of the residual) of 150, 200, 250, 300 (a typical range in reading studies), and an effect size ranging from 10 to 50 ms. For a justification of these estimates for the sample size, standard deviation, and effect sizes, see Jäger et al. (2017).

outcome under repeated sampling will be misleading: there will be a high proportion of inconclusive null results, and any significant effects will be due to misestimations of the true effect (Gelman and Carlin 2014). The frequentist method would of course furnish accurate inferences *in the long run* if there were no publication bias (if studies' results were published regardless of statistical significance), and meta-analyses were being carried out to synthesize evidence, as is routinely done in medicine (Higgins and Green 2008). One important pre-requisite for carrying out such meta-analyses is transparent data and code release along with the published paper, as recommended by Simmons et al. (2011), among others. Fortunately, modern open access journals in psycholinguistics, such as *Glossa: Psycholinguistics*, now have an Open Data Policy, which requires data and

code release. This policy decision is likely to have a positive impact on psycholinguistics, because it will allow for better-quality systematic reviews and evidence synthesis.

As long as one does not filter results by their statistical significance, the NHST paradigm works as you would expect: If power is low, most results will be regarded as uninformative, and the few significant results will be overestimates. But once you filter results by their statistical significance and power is low, all remaining results will be overestimates and the literature will be entirely misleading. Figure 3 illustrates this. Here, we assume that the true effect in a reading time experiment is 20 ms, and that standard deviation is 150. A paired t-test with 25 subjects will have

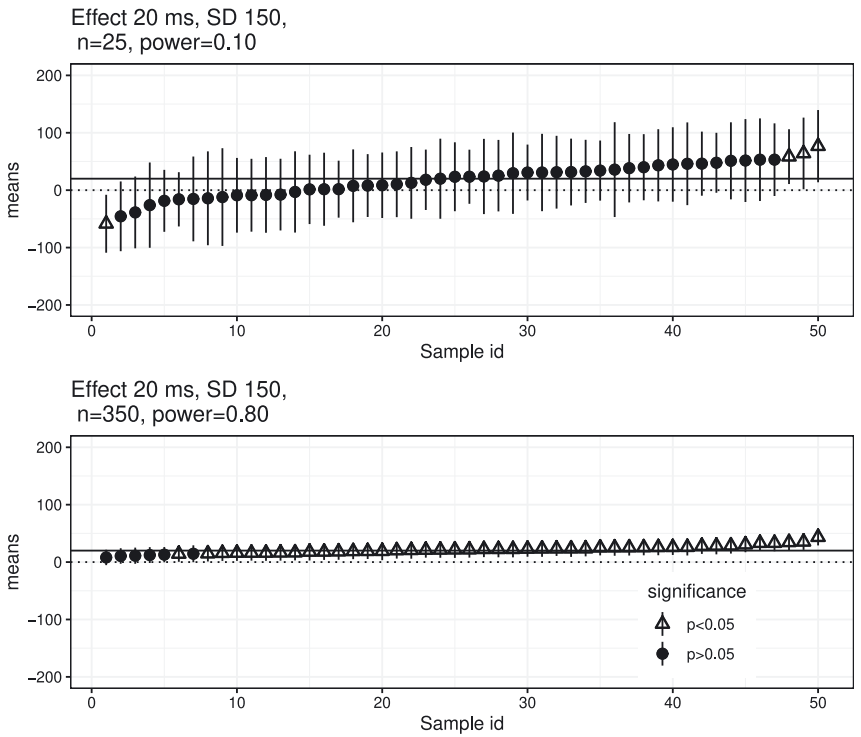

Figure 3: A simulation showing the effect of low versus high power on estimates of an effect, under repeated sampling (50 samples). Here, we assume that the data are generated from a normal distribution with mean 20 ms and standard deviation 150. The true mean is shown in each plot as a solid horizontal line. When power is low, every outcome is bad in a different way: either we get a lot of null results, or we get a significant outcome that results from a mis-estimate (a Type M or Type S error). By contrast, when power is high, significant results are meaningful: they are close to the true value.

approximate power 10%, and with 443 subjects, power will be approximately 80%. Statistical power is a continuum ranging from whatever the probability of committing a Type I error is (usually 5%) to 100%. By taking 10 and 80% power as representative low and high-power situations here, our aim is to show two edge cases.

Figure 3 shows that under low power, all outcomes are bad: there will be many uninformative null results, and any significant results will be based on misestimation of the true effect. Under high power, we get a high proportion of significant results, and, importantly, in each the estimated effect is close to the true value.

One important point to take away from this discussion is that the frequentist method can work well, but only under specific conditions; at the very least, power must be high. When power is low, relying on statistical significance or non-significance is not meaningful. When power is high, it can be useful to use statistical significance as one source of information (Wasserstein and Lazar 2016). But there are other sources of information that should not be ignored. We discuss this point next.

2 Accepting and quantifying uncertainty

So far, we have discussed some problems in the ways that the results of statistical tests are commonly misinterpreted. What are some alternative ways to proceed? We present some possibilities.

The most difficult idea to digest in data analysis – and one that is rarely taught in linguistics and psychology – is that conclusions based on data are almost always uncertain, and this is regardless of whether the outcome of the statistical test is statistically significant or not. This uncertainty can and must be communicated when addressing questions of scientific interest. The perspective we take is that the focus in data analysis should be on estimation rather than (or only on) establishing statistical significance or the like (Cumming 2014; Thompson 2002; Wasserstein and Lazar 2016).

One suggestion in the statistics literature is to "accept uncertainty and embrace variation" (Gelman 2018). But what does embracing variation mean in practice? By revisiting several published data-sets that investigate agreement attraction (the phenomenon discussed above), we illustrate how the results from data analyses can be presented in such a way that the focus is on estimation and uncertainty quantification, rather than on drawing overly confident (and often invalid) conclusions. We present some alternative ways in which uncertainty can

be given the importance it deserves when summarizing the results of a (psycho) linguistic analysis.

2.1 A case study: agreement attraction effects

Consider again the agreement attraction effect discussed in the introduction. What do the data tell us about this effect? To illustrate different possible approaches, we will use 10 published studies' data; the data are available online as part of a larger meta-analysis, reported in Jäger et al. (2017). Approach 1 is the standard one, which we have criticized above. Approaches 2–4 are alternatives one can adopt; they are not intended to be mutually exclusive. One can use all of them together, depending on the situation.

2.1.1 Approach 1: standard significance-testing

Suppose that we were to carry out a standard frequentist linear mixed model analysis (Bates et al. 2015) of each of the 10 data-sets on agreement attraction. The *t*-values from such an analysis are shown in Table 1. Here, we could have carried out paired t-tests; but because all the data are available publicly, we were able to fit varying intercepts and varying slopes by participant and by item, without any correlation parameters (Barr et al. 2013; Bates et al. 2018).

What stands out in Table 1 is that although a few studies manage to cross the significance threshold of an absolute *t*-value of 2, the results do not look too convincing if we compare the number of significant effects (four) to the number of null results (six). One can think of these studies as replication attempts. In summary, under the conventional α of 0.05, we obtain four significant and six non-significant results (a 40% replication rate). This should count as the beginning of a full-blown replication crisis in psycholinguistics, much like the famous psychology replication attempt in which only about a third to half (depending on the replication criterion) of the studies could be replicated (Open Science Collaboration 2015). As discussed above, this approach is fairly meaningless, for the reasons explained above. We turn next to some more meaningful approaches.

Table 1: *t*-values from 10 published studies on the agreement attraction effect.

1	2	3	4	5	6	7	8	9	10
−2.56	−2.25	−1.67	−1.83	−1.40	−2.22	−1.33	−0.22	−2.81	−1.74

2.1.2 Approach 2: display the estimates with uncertainty intervals

There is a better way to summarize these results than in terms of significant versus non-significant results. Figure 4 shows the estimated means and 95% confidence intervals in log milliseconds of the agreement attraction effect in the 10 studies.

Using confidence intervals to summarize the results leads to two observations: First, the mean effect across the studies is consistently negative. Looking for such consistency across multiple studies is referred to by Gelman and Hill (2007) as the "secret weapon"; we will presently show (Approach 3) how to formalize this suggestion. The second important observation is the noisiness of the estimates. For example, on the log scale, the largest estimate (study 1) has an effect (back transformed to milliseconds) of –75 ms, and a 95% confidence interval spanning [–133, –16] ms. Such a large confidence interval suggests that under repeated sampling, the sample mean will be highly variable. Indeed, a larger-sample replication attempt of study 1 (181 participants as opposed to 40 participants in the original study) shows a much narrower interval and a smaller mean effect estimate: –22 [–46, 3] ms (Jäger et al. 2020). The difference between Approach 1 and 2 is that in *t*- or (equivalently) *p*-value based reasoning, we only focused on how many effects were significant; there was no discussion about the estimate of the

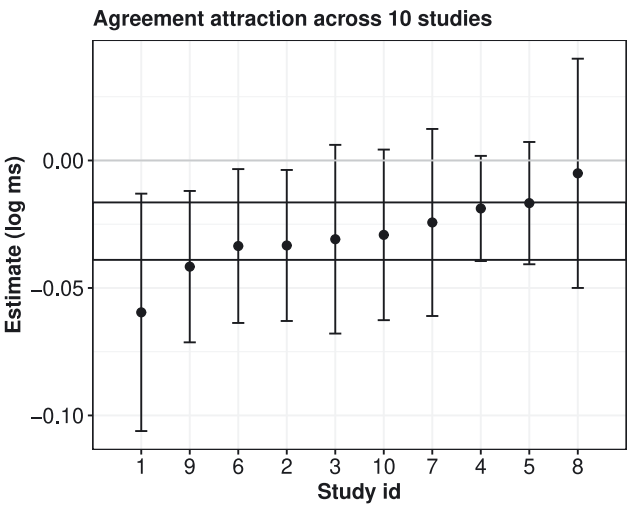

Figure 4: The mean agreement attraction effect and 95% confidence intervals from the frequentist analyses of 10 reading studies. The horizontal black lines show the 95% Bayesian credible interval of the meta-analysis estimate, computed by synthesizing the evidence from the 10 studies.

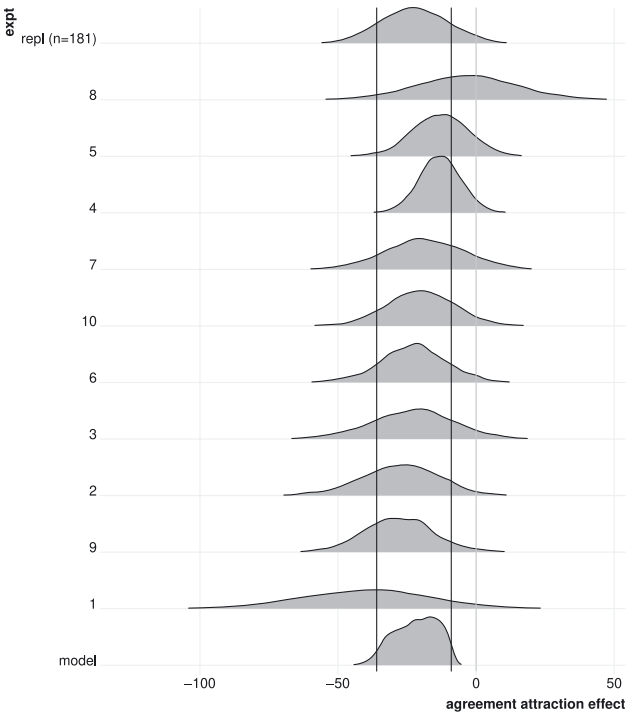

Figure 5: Ridgeplots showing the distributions of the effect of interest from 10 published reading experiments (eyetracking and self-paced reading) on agreement attraction; the studies are ordered by the magnitude of the mean effect. Also shown is the model's probability distribution of the predicted effect, computed using a large-sample (n = 181) data-set investigating agreement attraction (Engelmann et al. 2020; Jäger et al. 2020; Vasishth 2020); the model's prediction is labeled "model" in the figure. For reference, we also show the estimate of the agreement attraction effect in the large-sample study (this is labeled "repl (n = 181)"); this study was a replication attempt of study 1. The black vertical lines mark the 95% credible interval of a meta-analysis estimate computed using all published reading studies that were available in 2016 that investigated agreement attraction (Jäger et al. 2017).

magnitude of the effect and the uncertainty of the estimated difference in means. In Approach 2, the noisiness of the estimate is of central importance.

Even though the sample sizes in the 10 studies, given the experiment design and research question, are too small to give us sufficiently high power (Figure 2), by looking at the estimates and their 95% confidence intervals from the 10 studies side by side, we could still conclude that the data are at least consistent with the theoretical prediction that the effect should be negative in sign, with the qualification that the true value of the effect is likely to be much smaller, and therefore

strong conclusions should not be drawn from these data. The true effect is likely to be smaller because papers are generally published selectively based on significance, and if the studies reported are underpowered, Type M error becomes an issue.

2.1.3 Approach 3: conduct a meta-analysis

The graphically based reasoning we did above was an informal meta-analysis. It is possible to synthesize the information from the 10 studies formally. We can carry out a so-called random-effects meta-analysis (Gelman et al. 2014; Normand 1999; Sutton and Abrams 2001; Van Houwelingen et al. 2002). Such a meta-analysis produces an estimate of the effect given all the estimates from the studies, weighting (or partially pooling) each study's estimate by its uncertainty (standard error). The larger the standard error in a particular study, the less influence the study has on the meta-analysis mean.

Formally, a random-effects meta-analysis is simply a linear mixed model (Bates et al. 2015) with varying intercepts. We assume that the true unknown effect we are interested in (here, the agreement attraction effect) is the parameter θ. Given estimates d of the effect, along with their standard errors SE, from $i = 1, \ldots, n$ studies, assume that the observed estimates d are generated as follows:

$$d_i \sim Normal(\theta_i, SE_i)$$

θ_i refers to each individual study's true (unknown effect); this will differ from study to study due to differences in design, labs, protocols, etc., across the research groups conducting these studies. We further assume that the true effect θ generates these individual studies' true estimates, with some variability, represented by the standard deviation τ:

$$\theta_i \sim Normal(\theta, \tau)$$

For further details, and examples of meta-analyses in psycholinguistics, see Bürki et al. (2020), Jäger et al. (2017), Mahowald et al. (2016), and Nicenboim et al. (2018, 2020).

Figure 4 shows the meta-analysis confidence interval (black horizontal lines). These are actually not frequentist confidence intervals, but so-called Bayesian 95% credible intervals. They represent the range over which one can be 95% certain that the values of the effect lie, given the data and model. The 95% credible interval is going to be influenced by the data (if the data are biased, the interval will be too), and the model (if the model is incorrect, then this can affect the interval). So, being 95% certain about the range of plausible values of the effect doesn't necessarily entail that the interval reflects the true values of the effect. Nicenboim et al. (2018)

is a tutorial article that explains how to carry out such analyses, using an example from linguistics.

Thus, if data from multiple (low-powered) experiments exist, we can synthesize what we can learn from these via a meta-analysis. This is one way to realize the recommendation to "accept uncertainty and embrace variation" (Gelman 2018): focus on and interpret the uncertainty of the estimate from the accumulated evidence before drawing any firm conclusions about the effect. The meta-analysis estimates in Figure 4 show that the mean agreement attraction effect on the millisecond scale is −35, with 95% credible interval [−49, −21] ms. This estimate is consistent with the theoretical claim of a speedup. Whether this amounts to a discovery claim, i.e., whether there is evidence in favor of an effect, requires much more investigation, using formal hypothesis testing tools such as Bayes factors (Kruschke 2014; Schad et al. 2021; Wagenmakers et al. 2018).

Once we have such a theoretically predicted range of effects, we can use it to interpret future data. We turn to this approach next.

2.1.4 Approach 4: use a region of practical equivalence

Sometimes, quantitative predictions for an effect are available. These could be the meta-analysis estimates available from existing work, or they could be derived from a computational model. Figure 5 shows the estimates from a larger meta-analysis than the one done above (Jäger et al. 2017), as well as the predicted range of effects from the computational model for agreement attraction mentioned earlier (Jäger et al. 2020; Vasishth 2020). In Figure 5, the meta-analysis range is shown as black vertical lines and the model predictions and the estimates from the individual studies are shown as probability distributions.

Given the model's predicted range of values for the agreement attraction effect, we can see that the meta-analysis estimate, and estimates from the 10 studies are consistent with the predicted range. The meta-analysis credible interval overlaps almost exactly with the model's predictions. From this, we would conclude that the evidence from published studies on agreement attraction is at least consistent with model predictions. A future study could use the model's predictions as well as the meta-analysis estimates to interpret their data in the context of the theory's predictions.

Comparing the estimates derived from individual studies to a predicted range of effects is not a new idea (Freedman et al. 1984; Spiegelhalter et al. 1994). In recent years, this idea has been re-introduced into psychology by Kruschke (2014) as the region of practical equivalence (ROPE) approach. The essential idea behind interpreting data using a ROPE is summarized in Figure 6. Assume that we have a model prediction spanning [−36, −9] ms; this is in fact the model prediction

reported in Jäger et al. (2020). Then, if we aim to run our experiment until we have the same width as the predicted range (here, −9−(−36) = 27 ms), then there are five possible intervals that can be observed. These uncertainty intervals are not frequentist confidence intervals, but Bayesian 95% credible intervals; they demarcate plausible ranges of values for the effect, given the model and data.

The observed uncertainty interval can be:

A. entirely to the right of the predicted interval.
B. entirely to the left of the predicted interval.
C. to the right of the predicted interval but overlapping with it.
D. to the left of the predicted interval but overlapping with it.
E. within the predicted range (this is the case in Figure 5).

Only situation E shows a convincing consistency with the quantitative prediction. A and B are inconsistent with the model prediction; and C and D are also consistent with the quantitative prediction, but unlike E are inconclusive. If, for some reason, one cannot reach the desired precision (width of 27 ms), there is a sixth possibility: the observed interval may overlap with the predicted range but may be much wider than it (here, the width of the predicted range is 27 ms). That would be an uninformative, low-precision study.

In contrast to the region of practical equivalence approach described above, what linguists usually predict is the sign of an effect, but they do not attend to the magnitude or the uncertainty. But a prediction like "the effect will be negative in sign" is not particularly useful because this implies that an effect with mean −500 ms that is statistically significant would validate the prediction just as well as a significant −10 ms effect. As discussed above, under low power, statistically significant large effects are very unlikely to be an accurate estimate due to Type M error (Gelman and Carlin 2014).

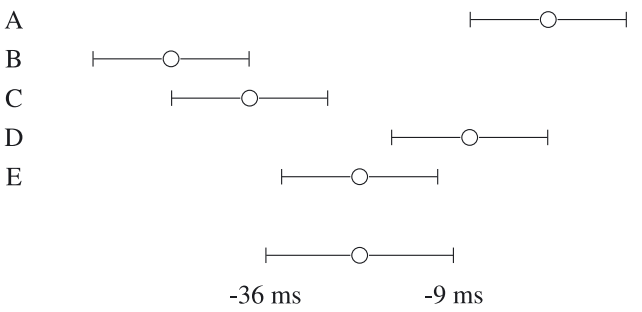

Figure 6: The five possible outcomes when using the null region or "region of practical equivalence" method for decision-making (Kruschke 2014). Outcomes A and B are inconsistent with the quantitative predictions of the theory; C and D are inconclusive; and E is consistent with the quantitative theoretical prediction.

The region of practical equivalence approach is also relevant to more general issues relating to model/theory evaluation. As Roberts and Pashler (2000) have pointed out in their classic article, a vague theoretical prediction (e.g., "the effect is predicted to have a negative sign") and/or a very uncertain estimate from the data (an effect with a very wide 95% confidence interval) both lead to very weak support for the theoretical prediction. In psychology and linguistics, the Roberts and Pashler (2000) discussion on what constitutes a persuasive evaluation of a model has not yet received the attention it deserves. The essential idea in their paper is summarized in Figure 7. A vague theory will allow a broad range of predictions, and a data-set which has a lot of uncertainty associated with the estimate will be uninformative when testing a prediction. In order to argue that the data are consistent with a theory, it is necessary to have both a constrained quantitative prediction, and a high-precision estimate of the effect.

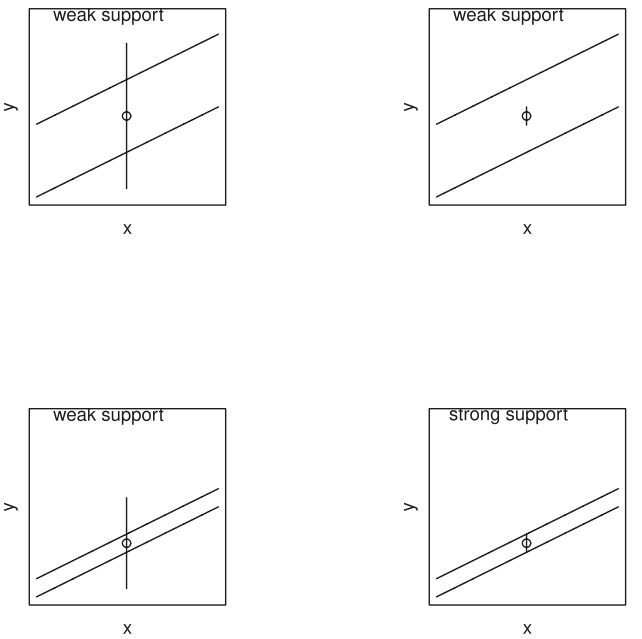

Figure 7: A schematic summary of the Roberts and Pashler (2000) discussion regarding what constitutes a good fit of a model to data. If a model predicts a positive correlation between two variables *x* and *y*, the strong support for the model can only be argued for if both the data and the model predictions are highly constrained: the model must make precise predictions, and the data must have low uncertainty associated with it. The figure source: 10.6084/m9.figshare.14241869.

In summary, with the region of practical equivalence approach, the focus is on graphically visualizing the uncertainty of the estimates from different experiments, *with reference to a predicted range of effects*. The Roberts and Pashler (2000) criteria for deciding what constitutes a good fit is closely related to this approach, because they also place the focus on the range of quantitative predictions made by the model, and the uncertainty associated with the estimate of the effect in the data.

3 Planning future studies using available information

One important point that we emphasized in the above discussion is the importance of running an informative experiment (when feasible). This involves ensuring that there is as little measurement error as possible (Loken and Gelman 2017), that the experiment design is thought out well so as to have a good chance of detecting the effect (Gelman and Carlin 2014), and that sample size (number of participants and items) is high enough to have a reasonably good chance of detecting the effect of interest (Cohen 1988).

In practice, how can one plan a study such that one ends up with an informative experiment? One approach, which focuses on achieving a tight enough confidence interval to be informative for the research question at hand, is to define a ROPE based on a meta-analysis, quantitative predictions from a model, or expert judgement (O'Hagan et al. 2006). For an example using judgement about expected ranges of effect sizes for deciding on a sample size, see Vasishth et al. (2018). Another possible approach is Bayes factor design analysis (Schönbrodt and Wagenmakers 2018); for an example, see Montero-Melis et al. (2019) (although the way that these authors compute Bayes factors is not really appropriate; for further details, see Schad et al. (2021)). The adaptive Bayesian methods developed for clinical trials (Berry et al. 2010; Spiegelhalter et al. 2004) also have a lot of potential applications in linguistics and psychology.

An alternative (purely frequentist) approach to ensuring that one has precise enough estimates is to conduct a power analysis. One can use quantitative predictions based on a meta-analytic estimate in the following way for planning a future study. As an example, consider a study with two conditions. We want to plan a new, higher-powered study, assuming that a meta-analysis estimate (along with its 95% confidence interval) reflects the best guess we have of the effect. We proceed as follows (all the code for reproducing these analyses is shown in the appendix):

1. Extract all the parameter estimates from the linear mixed model used to analyze an existing study (or studies, if more than one is available). This includes all the variance components estimated by the linear mixed model.
2. Using the above estimates, generate simulated data 100 times (or more) repeatedly using the meta-analysis estimates. Using such a simulation, compute the proportion of times that the null hypothesis is rejected; this gives us the estimated range of power for the meta analysis mean and 95% confidence interval.
3. Use the above simulation technique to work out the range of participants that would be needed to achieve at least 80% power.

When we carry out such a simulation-based computation using study 1's data, what we find is that for the sample size of 40 participants and 48 items in study 1, our estimated power ranges from 0.25 to 0.77. We can now easily compute the power for, e.g., 300 participants: for the mean estimate from the meta-analysis, the estimated power is 1, with lower and upper bounds ranging from 0.77 to 1. The wide range of uncertainty in the power calculation arises due to the uncertainty implied by the 95% confidence interval of the meta-analysis estimate.

We carried out the power analysis above "by hand", i.e., by writing custom code that generated simulated data. There are ready-made functions/packages available that can automate the process to a large extent: see the packages `simr` (Green et al. 2021) and `designr` (Rabe et al. 2021). Accessible tutorials for automating the simulation-based power computation process are also available (Brysbaert and Stevens 2018; DeBruine and Barr 2021). For a Bayesian perspective on power analysis, see Kruschke and Liddell (2018) and Schad et al. (2021).

Our discussion here is heavily focused on statistical power. Of course, power is not the only important issue in experimental science: other factors like measurement error and a strong theoretical foundation are also very important. But it is important to understand that without adequate power, the significance testing paradigm breaks down. This is why power calculations need to become an integral part of the data analysis workflow.

4 Some potential objections

We encounter various objections to the points we have raised in this paper. We discuss some of these next.

4.1 Is there a danger of "uncertainty overshoot"?

"Uncertainty overshoot" could be a danger: we may become overly conservative when drawing conclusions from data. In the practical running example in this paper, we have discussed the conditions under which strong support for a theory can be argued for: both the theory and the data have to be sufficiently informative (Roberts and Pashler 2000). In all other situations, uncertainty undershoot is not very likely; far more likely is "certainty overshoot". In practice, what we see in the literature are over-confident claims that fail to be validated upon closer scrutiny.

4.2 Will over-cautious reporting make papers difficult to publish?

Researchers sometimes object to proposals demanding weaker claims in published articles with the argument that it would make papers more difficult to publish if one does not make a decisive claim. We consider it a questionable research practice to make a decisive claim when none is warranted statistically. Nevertheless, these concerns do have some basis: sometimes journals, editors, and reviewers explicitly state that they want certainty or "closure" in a paper, and that expressing uncertainty about the conclusions does sometimes lead to rejection. However, our experience in recent years has been that the situation is changing. Editors and reviewers have started to appreciate open discussion of uncertainty, especially if one has done one's best to get to the facts (e.g., through many replication attempts, or large sample studies; usually both). Here are some examples of papers that explicitly express uncertainty about the findings and were nevertheless published in a major psycholinguistics journal:

- In Vasishth et al. (2018), one out of seven experiments showed an effect that was consistent with a theoretical claim, but was nevertheless unexpected because no other study had found such an effect in the language. In the conclusion, the authors wrote:

  > One interesting suggestion from this 100-participant study is that the … effect that is predicted by [the theoretical account under study] may have some weak support. Since this is, to our knowledge, the first time that any evidence for [the theoretical claim] has been seen in German, clearly further investigation is needed.

- In a single large-sample eyetracking study reported in Jäger et al. (2020), in total reading times the authors found effect estimates consistent with a particular theory of sentence processing. But in first-pass regressions, they

also found effects not consistent with this theory's predictions. It is not clear which dependent measure one should rely on. Accordingly, in the paper, the authors openly discuss the support (or lack thereof) for the theoretical claim, conditional on the dependent measure considered. The paper does not end with a clear conclusion.

Other researchers have also published papers in which power analyses were carried out to understand the properties of the experiment design, and/or the results framed appropriately without overly strong conclusions. Some examples are Montero-Melis et al. (2017, 2019), Xie and Jaeger (2020), and Xie et al. (2021).

Despite all the positive developments exemplified above, papers (including those from the first author's lab) do continue to be rejected for not providing sufficiently conclusive results. We hope that this situation will change some day. A major goal of the present paper is to help towards normalizing openness in expressing our uncertainty about our conclusions. The alternative to maintaining uncertainty about our conclusions is a proliferation of exaggerated conclusions that will probably not hold up to closer scrutiny. This is in fact what has happened in social psychology and other areas: claims have been published that are non-replicable. Linguistics can learn from these past mistakes in other fields, and develop a culture of accepting and quantifying uncertainty about the conclusions that can be drawn from a particular study.

It is important to stress here that our point is not that researchers should only publish high-powered studies. Often, it is impossible to run a sufficiently powered study; examples are experiments involving field work in remote regions of the world, and studies on aphasia. Science is an incremental process, and eventually enough information can accumulate (e.g., through meta-analyses) about a research topic. As Simmons et al. (2011) and many others have pointed out, open availability of data, and reproducible code and analyses, will be important drivers such an incremental evidence-accumulation process.

Our goal in this paper is merely to stress the point that we should not present underpowered studies as furnishing clear evidence for or against a theoretical claim, otherwise we risk flooding the field with non-replicable results.

4.3 Will increasing the number of replicates per subject and keeping sample size small solve the power problem?

Psychologists (Smith and Little 2018) have recommended so-called small-N studies as a response to the replication crisis: obtain many repeated measurements from each participant, but use only a few participants. This approach can be effective in

obtaining accurate estimates in certain specific types of scientific inquiries; for example, Ebbinghaus discovered several laws of memory with a single subject (himself), and in vision science it is common to use only a few participants. Small-N studies only make sense when it is known that between-subject variability is low or the effect size is so large that the effect is easy to detect. This situation only rarely arises in linguistics and psycholinguistics. One extreme example where a small-N study would yield robust (i.e., replicable) results is asking subjects to rate the acceptability of strings like *The boy saw the girl* and *Boy the girl the saw*. Most linguistic and psycholinguistic studies today investigate much more subtle questions with small effect sizes, and these can show important between-subject variability. In such cases, if the goal is to generalize from a sample to a population, there is no way around properly powering the study if one wants to obtain accurate estimates.

4.4 Can some effects already be detected with small sample studies?

There is a commonly-encountered fallacy relating to sample sizes that Loken and Gelman (2017) summarize as "that which does not kill statistical significance makes it stronger." Some researchers think that if one observes a significant effect with a small sample size, that effect is all the more convincing. For example, Kuang et al. (2007) state (footnote 11): "… the fact that we get significant differences in spite of the relatively small samples provides further support for our results."

Such misunderstandings can easily be dispelled through simulation-based investigation of one's experiment design. To take a concrete example, Gibson and Wu (2013) obtained a significant effect in a two-condition repeated measures design with a sample size of 37 participants and 15 items (originally there were 16 items, but one item was removed). One can adapt the simulation code shown in the appendix to establish that the significant effect was likely a Type M error, arising from an underpowered study. In the Gibson and Wu (2013) study, the estimate of the difference between the conditions was approximately 120 ms (they analyzed the data on the raw ms scale; we follow their approach here). Although this estimate is larger than the approximately 100 ms difference found in English relative clauses (Grodner and Gibson 2005), let's assume that the true difference between relative clause processing times in Chinese is in fact 120 ms. If we were to repeatedly sample data from such a design (the appendix shows how this can be done), with sample size 40 subjects and 16 items, we would find that almost all the statistically significant effects are driven by effect estimates larger than 120 ms. As shown in Figure 8, 89% of the significant effects are based on overestimates of the

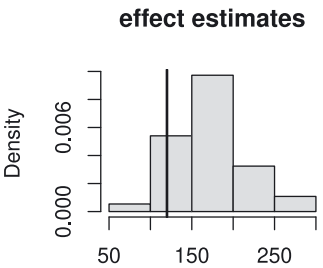

Figure 8: The distribution of effect estimates that are statistically significant when the true value of the effect in the Gibson and Wu (2013) data-set is 120 ms (shown by the vertical line). The histogram shows that most of the estimates that are statistically significant under repeated sampling using simulated data are overestimates, i.e., they are Type M errors.

effect (the significant estimates can be as much as 2.4 times larger than 120 ms). If the true effect had been 60 ms, the probability of overestimating the effect size given a significant result is 100%, with the estimate being as much as 3.2 times larger than 60 ms. This kind of simulation is an easy way to establish that a significant effect based on a small sample size is not very convincing, because it is based on an overestimated effect size.

In summary, the importance of power cannot be stressed enough. Power should be seen as the ball in a ball game; it is only a very small part of the sport, because there are many other important components. But the players would look pretty foolish if they arrive to play on the playing field without the ball. Of course, power is not the only thing to consider in an experiment; no amount of power will help if the design is confounded or introduces a bias in some way.

5 Concluding remarks

We have argued that statistical analyses in linguistics and related areas should follow the best practice recommendations of statisticians and psychologists: they should focus on uncertainty quantification rather than just conducting null hypothesis significance testing and drawing overly strong conclusions from data. We presented specific examples that showed how this could be done in practice, and the advantages that come with using such an approach as regards theory evaluation.

Acknowledgments: Thanks go to Garrett Smith, Kate Stone, João Veríssimo, Dorothea Pregla, Lukas Sönning, Reinhold Kliegl, Timo Roettger, Daniela Mertzen, Bruno Nicenboim, and Titus von der Malsburg for useful comments. We also thank the reviewers, among them Florian Jaeger, for useful feedback. The research reported here was partly funded by the Deutsche Forschungsgemeinschaft (German Science Foundation), Collaborative Research Center - SFB 1287, project number 317633480 (*Limits of Variability in Language*) through project Q (PIs: Shravan Vasishth and Ralf Engbert), and the US Office of Naval Research, through grant number N00014-19-1-2204.

Appendix A

Generating simulated data to compute power

Here we provide code for generating simulated data, and for computing power for a two-condition experiment.

A.1 Function for generating simulated data

First, we write a function for producing data from a Normal likelihood, assuming a varying intercepts and varying slopes model, for participants and items. The underlying model assumed is as follows.

Let j index participant id, and let k index item id. The variable cond is a sum-coded contrast (Schad et al. 2020), where +1/2 represents one condition a and –1/2 the other condition b. Thus, a negative sign on the β coefficient would be consistent with a theoretical prediction of a speedup in condition a versus b.

$$y_{kj} \sim Normal(\alpha + u_{0j} + w_{0k} + (\beta + u_{1j} + w_{1k}) \times cond_{kj}, \sigma)$$

with the following sources of variability:

- $u_{0j} \sim Normal(0, \sigma_{u0})$
- $u_{1j} \sim Normal(0, \sigma_{u1})$
- $w_{0k} \sim Normal(0, \sigma_{w0})$
- $w_{1k} \sim Normal(0, \sigma_{w1})$

Here, we are assuming no correlation between the varying intercepts and slopes; if one wants to assume such a correlation, one can easily modify the code. See Jäger et al. (2020) and Vasishth et al. (2018) for example code.

Data from the above model can be generated using the following function:

```
library(MASS)
gen_fake_norm <- function(nitem=NULL,nsubj=NULL,
                          alpha=NULL,beta=NULL,
                          sigma_u0=NULL,
                          sigma_u1=NULL,
                          sigma_w0=NULL,
                          sigma_w1=NULL,
                          sigma_e=NULL){
## prepare data frame for two condition in a Latin square design:
g1<-data.frame(item=1:nitem,
               cond=rep(c("a","b"),nitem/2))
g2<-data.frame(item=1:nitem,
               cond=rep(c("b","a"),nitem/2))

## assemble data frame in long format:
gp1<-g1[rep(seq_len(nrow(g1)),
            nsubj/2),]
gp2<-g2[rep(seq_len(nrow(g2)),
            nsubj/2),]

fakedat<-rbind(gp1,gp2)

## add subjects:
fakedat$subj<-rep(1:nsubj,each=nitem)
fakedat<-fakedat[,c(3,1,2)]
## contrast coding:
fakedat$cond<-ifelse(fakedat$cond=="a",1/2,-1/2)

## subject random effects:
u0<-rnorm(n=length(unique(fakedat$subj)),
            mean=0,sd=sigma_u0)
u1<-rnorm(n=length(unique(fakedat$subj)),
            mean=0,sd=sigma_u1)
## item random effects
w0<-rnorm(n=length(unique(fakedat$item)),
            mean=0,sd=sigma_w0)
w1<-rnorm(n=length(unique(fakedat$item)),
            mean=0,sd=sigma_w1)

## generate data row by row:
N<-dim(fakedat)[1]
rt<-rep(NA,N)
for(i in 1:N){
  rt[i] <- rnorm(1,alpha +
                     u0[fakedat[i,]$subj] +
                     w0[fakedat[i,]$item] +
                     (beta+u1[fakedat[i,]$subj]+
                     w1[fakedat[i,]$item])*fakedat$cond[i], sigma_e)}
```

(continued)

```
fakedat$rt<-rt
fakedat$subj<-factor(fakedat$subj)
fakedat$item<-factor(fakedat$item)
fakedat
}
```

A.2 Extract parameter estimates from fitted model

Given a data-set `dat` containing a predictor `cond` with two levels, fit a so-called maximal model (Barr et al. 2013), and then write a function to extract all parameter estimates from the model as a list.

An example data-set is the Gibson and Wu (2013) Chinese relative clause data, which has 37 participants and two conditions, subject and object relatives. Originally, there were 16 items, but one was removed by the authors, leaving 15 items. We analyze the data on the log ms scale because the normality assumption of the residuals is violated with raw reading times.

First, we load and pre-process the data, and choose the relevant subset of the data for analysis (this is the head-noun region in the sentence; see Gibson and Wu (2013) for details).

```
## install from: https://github.com/vasishth/lingpsych
library(lingpsych)
data("df_gibsonwu")
## sum-contrast coding of predictor:
gw$cond <- ifelse(
gw$type%in%c("subj-ext"),-1/2,1/2)
## subset critical region
dat<-subset(gw,region=="headnoun")
```

Next, we fit a linear mixed model, with a full variance-covariance matrix. This model is overparameterized: the correlation parameters are not estimable. The reason we include the correlations in the model even though they are not estimable is just for convenience in extracting the variance components: the `extract_pars-ests_lmer` function below happens to assume a full variance-covariance matrix model. In our simulations below, we will not attempt to estimate the correlation parameters when we repeatedly generate simulated data.

```
m<-lmer(log(rt) cond+(1+cond|subj)+(1+cond|item),dat,
        control=lmerControl(calc.derivs=FALSE))

## function for extracting all parameter estimates:
extract_parests_lmer<-function(
  mod=m){
  alpha<-summary(mod)$coefficients[1,1]
  beta<-summary(mod)$coefficients[2,1]
## extract standard deviation estimate:
sigma_e<-attr(VarCorr(mod),"sc")
## assemble variance covariance matrix for subjects and items:
subj_ranefsd<-attr(VarCorr(mod)$subj,"stddev")
sigma_u0<-subj_ranefsd[1]
sigma_u1<-subj_ranefsd[2]
item_ranefsd<-attr(VarCorr(mod)$item,"stddev")
sigma_w0<-item_ranefsd[1]
sigma_w1<-item_ranefsd[2]
## return list of params:
list(alpha=alpha,beta=beta,sigma_e=sigma_e,
    sigma_u0=sigma_u0,sigma_u1=sigma_u1,
     sigma_w0=sigma_w0,sigma_w1=sigma_w1)
}
```

The usage of this function will take as input the model that we want to extract the parameters from:

parest<-extract_parests_lmer(mod=m)

A.3 Function for computing power

Next, we write a function, `compute_power`, that (i) takes the parameter estimate values extracted above, (ii) generates simulated data using the `gen_fake_norm` function shown above, with 48 subjects and 40 items, (iii) fits a linear mixed model to the simulated data, (iv) extracts the *t*-value of the effect from the model, and (v) computes the proportion of absolute *t*-values that are larger than the critical value of 2. This is our estimated power.

```
compute_power<-function(nsim=100,
                        alpha=parest$alpha,
                        beta=parest$beta,
                        sigma_e=parest$sigma_e,
                        sigma_u0=parest$sigma_u0,
                        sigma_u1=parest$sigma_u1,
sigma_w0=parest$sigma_w0,
```

(continued)

```
                        sigma_w1=parest$sigma_w1,
                        nsubj=48,
                        nitem=40){
tvals<-c()
for(i in 1:nsim){
fakedat<-gen_fake_norm(nitem=nitem,
                       nsubj=nsubj,
               alpha=alpha,
               beta=beta,
               sigma_u0=sigma_u0,
               sigma_u1=sigma_u1,
               sigma_w0=sigma_w0,
               sigma_w1=sigma_w1,
               sigma_e=sigma_e)
m<-lmer(rt cond+(1+cond||subj)+(1+cond||item),
        fakedat,
        control=lmerControl(calc.derivs=FALSE))
tvals[i]<-summary(m)$coefficients[2,3]
}
mean(abs(tvals)>2)
}
```

The function can now be used as follows. Suppose we want to know what the power is for an effect size of −0.02 (log ms scale) given our sum-contrast parameterization.

```
compute_power(beta=-0.02)
```

One can compute the power for different sample sizes (number of participants or items):

```
compute_power(nsubj=50,beta=-0.02)
compute_power(nitem=80,beta=-0.02)
```

The code shown above can easily be extended for more complex models and for different likelihoods. For examples, see Jäger et al. (2020) and Vasishth et al. (2018). A more sophisticated Bayesian approach is discussed in Schad et al. (2021).

References

Arunachalam, Sudha. 2013. Experimental methods for linguists. *Language and Linguistics Compass* 7(4). 221–232.

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

Bates, Douglas M., Martin Maechler, Ben M. Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67. 1–48.

Bates, Douglas M., Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2018. Parsimonious mixed models. Unpublished manuscript. https://arxiv.org/abs/1506.04967.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, Eric-Jan Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Held Leonhard, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Imbens Guido, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Rouder Jeff, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman & Valen E. Johnson. 2018. Redefine statistical significance. *Nature Human Behaviour* 2(1). 6–10.

Berry, Scott M., Bradley P. Carlin, J. Jack Lee & Peter Muller. 2010. *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: CRC Press.

Brehm, Laurel E. & Matthew Goldrick. 2017. Distinguishing discrete and gradient category structure in language: Insights from verb-particle constructions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43(10). 1537–1556.

Brysbaert, Marc & Michael Stevens. 2018. Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition* 1(1). 9.

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

Cohen, Jacob. 1962. The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology* 65(3). 145–153.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edn. Hills-dale, NJ: Lawrence Erlbaum.

Cumming, Geoff. 2014. The new statistics: Why and how. *Psychological Science* 25(1). 7–29.

DeBruine, Lisa M. & Dale J. Barr. 2021. Understanding mixed effects models through data simulation. *Advances in Methods and Practices in Psychological Science* 4. 1–15.

Draper, Norman R. & Harry Smith. 1998. *Applied regression analysis*. New York: Wiley.

Engelmann, Felix, Lena A. Jäger & Shravan Vasishth. 2020. The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science* 43. e12800.
Freedman, Laurence S., D. Lowe & Petra Macaskill. 1984. Stopping rules for clinical trials incorporating clinical opinion. *Biometrics* 40(3). 575–586.
Gelman, Andrew. 2018. Ethics in statistical practice and communication: Five recommendations. *Significance* 15(5). 40–43.
Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
Gelman, Andrew & John B. Carlin. 2014. Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651.
Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Vehtari Aki & Donald B. Rubin. 2014. *Bayesian data analysis*, 3rd edn. Boca Raton, FL: Chapman and Hall/CRC Press.
Gelman, Andrew & Sander Greenland. 2019. Are confidence intervals better termed uncertainty intervals? *British Medical Journal* 366. l5381.
Gibson, Edward & H.-H. Iris Wu. 2013. Processing Chinese relative clauses in context. *Language and Cognitive Processes* 28(1–2). 125–155.
Green, Peter, Carriona MacLeod & Phillip Alday. 2021. *simr: Power analysis for generalised linear mixed models by simulation*. R package version 1.0.5.
Greenland, Sander. 2017. Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology* 186(6). 639–645.
Grodner, Daniel & Edward Gibson. 2005. Consequences of the serial nature of linguistic input. *Cognitive Science* 29. 261–290.
Higgins, Julian & Sally Green. 2008. *Cochrane handbook for systematics reviews of interventions*. New York: Wiley-Blackwell.
Hoekstra, Rink, Richard D. Morey, Jeffrey Rouder & Eric-Jan Wagenmakers. 2014. Robust misinterpretations of confidence intervals. *Psychonomic Bulletin and Review* 21. 11571164.
Jäger, Lena A., Daniela Mertzen, Julie A. Van Dyke & Shravan Vasishth. 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language* 111. 104063.
Jäger, Lena A., Felix Engelmann & Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language* 94. 316–339.
Kruschke, John. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General* 142(2). 573–603.
Kruschke, John. 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London: Academic Press.
Kruschke, John & Torrin M. Liddell. 2018. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25(1). 178–206.
Kuang, Xi Jason, Roberto A. Weber & Jason Dana. 2007. How effective is advice from interested parties?: An experimental test using a pure coordination game. *Journal of Economic Behavior & Organization* 62(4). 591–604.
Loken, Eric & Andrew Gelman. 2017. Measurement error and the replication crisis. *Science* 355(6325). 584–585.
Mahowald, Kyle, Ariel James, Richard Futrell & Edward Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language* 91. 5–27.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert & Jennifer L. Tackett. 2019. Abandon statistical significance. *The American Statistician* 73(1 Suppl). 235–245.

Montero-Melis, Guillermo, Jeroen van Paridon, Markus Ostarek & Emanuel Bylund. 2019. Does the motor system functionally contribute to keeping words in working memory? A pre-registered replication of Shebani and Pulvermüller (2013, cortex). https://doi.org/10.31234/osf.io/pqf8k.

Montero-Melis, Guillermo, Sonja Eisenbeiss, Bhuvana Narasimhan, Iraide Ibarretxe-Antuñano, Sotaro Kita, Anetta Kopecka, Friederike Lüpke, Tatiana Nikitina, Ilona Tragel, T. Florian Jaeger & Jürgen Bohnemeyer. 2017. Satellite- versus verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations. *Cognitive Semantics* 3(1). 36–61.

Navarro, Daniel. 2013. *Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.5*. Australia: University of Adelaide Adelaide.

Nicenboim, Bruno, Shravan Vasishth & Rösler Frank. 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.

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

Normand, Sharon-Lise T. 1999. Tutorial in biostatistics meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine* 18(3). 321–359.

O'Hagan, Anthony, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garth-Waite, David J. Jenkinson, Jeremy E. Oakley & Tim Rakow. 2006. *Uncertain judgements: Eliciting experts' probabilities*. New York: John Wiley & Sons.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). aac4716.

Phillips, Colin, Matthew W. Wagers & Ellen F. Lau. 2011. Grammatical illusions and selective fallibility in real-time language comprehension. In Jeffrey Runner (ed.), *Experiments at the interfaces* (Syntax and Semantics 37), 147–180. Leiden: Brill.

Rabe, Maxmilian, Reinhold Kliegl & Daniel J. Schad. 2021. *designr: Balanced factorial designs*. R package version 0.1.11.

Roberts, Seth & Harold Pashler. 2000. How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107(2). 358–367.

Roettger, Timo B., Bodo Winter & Harald Baayen. 2019. Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics* 73. 1–7.

Schad, Daniel J., Nicenboim Bruno, Paul-Christian Bürkner, Michael Betancourt & Shravan Vasishth. 2021. Workflow techniques for the robust use of Bayes factors. arXiv: 2103.08744v2.

Schad, Daniel J., Shravan Vasishth, Sven Hohenstein & Reinhold Kliegl. 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language* 110. 104038.

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

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11). 1359–1366.

Smith, Philip L. & Daniel R. Little. 2018. Small is beautiful: In defense of the small-n design. *Psychonomic Bulletin & Review* 25(6). 2083–2101.

Spiegelhalter, David J., Keith R. Abrams & Jonathan P. Myles. 2004. *Bayesian approaches to clinical trials and health-care evaluation*, vol. 13. New York: John Wiley & Sons.

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

Stack, Caoimhe M. Harrington, Ariel N. James & Duane G. Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition* 46(6). 864–877.

Sutton, Alex J. & Keith R. Abrams. 2001. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 10(4). 277–303.

Thompson, Bruce. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31(3). 25–32.

Van Houwelingen, Hans C., Lidia R. Arends & Theo Stijnen. 2002. Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine* 21(4). 589–624.

Vasishth, Shravan. 2020. Using approximate Bayesian computation for estimating parameters in the cue-based retrieval model of sentence processing. *MethodsX* 7. 100850.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.

Vasishth, Shravan, Nicenboim Bruno, Felix Engelmann & Burchert Frank. 2019. Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences* 23. 968–982.

Verhagen, Josine & Eric-Jan Wagenmakers. 2014. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General* 143(4). 1457.

Wagenmakers, Eric-Jan, Maarten Marsman, Tahira Jamil, Ly Alexander, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F. Gronau, Smíra Martin, Sacha Epskamp, Dora Matzke, Jeffrey N. Rouder & Morey Richard D. 2018. Bayesian inference for psychology. Part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review* 25(1). 35–57.

Wagers, Matthew W., Ellen F. Lau & Colin Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61(2). 206–237.

Wasserstein, Ronald L. & Nicole A. Lazar. 2016. The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician* 70(2). 129–133.

Winter, Bodo. 2019. *Statistics for linguists: An introduction using R*. New York, NY: Routledge.

Xie, Xin, Linda Liu & T. Florian Jaeger. 2021. Cross-talker generalization in the perception of non-native speech: A large-scale replication. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001039 (Epub ahead of print).

Xie, Xin & T. Florian Jaeger. 2020. Comparing non-native and native speech: Are L2 productions more variable? *The Journal of the Acoustical Society of America* 147(5). 3322–3347.

Yarkoni, Tal. 2020. The generalizability crisis. *The Behavioral and Brain Sciences*. 1–37. https://doi.org/10.1017/s0140525x20001685.

Zormpa, Eirini, Antje S. Meyer & Laurel E. Brehm. 2019. Slow naming of pictures facilitates memory for their names. *Psychonomic Bulletin & Review* 26. 1–8.