Model selection and phonological argumentation

James Kirby and Morgan Sonderegger

May 1, 2017

Abstract

Statistical and empirical methods are in widespread use in present-day phonological research. Researchers are often interested in the problem of MODEL SELECTION, or determining whether or not a particular term in a model is *statistically* significant, in order to make a judgement about whether or not that term is *theoretically* significant. If a term is not significant, it is often tempting to conclude that it is not relevant. However, such inferences require an assessment of statistical POWER, a dimension independent from significance. Assessing power is more difficult than assessing significance because it depends on factors including the true (or expected) effect size, sample size, and degree of noise. In this paper, we provide a non-technical introduction to the issue of power, illustrated with simulations based on experimental investigations of incomplete neutralization, to illustrate how not all null results are equally informative. In particular, depending on the statistical power, a non-significant result can either be uninformative, or reasonably interpreted as providing evidence consistent with a small or zero effect.

Keywords: power, model selection, significance, null result, effect size, incomplete neutralization

1 Introduction

Linguistic analysis frequently involves a proces of MODEL SELECTION: making a choice between several theoretical models, on the basis of empirical data. As quantitative methods have become ever more prominent elements of the modern phonologist's toolbox, the theory and practice of model selection have become increasingly important—both in experimental work, where statistical methods have always been essential, as well as in theoretical work, where experimental results are more and more frequently being used as evidence for supporting or advancing a particular theoretical position. In daily practice, phonologists are frequently interested in

To appear in Brentari, Diane and Jackson Lee (eds.), *Shaping Phonology.* Chicago: University of Chicago Press. Please do not cite this version without permission.

assessing the importance (either in a technical or non-technical sense) of a model parameter—does the underlying [voice] status of a German coda obstruent affect its surface realization, say, or does word frequency play a role in how likely trisyllabic shortening is to occur in English? Statistical results are then used to decide between different theories of the phenomenon: for example, is it necessary for phonological representations of English words to include information about word frequency, or can the effects of frequency be accounted for in other ways? In practice, this often ends up meaning: are there some factors that are found to be *significant* in a statistical model? As such, the issues of choosing between alternative statistical models and interpreting model parameters are becoming increasingly important for research on sound structure, and for linguistics as a whole.

Any model selection problem has two important aspects.[1] Given any pair of (nested) models in a set of candidate models, we want to assess (1) whether or not a particular term is justified ("is there an effect?"); and (2) the value of the term (its EFFECT SIZE: "how big is the effect?")? Both issues play into linguistic argumentation; here we focus on the first of them, which currently plays a larger role in the practical application of statistical methods in linguistics. (We return briefly to the second in the conclusion.) We also focus on the case of a single term, in a model which may contain many terms. The issue is then: given the results of an experimental study, do we conclude based on a statistical model that there is an effect of X (which would support Theory A) or not (which would support Theory B)?

This choice is independent of whether there is, in reality, an effect of X or not. This means that there are two types of errors the researcher can make: falsely concluding there is an effect when none exists (TYPE I ERROR), or falsely concluding there is no effect when one in fact exists (TYPE II ERROR). The first is arguably more familiar, and much more common in everyday statistical practice. If a term is found to be important (e.g. assessed via a p-value being below some cutoff, such as 0.05), in the sense of a Type I error being unlikely, then we conclude there is a *significant* effect, which supports Theory A. Although it is well known that a significant p-value does not support Theory B, in many contexts researchers are understandably tempted to interpret the finding that a coefficient is not significantly different from zero, because ultimately the goal is to make a choice between two theories, with potentially important linguistic ramifications.[2] Null results *can* in fact be interpreted, but only by taking POWER (Type II error) into account, in addition to significance. Power is less commonly covered than significance in introductory courses, and is trickier to think about, because it depends on considerations of sample size, effect size, and noise. Intuitively, a significant result is less likely to be found for an experiment with a smaller sample, where the effect is small, and/or where the variance is

high, even where a true effect exists.

The purpose of this paper is to provide a brief and accessible illustration of power analysis for a case study of phonological interest, motivated by two questions:

(Q1) What are we licensed to conclude on the basis of an individual study?

(Q2) What are we licensed to conclude from a body of studies?

2 Background

2.1 Incomplete neutralization

As a case study, we consider the issue of so-called INCOMPLETE NEUTRALIZATION (IN) of word-final voicing in languages like German, Catalan, or Dutch. An example from German is given in (1): in final position, the voicing contrast in stops is neutralized, leading to apparent homophony between *Rat* 'council' and *Rad* 'wheel'.

(1) a. *Rat* /ʁaːt/ > [ʁaːt] 'council', *Räte* [ʁɛːtə] 'councils'
b. *Rad* /ʁaːd/ > [ʁaːt] 'wheel', *Räder* [ʁɛːdɐ] 'wheels'

Word-final neutralization of this type has often been used as a textbook example of an exceptionless phonological rule. Beginning in the early 1980s, however, this picture was blurred by phonetic studies claiming to show a small but significant difference in the phonetic realizations of underlyingly voiced and voiceless obstruents, usually in terms of their effect on the durations of the burst, closure, and/or preceding vowel (e.g. Mitleb 1981; Fourakis & Iverson 1984; Port & O'Dell 1985; Port & Crawford 1989; Jassem & Richter 1989; Piroth & Janker 2004; Warner *et al.* 2004; Warner *et al.* 2006; Roettger *et al.* 2014). Much of the subsequent debate has been about methodological issues (see esp. Winter & Röttger 2011; Kohler 2012), but what has primarily captured the interest of phonologists are the implications for phonological theory. For some, the existence of IN effects would have important theoretical ramifications, entailing either a major modification to the notion of contrast in order to accommodate these small but consistent differences (Braver 2013; van Oostendorp 2008; Yu 2011), or the incorporation of a large set of (possibly speaker-dependent) articulatory features into phonology (Port & Crawford 1989). This stands in opposition to the traditional position that phonological contrasts are amenable to discovery on the basis of native speaker and/or analyst intuitions (Manaster-Ramer 1996), and that the phonetic differences attribute to IN are the result of orthographic confounds, task effects, hyperarticulation, or other factors outside the purview of the phonology. Some have gone so far as to claim that the existence of IN would pose "a threat to phonological theory" (Port & Crawford 1989:257) requiring that the

field "rethink the whole process of collecting and evaluating claims of fact about the phonetics and phonology of the world's languages and dialects" (Manaster-Ramer 1996:480), suggesting that the stakes for getting the model selection right are quite high.

Here, our focus is not on whether or not incomplete neutralization is real, or the theoretical implications in either case. Rather, we are interested in IN as a good example of an instance where the *absence of evidence* has been repeatedly interpreted by researchers as relevant for phonological argumentation. In the IN literature, we find studies which find statistically significant evidence for acoustically incomplete neutralization (Port & O'Dell 1985; Port & Crawford 1989; Roettger *et al.* 2014) alongside those that do not (Fourakis & Iverson 1984; Jassem & Richter 1989). What may we conclude from a single study that fails to find an effect? And, how should we interpret a body of results, some of which find an effect, and some which do not?

2.2 Power

Interpreting a significant result requires consideration of one hypothetical scenario: if there were in reality no effect (the "true effect size" is zero), how likely would the result be? Interpreting the *lack* of a significant effect requires the researcher to consider a different hypothetical scenario: if there were a real effect of a given size, for a dataset like this one, how likely would we be to detect it? This is statistical POWER, which is conceptually independent from significance. Power is the probability of committing a Type II error (falsely concluding that there is no effect when in fact one exists); thus, we can describe an experimental design where there is a low probability of committing a Type II error as having HIGH POWER, and one with a high probability of committing a Type II error has having LOW POWER. Although the concept of power is amply covered elsewhere (for a recent overview from a psycholinguistic standpoint, see Vasishth & Nicenboim 2016a), we give a motivating example here in the context of IN to provide some intuition.

Fourakis & Iverson (1984) examined production of German word-final stops for 6 repetitions of 5 voiced/voiceless pairs by 4 speakers, and found that while the mean value of acoustic parameters for voiced and voiceless stops were in the expected direction (e.g. vowel duration preceding voiced stops was 3.8 msec longer on average), this difference was not significant (assessed via *t*-tests). They concluded that "...the traditional position that German devoicing constitutes phonologically irrecoverable merger is supported."[3] This conclusion, however, depends on how surprising it is to have *not* found a real effect if it exists—the power—which is quite low in this case (see Sec. 4.1). Why is the conclusion not surprising, and what would change this?

Intuitively, not detecting a 3.8 msec difference is less surprising than not detecting a 30 msec difference (as found in e.g. English); not finding a significant result would be more surprising with 20 subjects instead of 4; and finding the effect would be less surprising in natural versus in read speech. These three factors—the true effect size, sample size, and the amount of noise—all affect an experiment's power. (A further factor, the data analysis method used, is discussed further below.)

The dependence of power on these factors is often illustrated in the context of relatively simple examples, such as *t*-tests, and it is not necessarily obvious how power analysis should proceed in the more complex scenarios in which researchers now typically find themselves, such as nested model comparison between generalized linear mixed models with large numbers of interactions (but see Matuschek *et al.* 2015; Vasishth & Nicenboim 2016a for related illustrations in the context of psycholinguistic studies). Our aim here is to provide an illustration of power analysis in a mixed-model setting, for a relatively simple case (one term of interest). In doing so, we illustrate how considerations of power interact with model selection strategies, and the impact this has on the interpretation of both single studies and bodies of studies.

3 Methods

We proceed via a simulation study, using a dataset from a moderately-powered study of IN in German where a significant effect was found. We use the results of this study to simulate datasets where two factors are varied—the sample size, and the true effect size—while holding all others constant; we also vary the criterion used to decide if there *is* IN. This allows us to explore how power would be affected if the same experiment had been run, but with less data; if the true effect size were different; or if the data were analyzed differently.

3.1 Dataset

The case we consider is Experiment 1 of Roettger *et al.* (2014), using the dataset from this study.[4] Roettger et al. recorded 16 native speakers of German producing singular forms of nonwords (e.g., [goːp]) in which the target consonant was in final position in response to auditory primes containing a voiced variant (e.g., [goːbə]). Each speaker produced one repetition of 24 critical items (a pair such as [goːbə]/[goːpə]). The statistical analysis, using a linear mixed-effects model, modeled duration of the vowel preceding the stop as a function of the stop's underlying voicing, as well as a number of control predictors. By-subject and by-item random intercepts and random slopes for voicing were included. The key result for our purposes is that speakers produced longer vowels before underlyingly voiced stops: the difference (correspond-

ing to the voicing fixed-effect coefficient) was statistically significant in a likelihood-ratio test ($\chi^2(1) = 13.76, p < 0.0002$), and estimated to be 8.6 msec (SE = 2.03 msec).

Since we do not know if the incomplete neutralization effect is "real", by definition we also do not know its true size. However, since a number of studies of this effect have now been conducted, we have some basis for guessing what the true size of the effect might be. For German, published estimates have ranged from around 4 msec (Port & Crawford 1989) to over 20 msec (Mitleb 1981). For present purposes, these estimates will suffice to give us a range in which to explore the ramifications of effect size on power in a mixed-model setting.

3.2 Simulations

In our simulations, we varied three factors[5] to understand their effect on power: the sample size, the effect size, and the model selection criteria (Table 1).[6] In terms of sample size, we altered the number of subjects, items, and repetitions, as these are the primary differences between studies of nominally the same phenomenon, both in the IN literature and in experimental work more generally. These parameters were varied in a range of values corresponding to previous work. Sweeping the effect size is important as well, because in a real-world study, we never know *a priori* the size of the effect we are looking for; we can only make an inference about likely effect size based on related work (see discussion in Vasishth & Nicenboim 2016a). The effect size was swept from values corresponding to no effect (β=0) to a moderate IN effect (10 ms). Finally, we considered power under three different model selection criteria (discussed below), to show how different choices about data analysis might also effect a researcher's ability to detect an effect.

[Table 1 about here.]

3.2.1 Simulation procedure

For a given set of parameter values, a single simulation run was performed as follows.[7] A random subset of the original dataset was taken corresponding to n_s subjects and n_i items; this data was concatenated n_r times, making a "resampled" dataset. The fixed-effect coefficients of the original model were used to predict a vowel duration for each data point, excluding the effect of voicing. The random-effect parameters (variance components) of the original model were used to sample new intercept and slope offsets for each subset and item in the resampled dataset. These, together with the desired true effect size (β), were used to adjust predicted vowel durations for each subject and item—including the effect of voicing. Finally, the estimated

residual error was used to add observation-level noise to each prediction. The resulting dataset can be thought of as one possible "smaller" version of the original dataset, accounting for by-speaker and by-item variability, and with an adjusted effect size for voicing. Two statistical models were fitted to this new dataset: the original statistical model (the *superset model*), and this model with the fixed effect of voicing removed (the *subset model*).

Given these two models, three model selection criteria were applied to decide whether the a model with the voicing term was justified by the data:

1. *Likelihood-ratio test* (LR): Assess the significance (p) of the difference in log-likelihood between the two models, using a χ^2-test. Choose the superset model if $p < 0.05$, and the subset model otherwise.
2. *Akaike Information Criterion* (AIC): Assess the tradeoff between the log-likelihood of the observed data under a proposed model ($\mathcal{L}$), and the number of parameters in the model ($\mathcal{Q}$), as $-2\mathcal{L} + 2\mathcal{Q}$. Choose the model with the lower AIC.
3. *Bayesian Information Criterion* (BIC): Assess a similar tradeoff, taking into account as well the number of observations in the dataset N, as $-2\mathcal{L} + \ln(N)\mathcal{Q}$. Choose the model with the lower BIC.

All three methods measure the tradeoff between model complexity and fit to the data in some way. As N increases, BIC tends to favor simpler models than AIC, due to the $\ln(N)$ term which imposes a higher penality for each additional model parameter. In practice, BIC is expected to often be more conservative than AIC or LR: it will have lower Type I error, but also lower power (Type II error). Both LR and AIC are widely used in practice (in linguistic research in particular) as model selection criteria (AIC, for example, is often used in "stepwise" methods); we include BIC as well to help illustrate the effect of model selection criterion on power.

Thus, for each simulation run, we have three decisions on whether the voicing term is justified or not; concluding it is not would be a Type II error (unless $\beta = 0$). By performing n_{sim} runs and computing the fraction in which the subset model is chosen, we obtain an estimate of the power under each model selection strategy, for a given set of parameter settings. We performed simulations with $n_{sim} = 500$. Although our results show the curves resulting from sweeping several parameters, note that a common cutoff for a "high power result", corresponding to the common 0.05 cutoff for a "significant" result, is 0.8 (i.e., at least an 80% chance of detecting the effect).

4 Results

[Figure 1 about here.]

Figure 1 shows the results of these simulations: how power (on the y-axis) varies as a function of the true effect size β (on the x-axis) for several different sample sizes and model selection criteria. Each curve in the figure thus represents a different study design—with different choices for the number of subjects, items, and repetitions—as well as a choice of model selection criterion, and can be used to determine the power of the experiment as a function of the true effect size (which, in general, is not known to the analyst). The general pattern is quite clear: as the sample size and the true size of the effect increase, so too does power. The model selection criterion used also makes a difference; in general, using AIC will lead to greater power, followed by LR, and finally BIC. As noted above, this is expected based on the properties of these criteria: the AIC tends to favor predictive accuracy over model complexity, while the BIC more heavily penalizes model parameters, particular as the number of observations increase. The greater power of AIC leads it to have the highest Type I error (i.e., to wrongly conclude there is an effect when there is not), as can be seen by examining the power curves as the true effect size approaches zero.

Recall that we are interested in two main questions: (Q1) What are we licensed to conclude on the basis of an individual study? and (Q2) What are we licensed to conclude from a body of studies? To gain some intuition for the patterns in this figure with respect to these two questions, we will consider three regimes in detail (low, mid, and high-power), roughly corresponding to three studies from the existing IN literature. In each case, we pose the following question: suppose we re-ran the Roettger et al. study with a different sample size; what should we conclude in case of different outcomes (Q1)? We then we consider some examples of how to interpret results from several studies together (Q2), assuming they are from different regimes. We then offer one possible interpretation of the German incomplete neutralization literature.

[Figure 2 about here.]

4.1 Low-power regime

To illustrate a low-power regime, we select the power curves from simulations with 6 subjects, 8 items, and 2 repetitions per item. In this regime, power is always below 80%, regardless of model selection criteria or effect size (assuming that the effect is $\leq$ 10 ms), and is below 50% for most values of β. The logic of interpretation is different in different cases (keeping in mind that, in an experimental setting, we don't know what the true size of the effect is; it could be zero). If we find a significant result (e.g., $p < 0.05$), we may conclude that observing an effect of this magnitude, or larger, is unlikely to have occurred if the true contribution of β is in fact zero. If we

don't find a significant result, on the other hand, we should not be surprised; but we cannot interpret this lack of effect as evidence in favor of anything: a non-significant result would have been likely to occur whether there is in reality a true effect of $\leq$10 msec (low power) or not (high p-value). In a low-powered study, then, a null result is not informative.

In terms of the IN literature, a possible analog is the "elicitation condition" study of Fourakis & Iverson (1984), with 4 subjects and 6 repetitions; t-tests are reported for subsets corresponding in our terms to 1–2 items, and none of these tests are significant.[8] Approximate power calculations for these t-tests can be carried out using the information in their Table 2; even assuming a 10 msec true effect size (much larger than that reported), power is below 0.35 for all tests. Given what we might reasonably assume about the true size of the effect, the null result does not provide evidence to "falsify the claim that final obstruent devoicing is not neutralizing in German", neither can it be claimed that "the traditional position that German devoicing constitutes phonologically irrecoverable merger is fully supported" (Fourakis & Iverson 1984, p. 149). When power is this low, a null result does not by itself contribute one way or the other to our understanding of the phenomenon under study.

4.2 Mid-power regime

The mid-power regime is illustrated by power curves for simulations with 8 subjects and 3 repetitions each of 12 items (Figure 2B). In this regime, power is above the 80% mark for a sufficiently large effect size using the less conservative model selection criteria (LR or AIC), but only for the upper range of effect size. Again, a significant result can be interpreted as meaningful (i.e., unlikely to happen by chance); but the interpretation of a null result is less straightforward. If we have reason to believe the true β is, say, 10 msec or higher, we may reasonably expect to have detected it. Therefore, *not* finding a significant effect can be interpreted as evidence that *if* an incomplete neutralization effect exists, it is going to be smaller than 10 ms. Note that this is *not* the same as saying we have evidence that there is no effect; rather, this is a statement about our ability to detect an effect *of a given size*.

In this regime, power is particularly sensitive to both the true effect size as well as the data analysis method. One practical consequence of this is that replications of the same experiment (or subsets of data from the same experiment) could well return a mix of significant and non-significant effects *even if there is a true effect*. Similarly, different results may obtain depending on the particulars of the data analysis method. A possible example from the literature is the study of Piroth & Janker (2004), who analyzed data from 3 repetitions of 9 pairs uttered by 6 German speakers from different dialect areas. They found that the 2 Southern German speakers

in their sample preserved acoustic differences in coda duration between underlyingly voiced/voiceless pairs, but that speakers from the other dialect regions did not. Given the power of the study, however, we should not necessarily be surprised that they failed to detect a (small) effect that may in fact be present. If the IN effect for the Southern German speakers is in reality larger than for speakers from other dialect areas, it will be easier to detect, all else being equal. So while we are licensed to conclude something about the Southern German speakers in this study, we have not really learned anything about other speakers, or about the ensemble of speakers as a whole.

4.3 High-power regime

Finally, we consider a design with 16 speakers, 24 items, and 6 repetitions (i.e., the design of Roettger *et al.* 2014, but with 6 repetitions instead of just one). As seen in Figure 2C, power is above the 80% mark for the majority of the range of possible effect sizes, at least for the AIC and LR model selection criteria. Once again, a significant result can again be interpreted as evidence of an incomplete neutralization effect—if the true effect size were zero, such a result would be unlikely to occur (modulo of course the possibility of Type I error). Unlike the low and medium-power regimes, however, here a null result is also meaningful: if there *were* a true effect in this range of effect sizes, we would be surprised to not detect it, while if the true effect size were zero, we would not be surprised if we failed to find it. Therefore, in a high-powered design, we *are* licensed to interpret a null result as evidence for complete neutralization—at least in the sense of, if there is an effect, it has to be quite small.

In the incomplete neutralization literature, the studies of Warner *et al.* (2004); Warner *et al.* (2006) are illustrative in this regard. Warner *et al.* (2004) found that the Dutch word-final voicing contrast was incompletely neutralized, with vowels preceding voiced stops 3.5 msec longer than those preceding voiceless stops. However, in their follow-on study (Warner *et al.* 2006), in which they carefully controlled for possible orthographic effects, they failed to find a significant effect. Because the 2006 study was sufficiently high-powered (as calculated by the authors, using the effect size from the 2004 study), the authors are licensed to interpret their findings as indicating that "the current manipulation does not produce an incomplete neutralization effect, *at least not one comparable in size to the 3.5 msec differences found in the previous work*" (2006: p. 290, emphasis added). Because of the high power of the design, they are able to interpret the null result in a theoretically meaningful way.

5 Discussion

As the preceding examples illustrate, the conclusions that can be drawn on the basis of a null result, with respect to deciding between different theoretical stances, are heavily dependent on the power regime. This, in turn, is dependent on factors including the true effect size, the sample size, and the particulars of the data analysis method. This contrasts sharply with what may be inferred from a significant result, which (at least to a reasonably good first approximation) does not depend on sample size or model selection criteria.[9] This does not mean that only significant results are meaningful: if a high-powered study returns a null result, as in the case of Warner *et al.* (2006) (Sec. 4.3), it can be interpreted as evidence consistent with the null hypothesis, in the sense that it supports the selection of a simpler (less parameterized) statistical model.

Given what we have reviewed, what can we conclude from a collection of studies of (more or less) the same phenomenon, in which some studies find significant effects and others do not (or equivalently, select a model containing a term T versus reject a model containing T on the basis of some model selection criterion)? In such a scenario, do we have evidence for theory A, theory B, or truly conflicting evidence?

As the above discussion suggests, the answer depends on the power of the studies involved. If all of the studies concerned have high power, then those which find significant results provide us with evidence favoring/consistent with Theory A, while those that find null results can be interpreted as favoring/consistent with Theory B. In this scenario, the results are truly conflicting, because we have evidence that supports different, presumably incompatible theoretical positions. If, on the other hand, the high-powered studies find significant results, but the low-powered studies find null results, we only have evidence that supports Theory A; the null results are not evidence for or against anything.

What consequences does this have for our interpretation of the German IN literature? The existence of a high-powered study on German IN similar to Warner *et al.* (2006)—which found a null effect, or a significant effect in the wrong direction—would indeed conflict with earlier findings. At least in the German case, however, the mostly highly powered study of which we are aware (Roettger *et al.* 2014) finds a small but significant IN effect in the expected direction, in spite of numerous experimental controls. Previous low- to medium-powered studies have either similarly found small but significant effects in this direction, or have returned null results. Thus, because null results are only informative in a high-power regime, it is not the case that there is conflicting evidence; rather, a reasonable interpretation of the existing literature is that there is, in fact, a small but real effect of IN in German, and the inconsistency between studies is due to lack of statistical power.

5.1 Further issues

Our discussion has focused on interpreting the results of an experiment (or a set of experiments) with respect to whether or not an effect is zero, corresponding to a choice between two linguistic theories. There are of course further considerations that enter into the interpretation of experimental results, some of which might suggest that our concerns are unwarranted (at least for the IN case). Here, we discuss two of these: PRIOR BELIEF in one linguistic theory versus another, and (estimated) EFFECT SIZE.

First: isn't a null result informative if it agrees with (or fails to contradict) a more plausible theory? A researcher may have significant evidence (e.g. decades of agreement among experimentalists, or an assumption about phonological representation working well in daily practice) that leads them to think Theory A (e.g. complete neutralization) is more plausible than Theory B (e.g. incomplete neutralization). When an experiment then searches for evidence in favor of Theory B, and finds a null result, it is tempting to conclude that this supports the researcher's strong prior belief in Theory A. This is an implicitly Bayesian view of the scientific world: a researcher (or the field as a whole) has degrees of belief in different hypotheses, which are updated based on new information. This view makes intuitive sense, but is not consistent with the null hypothesis significance testing (NHST) statistical framework used almost exclusively in experimental studies in linguistics (including all studies of IN cited here). The NHST framework does not take prior beliefs into account, and a null result from an NHST statistical method cannot be interpreted as supporting the null, regardless of its plausibility. There are ways to explicitly combine prior beliefs with new evidence in performing statistical analysis, using Bayesian data analysis methods (e.g. Jaynes 2003; Gelman *et al.* 2014), the main alternative to NHST. Using these methods, a null result from a low-powered study tends to offer evidence both for and against a researcher's prior belief, since the observed outcome would be fairly likely under a number of different priors (Vasishth & Nicenboim 2016b). However, even within this framework, *a null result from a low-powered study is still uninformative*—as demonstrated in our simulation study.

Second, even if an effect can be demonstrated, it may not be of sufficient magnitude to warrant attention. For example, IN effects have often been suggested to arise from orthographic, hypercorrection and/or task effects (e.g. Fourakis & Iverson 1984; Warner *et al.* 2006; Kharmalov 2014), and in any event to be too small to be of any communicative relevance (Kohler 2012). This gets at a broader issue in interpreting experimental results: the estimated *size* of an effect is just as important as whether it is significantly different from zero or not, and these two things are independent. It is possible to have a tiny but significant effect (as in the Roettger *et al.*

2014 study) in a high-powered study, or a large but non-significant effect (in a low-powered study)—where 'small' and 'large' are always assessed relative to a particular domain (e.g. vowel duration differences between voiced and voiceless obstruents cross-linguistically). Thus, it is crucial to consider *both* the estimated size of effects and their significances when interpreting experimental data, as emphasized in modern approaches to data analysis across fields (e.g. Baayen 2008; Gelman & Hill 2007). An effect's estimated size (as for whether it is zero or not) is subject to error, and it is important to take into account in interpreting either a single experimental result or a body of studies. A thorough exposition is beyond the scope of this paper (but see Vasishth & Nicenboim 2016a for a discussion in the context of linguistic data), but power turns out to again be crucial: for low-powered studies, estimated effect sizes are likely to have the wrong magnitude or sign, *even for a significant result*; for high-powered studies, estimated effect sizes are likely to be reliable, *even for null results*. Thus, effect sizes from a low-powered study should be trusted less than effect sizes from appropriately-powered studies, which is important when interpreting experimental results in order to decide between competing theoretical models. In the IN literature, for example, a surprisingly large IN effect found in a low-powered study (such as the 20 msec reported in Mitleb 1981) should be given little weight, while a near-zero (and not significant) effect found in a high-powered study (such as Warner *et al.* 2006, for Dutch) should be taken seriously.

6 Conclusion

Model selection is a powerful tool for linguistics and provides a rigorous basis for scientific understanding, but it must be approached with cautious respect. In the preceding, we have tried to demonstrate how a failure to take statistical power into account can potentially lead to unlicensed inference with respect to the theoretical issue(s) at stake. Similarly, we have tried to illustrate how choices about data analysis, such as model selection criteria, can impact statistical inference and subsequent reasoning based on it.

References

Anderson, David R. 2008. *Model based inference in the life sciences: A primer on evidence*. New York: Springer Verlag.

Baayen, R.H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

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* 68.255–278.

Braver, Aaron, 2013. *Degrees of incompleteness in neutralization: Paradigm uniformity in a phonetics with weighted constraints*. Rutgers University-Graduate School-New Brunswick dissertation.

Fourakis, Marios, & Gregory K. Iverson. 1984. On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41.128–143.

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

Gelman, Andrew, John B Carlin, Hal S Stern, & Donald B Rubin. 2014. *Bayesian data analysis*. Taylor & Francis, 3rd edition.

Haller, Heiko, & Stefan Krauss. 2002. Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research* 7.1–20.

Jassem, Wiktor, & Lutoslawa Richter. 1989. Neutralization of voicing in Polish obstruents. *Journal of Phonetics* 17.317–325.

Jaynes, Edwin T. 2003. *Probability theory: the logic of science*. Cambridge University Press.

Kharmalov, Viktor. 2014. Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: Phonological, lexical, and methodological influences. *Journal of Phonetics* 43.47–56.

Kohler, Klaus J., 2012. The phonetics-phonology issue in the analysis of word-final obstruent voicing. http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2012_3/neutralization.pdf.

Manaster-Ramer, Alexis. 1996. A letter from an incompletely neutral phonologist. *Journal of Phonetics* 24.477–489.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, & Douglas Bates. 2015. Balancing Type I error and power in linear mixed models. *ArXiv e-prints* arXiv:1511.01864.

Mitleb, F. 1981. Temporal correlates of "voicing" and its neutralization in German. *Research in Phonetics* 2.173–191.

Piroth, Hans Georg, & Peter M. Janker. 2004. Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics* 32.81–109.

Port, Robert, & Penny Crawford. 1989. Pragmatic effects on neutralization rules. *Journal of Phonetics* 16.257–282.

Port, Robert F., & Michael O'Dell. 1985. Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13.455–471.

Roettger, Timo B., Bodo Winter, Sven Grawunder, James Kirby, & Martine Grice. 2014. Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics* 43.

van Oostendorp, Marc. 2008. Incomplete devoicing in formal phonology. *Lingua* 118.1362–1374.

Vasishth, Shravan, & Bruno Nicenboim. 2016a. Statistical methods for linguistic research: Foundational ideas - part I. *ArXiv e-prints* arXiv:1601.01126. In press at *Language and Linguistics Compass*.

——, & ——. 2016b. Statistical methods for linguistic research: Foundational ideas - part II. *ArXiv e-prints* arXiv:1602.00245. In press at *Language and Linguistics Compass*.

Warner, Natasha, Erin Good, Allard Jongman, & Joan Sereno. 2006. Orthographic vs. morphological incomplete neutralization effects. *Journal of Phonetics* 34.285–293.

——, Allard Jongman, Joan Sereno, & Rachèl Kemps. 2004. Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics* 32.251–276.

Winter, Bodo. 2015. The other N: the role of repetitions and items in the design of phonetic experiments. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: The University of Glasgow.

——, & Timo Röttger. 2011. The nature of incomplete neutralization in German: Implications for laboratory phonology. *Grazer Linguistiche Studien* 76.55–74.

Yu, A.C.L. 2011. Contrast reduction. In *The handbook of phonological theory*, ed. by J. Goldsmith, J. Riggle, & A.C.L. Yu, 291–318. Oxford: Blackwell Publishing, 2nd edition.

Notes

[1]We set aside here the equally important aspect of deciding on an appropriate set of candidate models; for more on this issue see Anderson (2008).

[2]This issue is widespread both in linguistics and beyond; see e.g. Vasishth & Nicenboim (2016a); Haller & Krauss (2002).

[3]We are considering Fourakis & Iverson's "elicitation task". For a second experiment, they do find a significant IN effect, but attribute it to the less communicatively-natural setting, and conclude that neutralization is complete in natural settings.

[4]We thank Timo Roettger and Bodo Winter for sharing this dataset with us.

[5]We also explored varying the "noise"—the residual variance, and amount of variability among subjects and items—which also affects power, but hold it constant in the simulations we report here in the interest of expositional clarity.

[6]For discussion of how these factors effect Type I error, see Barr *et al.* 2013; Matuschek *et al.* 2015; Winter 2015.

[7]Our methodology is a simplified version of the simulation-based power calculation method for mixed models described in Chapter 20 of Gelman & Hill (2007). Our simulation script is available upon request.

[8]Although the words in this study were not organized into pairs, the corresponding power calculation is very similar. Note that we are considering only t-tests conducted across all speakers—which Fourakis & Iverson (1984) focus on—and not those conducted within individual speakers.

[9]Note that this statement applies to the binary "is the effect zero?" question considered here, but not to the actual *estimate* of the effect size, which is affected by factors similar to those influencing power.

Affiliations

James Kirby, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, Scotland EH8 9AD United Kingdom
Morgan Sonderegger, Department of Linguistics and Centre for Research on Brain, Language, and Music, McGill University, Montreal, Quebec H3A 1A7 Canada

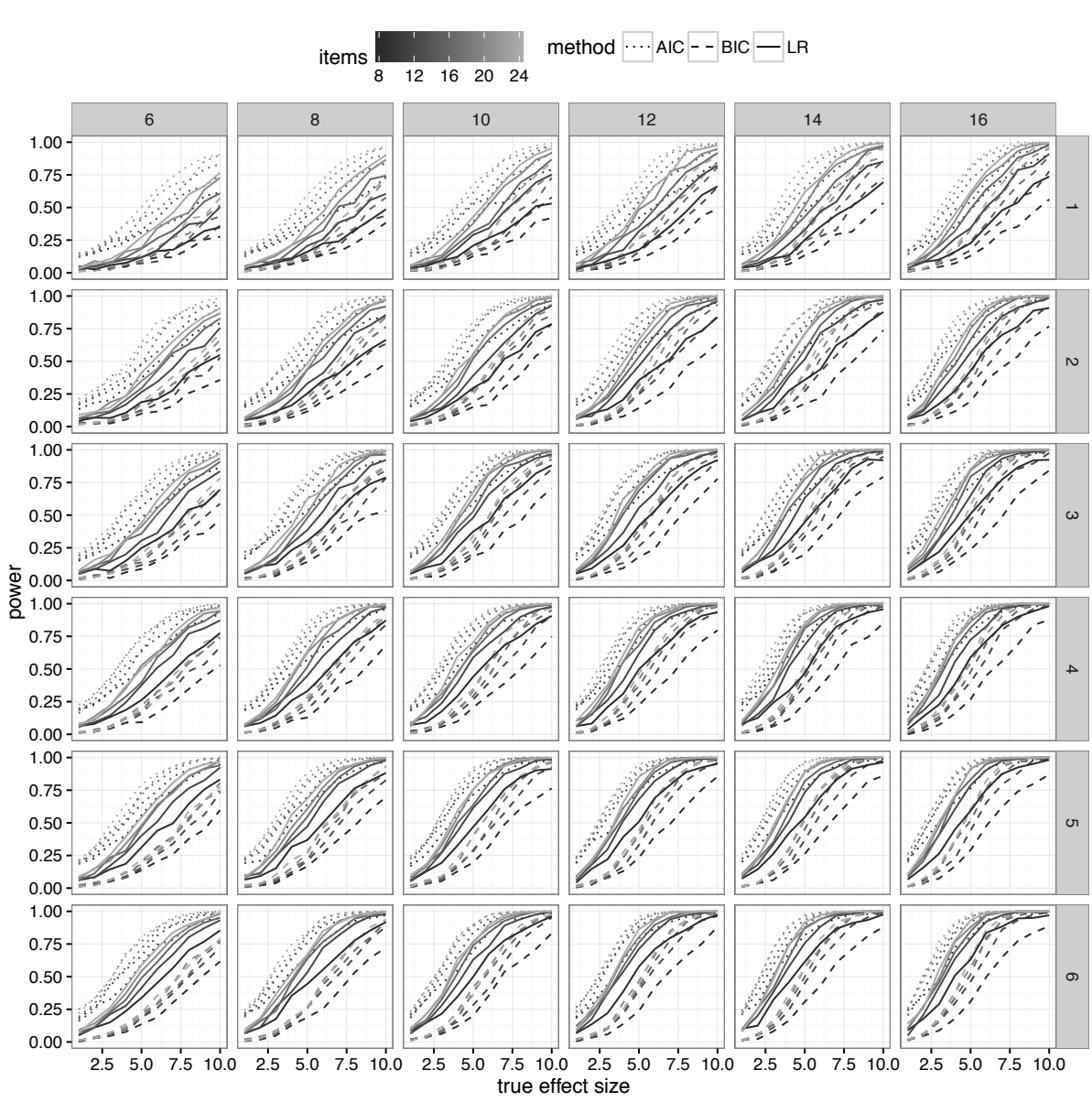

Figure 1: Power curves resulting from different sweeps simulated from the Roettger et al. data for three different model selection criteria. Columns show simulated number of subjects; rows show number of repetitions of each item per subject. Darker colored lines are for runs with fewer items, lighter colored lines for runs with more items.

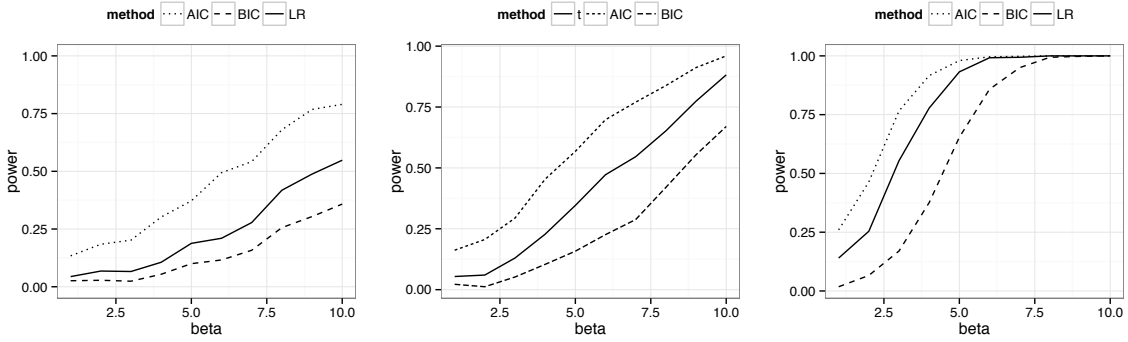

Figure 2: Left: Low power regime (6 subjects, 8 items, 2 repetitions). Center: Medium power regime (8 subjects, 12 items, 3 repetitions). Right: High power regime (16 speakers, 24 items, 6 repetitions).

Table 1: Parameters swept in simulation study.

parameter	*range*	*step*
number of subjects (n_s)	6-16	2
number of items (n_i)	8-24	4
number of repetitions (n_r)	1-6	1
true effect size (β)	0-10	0.5
model selection criterion	likelihood ratio, AIC, BIC	