DE GRUYTER Linguistics ...; ..(..):1–29

...

Shravan Vasishth* and Andrew Gelman

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

Embrace variation

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 significant results are driven by Type M error. 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 to learn 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: Uncertainty quantification; experimental linguistics

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 (highly) significant/robust effect of factor X on dependent variable Y." If researchers fail to find a significant effect, too often they

***Corresponding author: Shravan Vasishth,** Department of Linguistics, University of Potsdam, Potsdam, Germany
Andrew Gelman, Department of Statistics, Columbia University, New York, USA

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 we are trained 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. This way of thinking is fundamentally incorrect, a distortion of the underlying theory.

A major reason for these misunderstandings stems from the prefunctory education provided in statistics, in both linguistics and psychology programs worldwide. Learning statistical theory and practice are inseparable from scientific reasoning; and contrary to an increasingly popular belief in linguistics, experimentally grounded research is no guarantee that research will become more grounded in objective facts as opposed to subjective beliefs. 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. 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 book covering the theory that is specifically aimed at linguists is available (Winter 2019); also see the special issue *Emerging Data Analysis in Phonetic Sciences* edited by Timo Roettger, Bodo Winter, and Harald Baayen (https://bit.ly/2DsF8M6).

1.1 The logic of significance testing

The standard logic of significance-based testing is straightforward and 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, Lau, and Phillips 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$^{-plural}_{+subject}$ who met the trainers$^{+plural}_{-subject}$ were$\{^{plural}_{subject}\}$ ...

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

Several theoretical explanations have been proposed to account for this observed speedup. One of them (Engelmann, Jäger, and Vasishth 2019; Vasishth, Nicenboim, et al. 2019) 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. Whether this is the correct theoretical account or not is debatable (Hammerly, Staub, and Dillon 2019), but that is beside the point here. What is important is that a speedup is predicted at the auxiliary verb in (1a) vs. (1b), and that there exists at least one computational model that can produce quantitative predictions of this agreement attraction effect (Engelmann, Jäger, and Vasishth 2019; Vasishth, Nicenboim, et al. 2019).

Thus, based on the quantitive predictions (shown later, in Figure 5) of the model reported in Engelmann, Jäger, and Vasishth (2019), the research hypothesis is that the auxiliary verb in (1a) will be read faster than in (1b). The statistical test of this hypothesis is 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 $H_0 : \delta = \mu_a - \mu_b = 0$.

Even though the research hypothesis is that the sign of the difference is negative, the standard procedure is to set up a so-called two-sided hypothesis test: this assumes that the alternative to the null hypothesis is that the difference between the two means, δ, is not equal to 0 (it can be positive or negative). There is a very good reason for this convention of doing a two-sided test even if the research hypothesis is one-sided: in general, one cannot completely rule out the possibility that the sign of the effect is in the opposite direction to that predicted by theory. This becomes clearer when we consider a medical example: we might want to establish whether a particular drug is beneficial for patients; there may even be a theoretical prediction that leads to this expectation. But we usually cannot rule out a priori the possibility that the drug may turn out to be harmful. As Pocock

(2013, p. 206) puts it: "..., the use of one-sided tests is generally inappropriate since it prejudges the direction of treatment difference (usually new treatment better than standard) and there have been many trials where a new treatment fared worse." Thus, when we cannot categorically rule out the possibility that the sign of the effect could be either positive or negative, a two-sided test must be carried out (Rice 1995, p. 425).

Having set up the null hypothesis, we collect data from n participants for both (1a) and (1b). 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.[1] Given these data, we first compute a vector of pairwise difference scores d from each participant, and then compute the mean difference between the two conditions, $\bar{d}$.

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

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

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

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

Then, we compute the standard error:

$$SE = \frac{s}{\sqrt{n}} \tag{3}$$

The standard error tells us 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 $SE = s/\sqrt{n}$; 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

1 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, see Winter (2019).

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 $\delta = 0$.

$$t \cdot SE = \bar{d} - \delta \tag{4}$$

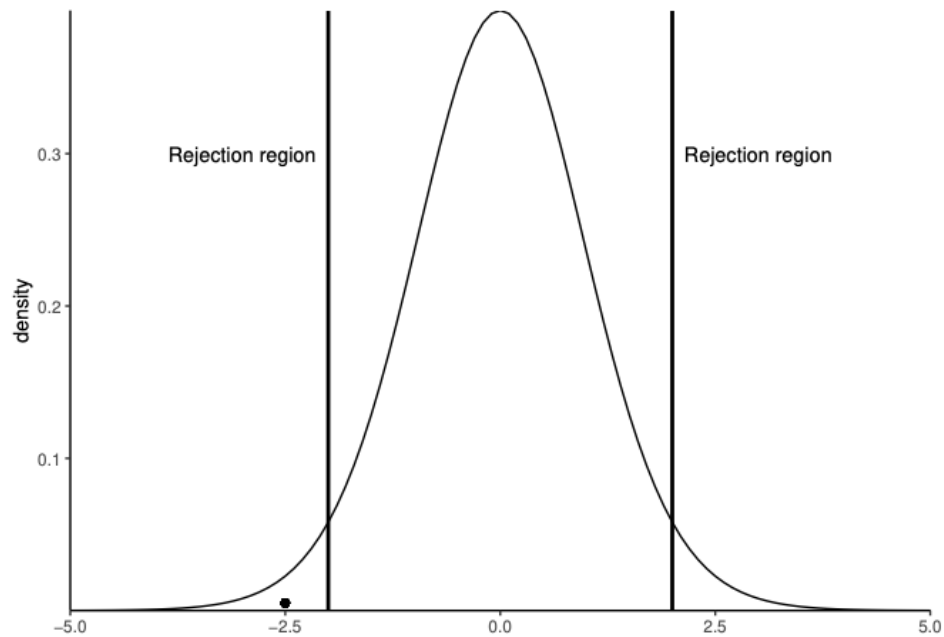

Fig. 1: An illustration of the two-sided t-test. If the observed t-value (the black dot) falls in either of the rejection regions in the tails, then the null hypothesis is rejected.

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 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 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.

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 last conclusion is technically invalid: the statistical test only rejects the null hypothesis, which allows all possible outcomes that are consistent with $\delta \neq 0$. In other words, a $\bar{d}$ with a positive sign (i.e., a slowdown) would be just as consistent with this rejection of the null, but such a positive $\bar{d}$ would be inconsistent with our research hypothesis. Rejecting a null hypothesis only gives us evidence against the null, which is not the same thing as evidence in favor of the specific alternative we are interested in. Nevertheless, the convention is that when we are able to reject the null hypothesis, we say that we have a "statistically significant" result, and that we have found evidence in favor of our specific research hypothesis.

A final note here is that 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: $\bar{d} \pm 2 \times SE$. The confidence interval has a rather convoluted meaning that is prone to misinterpretation (Hoekstra et al. 2014). In practice, 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 Gelman and Greenland (2019).

1.2 Some problems with significance testing

Thus, 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 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. Either we will obtain null results repeatedly if we run the study multiple times, or we will occasionally get significant or even highly significant effects that are gross 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, Wagers, and Lau 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 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 (Nieuwland et al. 2017; Vasishth, Mertzen, et al. 2018; Jäger et al. 2020). 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 1988; Cohen 1962) have long pointed out the importance of ensuring high statistical power for making discovery claims, but these recommendations have largely been ignored in linguistics, psychology, and psycholinguistics (some recent exceptions are Stack, James, and Watson 2018; Brehm and Goldrick 2017; Zormpa, Meyer, and Brehm 2019).[2]

Figure 2 shows power estimates (based on the meta-analysis in Jäger, Engelmann, and Vasishth 2017) for reading studies on agreement attraction and closely

2 In response to the replication crisis that (partly) resulted from underpowered studies (Open Science Collaboration 2015), several remedies have been suggested, such as reducing 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.

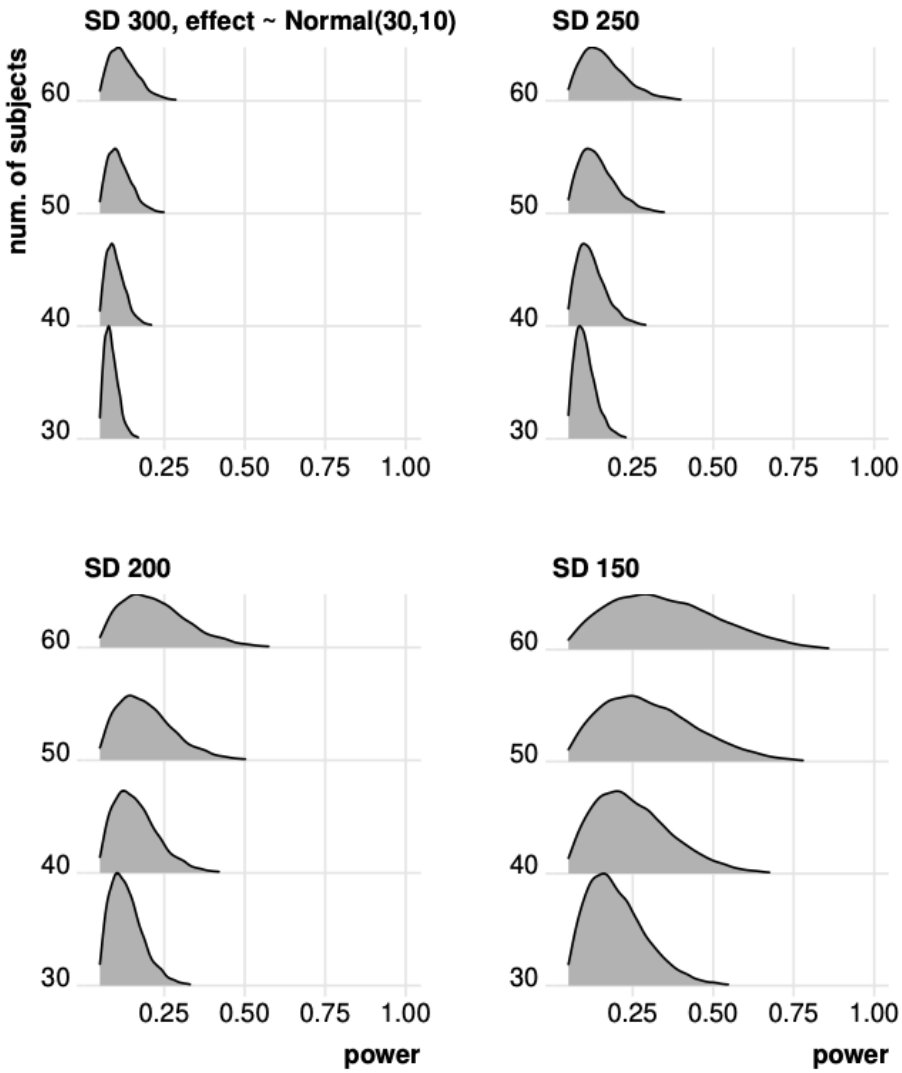

Fig. 2: Power estimates for different numbers of participants (30, 40, 50, 60), assuming a standard deviation (of the residual) of 150, 200, 250, 300 (a typical range in reading studies), and an effect size represented as a Normal distribution with mean 30 ms and standard deviation 10 (i.e., the effect ranges from approximately 10-50 ms with probability 0.95). For a justification of these estimates for the sample size, standard deviation, and effect sizes, see Jäger, Engelmann, and Vasishth (2017).

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 eyetracking), 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 methodology will *never* yield meaningful results because, as discussed above, every possible 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 mis-estimations of the true effect (Gelman and Carlin 2014).

The preceding sentence might seem like an exaggeration to the reader; one might object that as long as one does not filter results by statistical significance, the NHST paradigm is fine, regardless of what the power is. A quick simulation confirms that when power is low, all possible outcomes will be misleading if significance is used as a decision criterion. 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 approximate power 10%, and with 443 subjects, power will be approximately 80%.[3] This figure shows that under low power, there will be many null results, and any significant results will be based on mis-estimation of the true effect. Under high power, we get a high proportion of significant results, and 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.

3 Statistical power is a continuum ranging from whatever the 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.

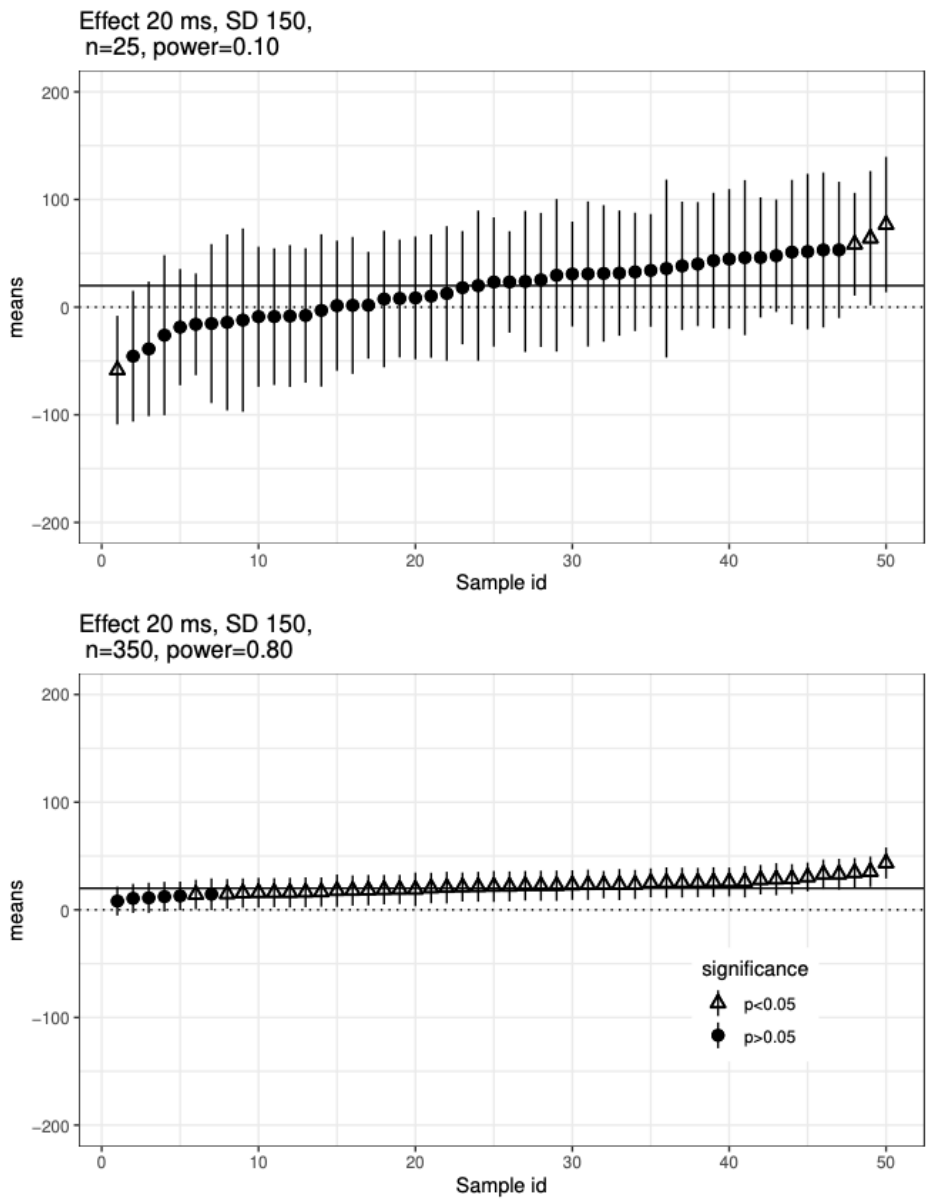

Fig. 3: A simulation showing the effect of low vs. high power on estimates of an effect, under repeated sampling (50 samples). Here, we assume that the data are generated from a normal distribution with mean 20 ms and standard deviation 150. The true mean is shown in each plot as a solid horizontal line. 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 result 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.

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 *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. Statisticians have repeatedly pointed out that the focus in data analysis should be on estimation rather than (or only on) establishing statistical significance or the like (Thompson 2002; Cumming and Fidler 2009; 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 one way that 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, Jäger, Engelmann, and Vasishth 2017). The approaches presented here 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, Maechler, 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, Kliegl, et al. 2015).[4]

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

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

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.[5]

The reader may be tempted to argue that had we done a one-sided t-test, the significance criterion would not be an absolute t-value of 2, but rather 1.65. The argument for a one-sided t-test (rather than a two-sided t-test) in the above case would be that the difference in means is predicted to be negative. Under a one-sided test, we would get seven out of 10 studies coming out significant; this is better than four out of 10 significant effects under the two-sided test. Apart from the fact that there are good reasons to almost never conduct one-sided tests (Pocock 2013), this kind of counter-argument misses the more general point here: when power is low, Type M error (mis-estimation) will be a persistent problem. There are also equally valid and very influential arguments (Benjamin et al. 2018) for reducing Type I error to 0.005, which would lead to an even more unpleasant conclusion than the one we obtained with Type I error 0.05: the absolute critical t-value would now be 2.8, making only *two* out the 10 studies significant.[6]

4 The published results from these studies reported mostly significant effects, but this was a consequence of data trimming, which we did not do here. Instead of trimming extreme values, we log-transformed the reading times. This downweights the extreme values without deleting possibly meaningful data.

5 There are two types of replication attempts. Direct replication attempts are experiments that aim to re-run a published or existing experiment exactly as it was originally conducted. This contrasts with conceptual replications, which are usually experiments that extend the original design in some way to obtain an effect that lead to conclusions consistent with those in the original study. In the present case, it is not technically correct to think of these as direct replications, because the experiments involve different languages, slightly different designs, and different labs. Nevertheless, they are similar enough to each other that they can be treated at least as conceptual replications.

6 There are independent arguments for reducing Type I error in a study: as Malsburg and Angele (2017) and others point out, many experiments involve dozens or hundreds

In summary, under the conventional Type I error 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).

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 vs. 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 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 the true estimate could be much smaller. Indeed, a larger-sample replication attempt of study 1 (181 participants as opposed to 40 participants in the original study) found a much smaller estimate of the effect: -22 $[-46, 3]$ ms (Jäger et al. 2020). Notice the difference between Approach 1 and 2: in t-value based reasoning, we only focused on how many effects were significant; there was no discussion about 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 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 consisent 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.

---

of statistical tests, only a small subset of which end up being reported in the published paper. This multiple comparisons problem inflates Type I error, requiring an adjustment such as the Bonferroni or the Šidák correction.

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, Carlin, et al. 2014). 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. Figure 4 shows the meta-analysis confidence interval (red lines) and mean (gray line).

Thus, if data from multiple experiments exist, we can also 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 estimates 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 -34, with 95% confidence interval [-49,-20] ms. This estimate is consistent with the theoretical claim of a speedup.

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 (source: Jäger, Engelmann, and Vasishth 2017), as well as the predicted range of effects from the computational model for agreement attraction mentioned earlier (Jäger et al. 2020; Vasishth 2019). The meta-analysis range is shown as black vertical lines and the model predictions are shown as a probability distribution.

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 confidence interval overlaps almost exactly with the model's predictions. From this, we would conclude that the evidence from published studies on agreement attraction is consistent with model predictions.

Comparing the posterior distributions individual studies to a predicted range of effects is not a new idea (Freedman, Lowe, and Macaskill 1984; Spiegelhalter,

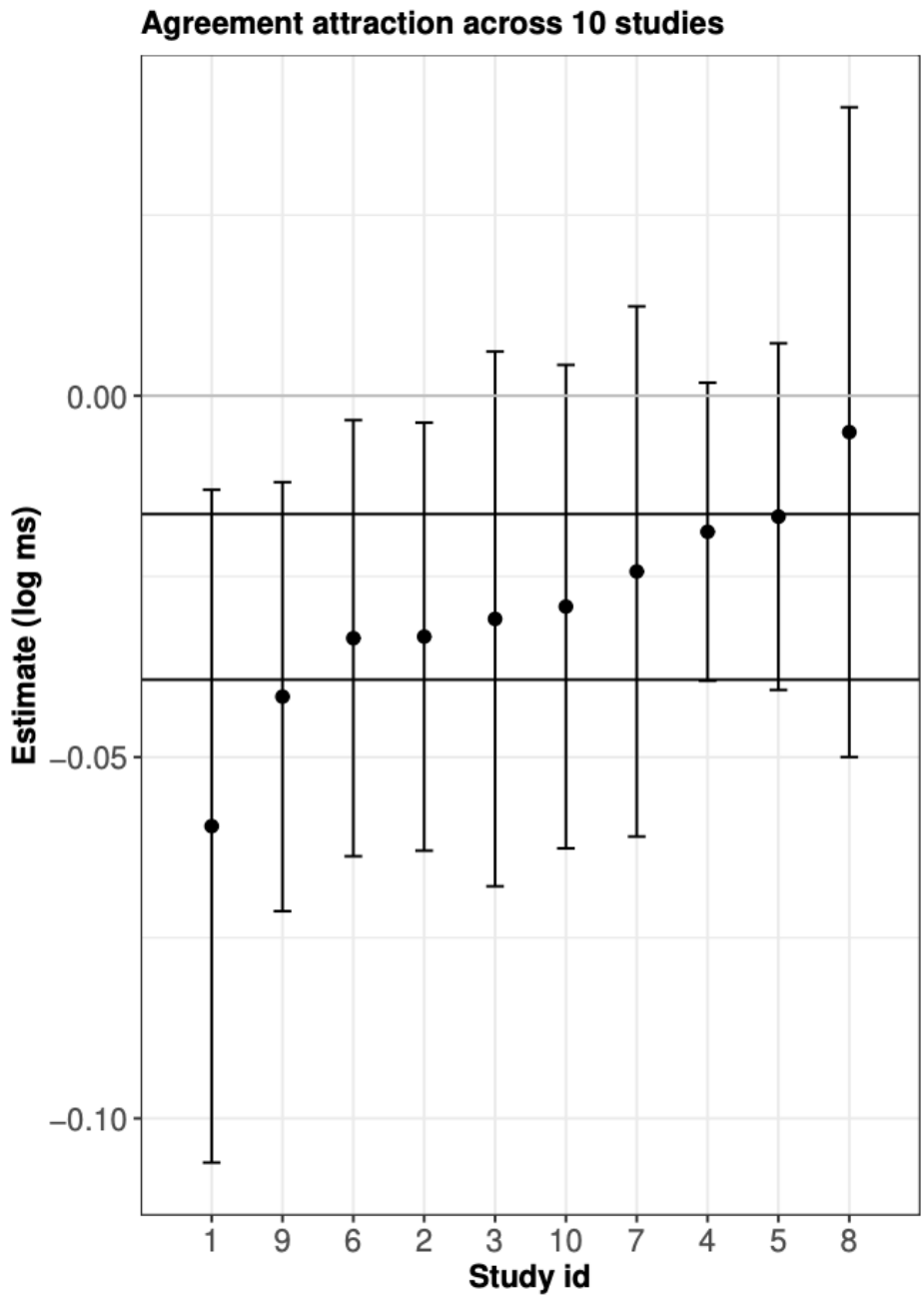

Fig. 4: The mean agreement attraction effect and 95% confidence intervals from the frequentist analyses of 10 reading studies. The horizontal black lines show the 95% confidence interval of the meta-analysis estimate, computed by synthesizing the evidence from the 10 studies.

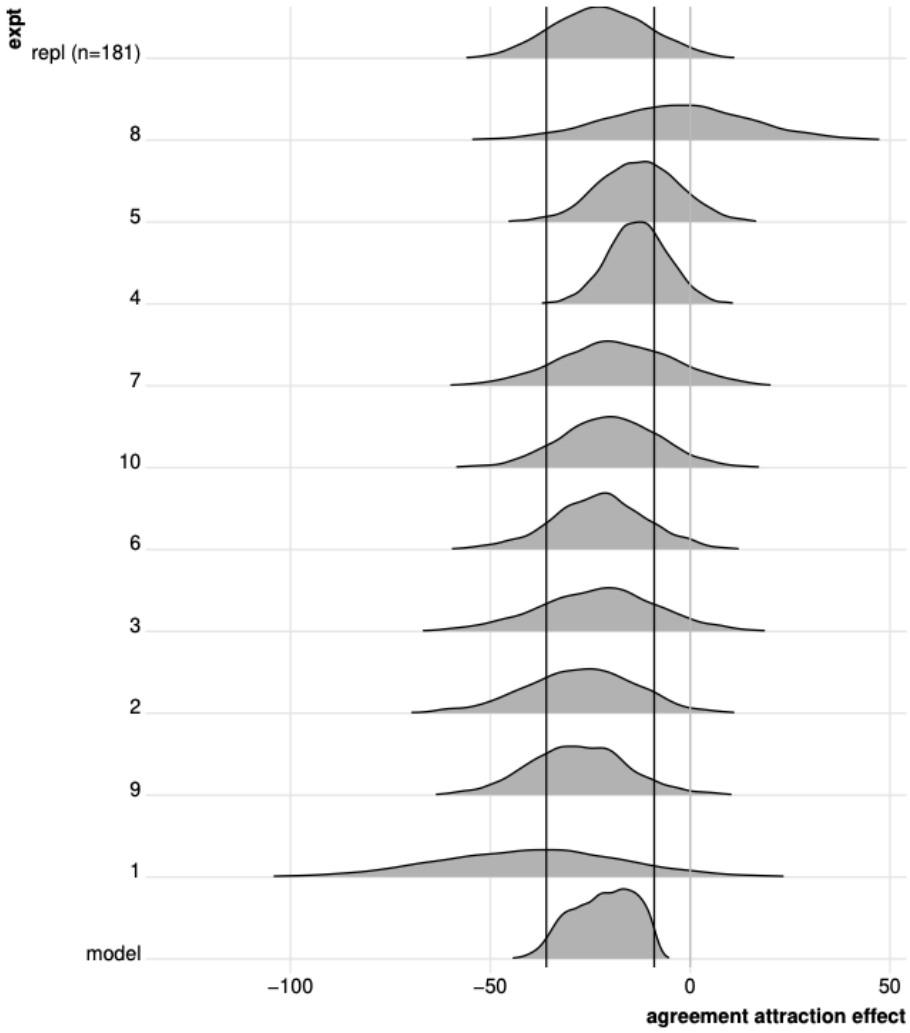

Fig. 5: Ridgeplots showing the distributions of the effect of interest from 10 published reading experiments (eyetracking and self-paced reading) on agreement attraction; the studies are ordered by the mean of the posterior. Also shown is the model's probability distribution of the predicted effect, computed using a large-sample (n=181) data-set investigating agreement attraction (Jäger et al. 2020; Engelmann, Jäger, and Vasishth 2019; Vasishth 2019); for reference, we also show the posterior distribution of the agreement attraction effect in the large-sample study. The black vertical lines mark the 95% confidence interval of a meta-analysis estimate computed using all published reading studies that were available in 2016 that investigated agreement attraction (Jäger, Engelmann, and Vasishth 2017).

Freedman, and Parmar 1994). In recent years, this idea has been re-introduced into psychology by Kruschke (2014) as the region of practical equivalence 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 run our experiment until we have the same width as the predicted range (here, 36+9=45 ms), then there are five possible uncertainty (confidence) intervals that can be observed. The observed interval can be:

A. to the right of the predicted interval.
B. 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 consistency with the quantitative prediction. A and B are inconsistent with the model prediction; and C and D are inconclusive. 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 45 ms). That would be an uninformative, low-precision study.

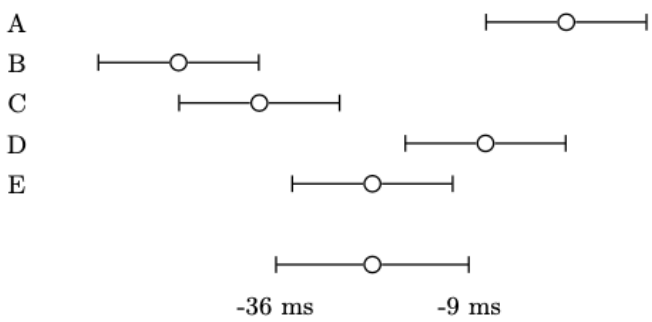

Fig. 6: The five possible outcomes when using the null region or "region of practical equivalence" method for decision-making (Kruschke, 2015). Outcomes A and B are inconsistent with the quantitative predictions of the theory; C and D are inconclusive; and E is consistent with the quantitative theoretical prediction.

In contrast to the region of practical equivalence approach described above, what linguists usually predict is the sign of an effect, but not 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).

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.

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. 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

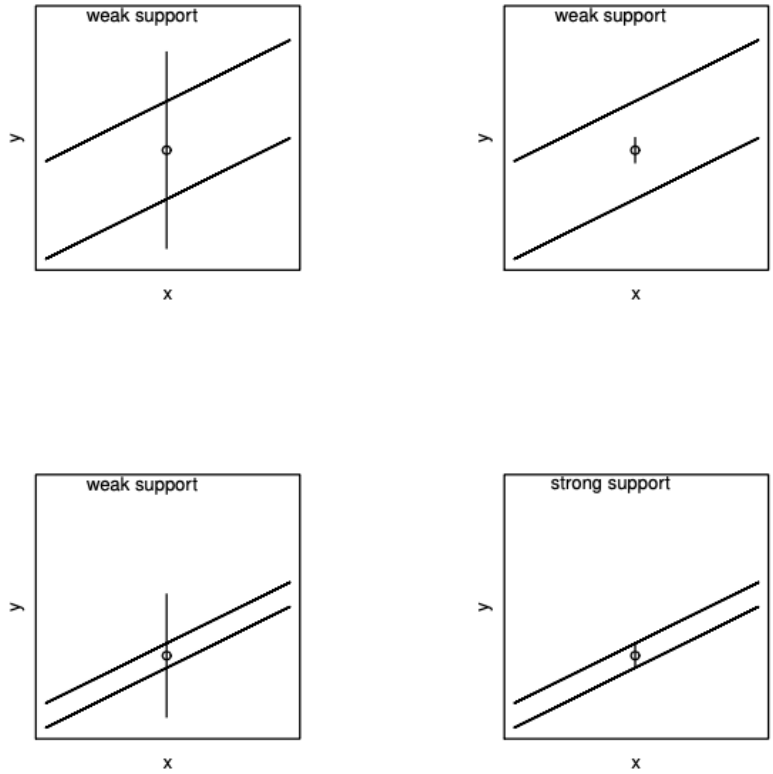

Fig. 7: A schematic summary of the Roberts and Pashler (2000) discussion regarding what constitutes a good fit of a model to data. If a model predicts a positive correlation between two variables x and y, the strong support for the model can only be argued for if both the data and the model predictions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

judgement (O'Hagan et al. 2006). For examples using this method for deciding on sample size, see Vasishth, Mertzen, et al. (2018) and Jäger et al. (2020).

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, we use study 1 as a starting point for planning a new, higher-powered study, assuming that the meta-analysis estimate (along with its 95% confidence interval) reflects the true value of the agreement attractione 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 (here, study 1). 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 (on the log milliseconds scale, these are: mean: -0.03, confidence intervals: [-0.04, -0.02]). 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, what we find is that for the sample size of 40 participants and 48 items in study 1, our estimated power ranges from 0.08 to 0.27. We can now easily compute the power for, e.g., 300 participants: for the mean estimate from the meta-analysis, the estimated power is 0.77, with lower and upper bounds ranging from 0.38 to 0.97. 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.

4 Some potential objections to using uncertainty quantification as an inferential tool

We encounter various objections to the alternative approaches that we have presented which can be used alongside or instead of NHST. 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. Neveretheless, 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, Mertzen, 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.

Despite the positive developments exemplified above, papers may continue to be rejected for not providing supposedly conclusive results.

A major goal of the present paper is to help in 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.

5 Concluding remarks

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

6 Acknowledgements

Thanks go to Lukas Sönning, Reinhold Kliegl, Timo Roettger, Daniela Mertzen, Bruno Nicenboim, and Titus von der Malsburg for useful comments. 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.

A Generating simulated data to compute power

Here we provide code for generating simulated data, and for computing power.

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, Hohenstein, et al. 2020), where $+1$ represents condition (1a) and -1 condition (1b) in the agreement attraction sentences. Thus, a negative sign on the β coefficient would be consistent with the theoretical prediction of a speedup.

$$y_{kj} \sim Normal(\alpha + u_{0j} + w_{0k} + (\beta + u_{1j} + w_{1k}) \times cond_{kj}, \sigma) \tag{5}$$

with the following sources of variability:

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

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

  ## 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$cond[i],
                    sigma_e)}
  fakedat$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, 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.

```
## maximal model:
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, (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,
                        sigma_w1=parest$sigma_w1,
                        nsubj=48,
                        nitem=40){
tvals<-c()
for(i in 1:nsim){
fakedat<-gen_fake_norm(nitem=nitem,
                       nsubj=nsubj,
               alpha=alpha, ## Dillon intercept
               beta=beta,
               sigma_u0=sigma_u0,
               sigma_u1=sigma_u1,
               sigma_w0=sigma_w0,
               sigma_w1=sigma_w1,
               sigma_e==sigma_e)
m<-lmer(rt~cond+(1+cond||subj)+(1+cond||item),
        fakedat,
        control=lmerControl(calc.derivs=FALSE))
tvals[i]<-summary(m)$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)

## [1] 0.1
```

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

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

## [1] 0.12
```

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

## [1] 0.22
```

The code shown above can easily be extended for more complex models and for different likelihoods, such as the LogNormal. For examples, see Vasishth, Mertzen, et al. (2018) and Jäger et al. (2020).

References

Barr, Dale J, Roger Levy, Christoph Scheepers, and 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., Reinhold Kliegl, Shravan Vasishth, and Harald Baayen, 2015. "Parsimonious mixed models." Unpublished manuscript.

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

Benjamin, Daniel J, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al., 2018. Redefine statistical significance. *Nature Human Behaviour* 2.1, 6.

Brehm, Laurel E and 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.

Cohen, Jacob, 1962. The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology* 65.3, 145.

— 1988. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum.

Cumming, Geoff and Fiona Fidler, 2009. Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology* 217.1, 15–26.

Draper, Norma R. and Harry Smith, 1998. *Applied Regression Analysis*. New York: Wiley.

Engelmann, Felix, Lena A. Jäger, and Shravan Vasishth, 2019. The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*. In Press. URL: https://osf.io/b56qv/.

Freedman, Laurence S., D. Lowe, and P. 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 and 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, Aki Vehtari, and Donald B. Rubin, 2014. *Bayesian Data Analysis*. Third. Boca Raton, FL: Chapman and Hall/CRC.

Gelman, Andrew and Sander Greenland, 2019. Are confidence intervals better termed "uncertainty intervals"? *British Medical Journal* 366, l5381.

Gelman, Andrew and Jennifer Hill, 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Greenland, Sander, 2017. Invited commentary: the need for cognitive science in methodology. *American journal of epidemiology* 186.6, 639–645.

Hammerly, Christopher, Adrian Staub, and Brian Dillon, 2019. The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive psychology* 110, 70–104.

Hoekstra, Rink, Richard D Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers, 2014. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 1–8.

Jäger, Lena A., Felix Engelmann, and Shravan Vasishth, 2017. Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language* 94, 316–339.

Jäger, Lena A., Daniela Mertzen, Julie A. Van Dyke, and Shravan Vasishth, 2020. Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language* 111. In Press. URL: https://psyarxiv.com/7c4gu/.

Kruschke, John, 2014. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Loken, Eric and Andrew Gelman, 2017. Measurement error and the replication crisis. *Science* 355.6325, 584–585.

Malsburg, Titus von der and Bernhard Angele, 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language* 94, 119–133.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert, and Jennifer L Tackett, 2019. Abandon statistical significance. *The American Statistician* 73.sup1, 235–245.

Nieuwland, Mante, Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, Sarah Von Grebmer Zu Wolfsthurn, Federica Bartolozzi, Vita Kogan, Aine Ito, et al., 2017. Limits on prediction in language comprehension: A multi-lab failure to replicate evidence for probabilistic pre-activation of phonology. *BioRxiv*, 111807.

O'Hagan, Anthony, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow, 2006. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science* 349.6251, aac4716.

Phillips, Colin, Matthew W. Wagers, and Ellen F. Lau, 2011. "Grammatical illusions and selective fallibility in real-time language comprehension." In: vol. 37. Emerald Bingley, UK, pp. 147–180.

Pocock, Stuart J, 2013. *Clinical trials: A practical approach*. John Wiley & Sons.

Rice, John A., 1995. *Mathematical statistics and data analysis*. Duxbury press Belmont, CA.

Roberts, Seth and Harold Pashler, 2000. How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107.2, 358–367.

Schad, Daniel J., Sven Hohenstein, Shravan Vasishth, and Reinhold Kliegl, 2020. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language* 110. In Press. URL: https://osf.io/7ukf6/.

Schad, Daniel J. and Shravan Vasishth, 2019. "The posterior probability of a null hypothesis given a statistically significant result." Submitted. URL: https://arxiv.org/abs/1901.06889.

Spiegelhalter, David J, Laurence S. Freedman, and Mahesh KB 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, and Duane G Watson, 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition* 46.6, 864–877.

Thompson, Bruce, 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher* 31.3, 25–32.

Vasishth, Shravan, 2019. "Using Approximate Bayesian Computation for estimating parameters in the cue-based retrieval model of sentence processing." Submitted.

Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger, and Andrew Gelman, 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103, 151–175. URL: https://osf.io/eyphj/.

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

Wagers, Matthew W., Ellen F. Lau, and Colin Phillips, 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61.2, 206–237.

Wasserstein, Ronald L. and 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*. Routledge.

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