Mixed-effects design analysis for experimental phonetics

Mixed-effects design analysis for experimental phonetics

James Kirby[1] & Morgan Sonderegger[2]

[1]University of Edinburgh
[2]McGill University

May 2018

DRAFT

Abstract

It is common practice in the statistical analysis of phonetic data to draw conclusions on the basis of statistical significance. While p-values reflect the probability of incorrectly concluding a null effect is real, they do not provide information about other types of error that are also important for interpreting statistical results. In this paper, we focus on three measures related to these errors. The first, *power*, reflects the likelihood of detecting an effect that in fact exists. The second and third, *Type M* and *Type S* errors, measure the extent to which estimates of the magnitude and direction of an effect are inaccurate. We then provide an example of *design analysis* (Gelman & Carlin, 2014), using data from an experimental study on German incomplete neutralization, to illustrate how power, magnitude, and sign errors vary with sample and effect size. This case study shows how the informativity of research findings can vary substantially in ways that are not always, or even usually, apparent on the basis of a p-value alone. We conclude by repeating three recommendations for good statistical practice in phonetics from best practices widely recommended for the social and behavioral sciences: report all results; design studies which will produce high-precision estimates; and conduct direct replications of previous findings.

1 Introduction

Statistical analysis is often used to reason about scientific questions based on a data sample, with the goal of determining "which parameter values are supported by the data and which are not" (Hoenig & Heisey, 2001, p. 4). Researchers in phonetics frequently reach such conclusions based on *significance*: the probability, or p-value, of obtaining an effect of the observed size (or greater), if the true effect were zero.

For example, consider a study of the effect of speech rate on Voice Onset Time (VOT) on short-lag stops in a particular language (e.g. Kessinger & Blumstein, 1997). The researcher fits a statistical model (say, a simple linear regression) in which the dependent variable is VOT, and the regression coefficient of interest β_1 is the slope of the regression line, representing an estimate of how a unit change in speech rate impacts VOT. A t-test is then conducted to assess whether this slope is different from zero. Judging from the literature, many researchers would conclude that there is an effect of rate if this difference is significant (i.e. if $p < 0.05$), and that if the difference is not significant ($p \geq 0.05$), VOT is unaffected by rate.

This focus on the p-value stems from a desire to avoid incorrectly rejecting the null hypothesis, when it is in fact true. This is obviously to be avoided, because we do not want to claim that an effect exists when it does not. However, p-values provide only limited information when interpreting studies, particularly if we are trying to interpret a study in relation to other work. To continue with the speech rate example, imagine two studies of the effect of speech rate on VOT, one of which finds a significant effect ($p \leq 0.05$) and one of which does not ($p > 0.05$). Given only the p-values, we are not in a position to assess which result is more plausible, since the p-value itself does not measure the probability that speech rate has a non-null effect on VOT. Moreover, the difference between the p-values may not itself be statistically significant (Gelman & Stern, 2006; Nieuwenhuis et al., 2011), so we cannot even conclude that there is a meaningful difference between the two studies.

In addition to interpreting significant effects, researchers are often interested in interpreting the *lack* of a significant effect, a so-called "null result". The temptation is often to conclude that if a coefficient is not significantly different from zero, that it does not have an effect on the dependent variable. Concluding from non-significance that there is no effect of an experimental manipulation is a well-known statistical fallacy; the p-value is *not* the probability that the null hypothesis is true, but rather the probability of observing an effect of a given magnitude, or larger, assuming that the null hypothesis *is* true. In order to avoid this pitfall, it is sometimes taught, or propagated in practice, that null results cannot be interpreted at all. However, this is not strictly speaking the case: null results can sometimes give information about likely parameter values or *effect size*—arguably the central goal of data analysis—but determining whether or not this is the case requires considering information other than the p-value of a test statistic.

In this paper, we discuss additional quantities that can give useful and complementary information to p-values: the probability of rejecting the null hypothesis assuming that it is false (statistical *power*) as well as errors of magnitude and sign in estimating effect size (*Type M* and *Type S* errors: Gelman & Tuerlinckx, 2000; Gelman & Carlin, 2014). Using simulation studies based on real experimental data, we illustrate three reasons researchers in phonetics should take into account power and effect size in addition to significance:

(1) Depending on statistical power, a non-significant result can still be informative.

(2) Errors in estimates of effect size can be substantial even when p-values are low.

(3) Estimates of effect size improve with power, and can be robust even when p-values are higher than a conventional threshold, e.g. $\alpha = 0.05$.

Using a case study of so-called *incomplete neutralization* (hereafter IN), we illustrate how (1)–(3) can affect conclusions drawn with respect to two questions, which are arguably always our goal in interpreting research studies: *what can we conclude about likely values of a parameter from a single study (Q1), as well as from a body of studies (Q2)?*

This exercise provides an example of *design analysis* (or *design calculations*; Gelman & Carlin, 2014): the use of statistical tools to reason about likely outcomes (= parameter values) of replications of a study—which is generally of greater interest than the statistical analysis of a single experiment.[1] Our focus here will be on design analysis for mixed-effects regression models, because these methods have become increasingly common for phonetic data analysis, and also because they can be somewhat more technically and conceptually challenging to implement. However, we note that the basic points (1)–(3) apply to most statistical methods commonly used to analyze phonetic data, including t-tests, classical ANOVA, classical regressions (without random-effect terms), and GAMMs.

None of the points we raise about power and effect size are novel (see e.g. Meehl, 1967; Cohen, 1988; Gigerenzer et al., 2004; Nieuwenhuis et al., 2011; Button et al., 2013; Colquhoun, 2014; Gelman & Carlin, 2014; Westfall et al., 2014; Vasishth & Nicenboim, 2016; Judd et al., 2017; Vasishth & Gelman, 2017; Brysbaert & Stevens, 2018, among others), but they are not typically addressed in interpretation of phonetic data. We believe that greater attention to these dimensions would improve the quality of phonetic research, both in terms of research design as well as interpretation. We hope the technical illustration provided in this paper will be of particular use to those researchers who are interested in performing power calculations and design analysis in the mixed-model context, but are unsure how to go about doing so.

The remainder of this paper is organized as follows. Section 2 provides some background on of power, effect size, and sign and magnitude errors, including the practical issue of how to compute them. Section 3 gives a case study of incomplete neutralization in word-final German stops, focusing on points (1)–(3), in the context of interpreting individual studies (Q1) and a body of studies (Q2), using power and effect size considerations in addition to significance. Finally, in Section 4 we conclude with some more general observations and recommendations.

To facilitate the use of power and effect size error calculations in phonetic research, code and data files for carrying out all analyses in this paper, as well as further worked examples, are archived as an Open Science Foundation project (Kirby & Sonderegger, 2018a).

[1] For example, in a study of whether there is a speech rate effect on VOT for lenis stops in English, we are less interested in whether the coefficient for this effect is significantly negative ($p < 0.05$) than in what can be concluded about the *true* value of the speech rate effect. By points (2) and (3), these are not the same thing.

2 Background

In this section, we define power and effect size before turning to considerations of power calculation, magnitude and sign errors, and design analysis. While there exist large literatures on each of these topics—in particular for the behavioral and social sciences—they are not usually discussed as part of mainstream statistical analysis of phonetic data. (For psycholinguistic data on the other hand, Vasishth & Nicenboim, 2016 cover similar topics, and our presentation is indebted to theirs.) We aim here to briefly summarize relevant concepts for our case study, and give relevant references where interested readers can follow up to learn more. Our case study (Section 3) provides a worked example showing one way these concepts can be applied to the analysis of phonetic data.

2.1 Power

In considering whether there is in reality an effect of a covariate or experimental manipulation, there are two essential types of errors a researcher can make: falsely concluding there is an effect when none exists (a *Type I error*, or "false positive"), or falsely concluding there is no effect when one in fact exists (a *Type II error*, or "false negative"). Type I errors are arguably more familiar, and everyday statistical practice places considerable emphasis on avoiding them. If a term is found to be statistically significant, many researchers would conclude from this that a Type I error is unlikely. The Type I error rate of a study is normally abbreviated as α, while the Type II error rate is β.[2]

Power, or one minus Type II error, is the probability of the statistical test correctly rejecting the null hypothesis when it is false. Like significance, power depends on several factors: the sample size, the true effect size, the acceptance threshold (α), and the amount of variability in the data. All else being equal, a significant result is less likely to be found for an experiment with a smaller sample, where the effect is small, where a more stringent cutoff is used (lower α), and/or where the variance is high. For a given statistical test, power is a function of these four quantities, and *power calculations* consist of determining power given values of the four quantities, or determining the necessary value of one quantity to achieve a given power level. Typically power calculations for research planning focus on sample size, because other determinants of power are less accessible (e.g. true effect size is determined by the phenomenon itself).

Note that while power and Type I error are related, one does not determine the other,

[2] In what follows, we refer to both "significance levels" (p-values) and Type I and Type II errors. Strictly speaking, this is incoherent: there are no p-values in the Neyman-Pearson hypothesis-testing paradigm, and notions such as power and alternative hypotheses are absent from Fisher's (1956) significance-testing framework. However, in practice these traditions are often conflated into a procedure sometimes called *null hypothesis significance testing* (NHST), where researchers discuss "significance levels" (a Fisherian concept) but treat p like an α threshold in the Neyman-Pearson framework, rejecting H_0 when p is less than some pre-specified value. In the tradition of late-period Fisher (1956), we recommend reporting exact p-values, effect sizes, and confidence intervals, rather than accepting or rejecting hypotheses on the basis of an α threshold; but at the same time, we believe that the Neyman-Pearson idea of power, as a way of formalizing precision, can also be informative. For an overview of these issues with some historical context, see Gigerenzer et al. (2004).

because of the roles of effect size, sample size, and variability. A common critical value for power, analogous to $\alpha = 0.05$ for rejecting the null, is $\beta = 0.2$, giving power of 80%.

Cohen (1988, 1992) and Hallahan & Rosenthal (1996) provide thorough but accessible introductions to power and related issues. Snijders (2005), Gelman & Hill (2007, Ch. 20), and Judd et al. (2017), as well as references therein, discuss power for mixed-effects models. Vasishth & Nicenboim (2016) provide a discussion of power in the context of linguistic data.

2.2 Effect size

The term *effect size* refers to any measure of the size of an effect. Unstandardized estimators, such as regression coefficients, are the simplest quantifications of effect size. Variables in a multiple regression model can be easily scaled to make different regression coefficients comparable (e.g. Gelman & Hill, 2007, Sec. 4.2).

Effect size can also refer to one of many standardized measures which quantify the magnitude of a treatment effect in a way that allows comparison across studies, or between effects in the same study. There are two broad families of effect size measures for data where the dependent variable is continuous: measures of association/variance explained (such as R^2, or the η^2 measure commonly used for ANOVAs), and standardized differences in means (see Kline, 2013, Ch. 5). We focus on the most common measure in the latter family: Cohen's d, defined as the difference between group means divided by a standard deviation appropriate for the data given the experimental design (Cohen, 1988) (which intuitively reflects the "amount of variability" in the data).

For mixed-effects models, different options exist for calculating standardized effect sizes, because these models contain several parameters (variance components) capturing different kinds of variability. Westfall et al. (2014) and Judd et al. (2017) show how to calculate Cohen's d as a function of experimental design for certain mixed-effects designs, and Gelman & Hill (2007, Ch. 20–21) demonstrate more general simulation-based effect size measures for mixed-effect models.

Like power, (observed) effect size is partially independent from significance. One point we wish to emphasize here is that accuracy of effect size estimates *does not automatically follow from (non)significance of model terms*. This means that some null results can give meaningful information about effect size, and reported effect sizes can be unreliable even for significant results, depending on power. We will illustrate these points further below.

For more on measures of effect size, see Cohen (1988, 1992), Kline (2013), and the references above.

2.2.1 Type M and Type S errors

Let β_x denote the true size of an effect of interest, and $\hat{\beta}_x$ the estimated effect size when an experiment is done to estimate it. Because $\hat{\beta}_x$ is a random variable, which will be different each time the experiment is run, the estimated effect size can be incorrect relative to the true effect size, in either magnitude or sign. Gelman & Carlin (2014)

define two corresponding measures of error of the estimated effect size, across different replications of the same experiment:

- The expected *Type M error* (or *exaggeration ratio*) is the expected value of $|\hat{\beta}_x/\beta_x|$: the extent by which the magnitude of the effect is exaggerated.
- The *Type S error* is the probability that the estimated effect has the wrong sign ($\text{sign}(\hat{\beta}_x) \neq \text{sign}(\beta_x)$).

Gelman & Carlin (2014) define Type M and Type S error as conditional on significance: how often we will be wrong about the direction or magnitude of an effect if only non/significant results are taken into account. It is also possible to consider *unconditional* Type M and Type S error—what these values would be if *all* results are considered, without regard to significance. Conditional Type M/S error are relevant for what can be concluded from any single study (our Q1), since the result of the study will reach significance or not. Unconditional Type M/S error are relevant for what can be concluded from an ensemble of studies (our Q2), assuming that both significant and non-significant results have been reported. We consider both conditional and unconditional Type M/S errors in our case study (Section 3).

For unbiased and normally distributed estimates, the relationship between Type M and Type S error and power can be reasonably approximated. Generally speaking, both types of estimates scale roughly with power; but Type M error increases faster than Type S error as power decreases. As demonstrated by Gelman & Carlin (2014), when power is low, Type M and S error can be surprisingly high, even for statistically significant results (see also Button et al., 2013; Ioannidis, 2008). Conversely, when power is high, Type M and S error will remain low. The existence of Type M and Type S errors means that it is not always the case that significant findings are "correct", in the sense of providing accurate estimates of the sign and/or magnitude of the effect of interest.[3]

For more on both Type M and Type S error, see Gelman & Tuerlinckx (2000); Gelman & Carlin (2014); Vasishth & Nicenboim (2016).

2.3 Power calculations and design analysis

Conducting a power analysis involves calculating power as a function of the experimental design and the data distribution. Power calculations are often carried out *a priori*, as part of experimental design, for example to assess the sample size necessary to detect an effect of a particular size. *Observed* or *post-hoc* power analysis—determining the power of an experiment that has already been conducted using the observed effect size—has (rightly) often been dismissed as pointless (Cox, 1958; Hoenig & Heisey, 2001; Senn, 2002; O'Keefe, 2007). This is because observed power can be computed directly from the p-value (given the study design, observed effect size, and variance);

[3] Cf. Kirby & Sonderegger (2018b), where it is somewhat misleadingly suggested that so long as an effect is significant, it can be trusted.

thus, power gives no additional information once the p-value is known (Hoenig & Heisey, 2001; Lenth, 2007; O'Keefe, 2007).[4]

However, there are situations in which power calculations made after the experiment has been conducted can still be informative. In particular, retrospective *design analysis* (Gelman & Carlin, 2014), where a range of plausible effect sizes are considered, can be helpful in addressing both our Q1 (what can be learned from a single study) as well as Q2 (what can be learned from a body of studies of a given topic).

There are two key differences between retrospective design analysis and an observed power analysis. First, in design analysis one considers a range of plausible effect sizes, which may or may not include the observed value. Second, the focus of a design analysis is not only to assess the power of the study, but also to determine the Type M and S errors. Design analysis can therefore tell us what we can learn from a study with a given design and sample size about likely values of a parameter of interest, regardless of whether or not the value of the parameter was statistically significant in our particular study.

Arguably, the most difficult part of design analysis (or *a priori* power analysis for that matter) is determining what range of effect sizes to consider. While the true effect size is generally unknown (that's why the researcher is conducting the experiment in the first place!), delimiting a range of plausible effect sizes is usually possible. For example, suppose you are studying incomplete neutralization of aspirated, ejective, and plain stops in Klamath, which are said to be neutralized in final position (Blevins, 1993). The effect of interest is the degree to which word-final consonant laryngeal class affects preceding vowel duration. While there is a literature on incomplete neutralization effects, there may be no previous phonetic work on Klamath, or on neutralization of an ejective/plain/aspirated contrast. A lower bound on plausible non-zero effect sizes could be obtained by a survey of the incomplete neutralization literature, to give a sense of how small vowel length contrasts are before they are considered perceptually neutralized (probably 4–15 ms), or using the just-noticeable-difference for vowel duration (about 5 ms: Nooteboom & Doodeman, 1980). An upper bound could be obtained by surveying studies of how laryngeal class affects previous vowel duration word-finally in languages without final neutralization (Chen, 1970), and taking the lowest values from reliable studies ($\sim$30 ms). Thus, a range of about 4–30 ms would be reasonable. Even for areas where less is known from previous related work, it should usually be possible to establish plausible effect sizes within an order of magnitude; Gelman & Carlin (2014, pp. 7–8) provide some useful guidelines.

2.4 Techniques

2.4.1 Calculating power

Tools for calculating power are now widely available for a number of experimental settings, including those typically used in phonetic research. Broadly, these fall into two types:

[4] Intuitively: if a significant effect was/was not found, you will compute that power must have been high/low to give an effect of the observed size.

Option 1: Closed-form solutions For many simple tests such as differences between sample means, power can be found by specifying sample size, effect size, and Type I error threshold α, as shown in textbook treatments of power analysis (e.g. Chow et al., 2008). Although most statistical analyses phoneticians are now doing are more complicated than such simple cases, they can still provide a rough estimate of power for a given term by approximating the relevant effect sizes and degrees of freedom.

For example, consider some hypothetical German incomplete neutralization data, of the type to be considered in our case study, in which vowel duration is measured before phonologically voiced and voiceless consonants. We can conduct a two-tailed test between two independent samples, with the goal of determining whether the group means are significantly different at the 0.05 level. Assume that we have $n = 32$ participants, 16 in each group, with normally distributed group means of $\mu_1 = 50$ ms and $\mu_2 = 45$ ms. For ease of exposition, we further assume that the population standard deviation σ is known and shared across groups; in this case, $\sigma_1^2 = \sigma_2^2 = 100$. This means we can use a z-test, instead of the more common t-test. The test statistic is then:

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\frac{2\sigma^2}{n}}}$$

Assuming normally distributed means, the formula for power is

$$1 - \beta = \Phi(z - z_{1-\alpha/2}) + \Phi(-z - z_{1-\alpha/2}),$$

where Φ is the standard normal cumulative distribution function with a mean of 0 and a standard deviation of 1, and $z_{1-\alpha/2}$ is the critical value of the test statistic corresponding to a two-tailed test with significance level α (here, $z_{1-\alpha/2} = 1.96$). For a fixed $n_1 = n_2 = 16$, this corresponds to power of just 0.29:

$$\begin{aligned} z &= \frac{50 - 45}{\sqrt{\frac{2 \times 100}{16}}} = \frac{5}{\sqrt{\frac{200}{16}}} = 1.41214 \\ 1 - \beta &= \Phi(1.41214 - 1.96) + \Phi(-1.41214 - 1.96) \\ &= 0.2926065 + 0.000370134 \\ &= 0.2929766 \end{aligned}$$

Doubling the sample size increases power to 0.52. To achieve power of 0.8, 63 participants per group would be required.

Carrying out such calculations by hand, as we have here, is usually not necessary. Most statistical software packages provide functions to perform power analyses for all basic types of hypothesis tests, such as comparison of proportions, chi-squared tests, t-tests, and the z-test example illustrated above. Some examples using the `pwr` library (Champely, 2017) for `R` (R Core Team, 2016) are provided in the OSF Project accompanying this article (Kirby & Sonderegger, 2018a).

It is possible to extend these simple calculations to more complex designs, including the mixed-effects regression models now common in phonetic data analysis. A recent example is that of Westfall et al. (2014), who propose analytic methods of power analysis for designs with a single fixed-effect term and crossed random factors, helpfully implemented in a user-friendly online calculator (`http://jakewestfall.org/power/`). Such tools are invaluable for prospective power analysis and experimental design, and we encourage researchers to make routine use of them to insure that experimental work is of high power (see Brysbaert & Stevens, 2018 for worked examples). However, such approaches usually make simplifying assumptions, such as balanced data and no covariates, which do not hold for many actual phonetic studies. This means that for the purposes of design analysis, a different or complementary approach is useful.

Option 2: Simulation For many studies which include complex designs, unbalanced observations, and multiple covariates, developing closed-form solutions for power can be difficult or impossible. In these cases, power can be approximated to a more or less arbitrary degree of accuracy through Monte Carlo simulation. The chief advantages of this approach are (a) flexibility and (b) the ability to more accurately reflect the underlying structure of the statistical model. Simulation also helps us to understand our ability to detect effects of arbitrary size, as it is simple to generate underlying data sets in which a given effect does or does not exist. The caveat is that simulation-based approaches to power are closely tied to the specific model and data used (see Section 3.5.2 below). For some other examples of simulation-based approaches to understanding statistical properties of experimental design see e.g. Gelman & Hill (2007); Baayen et al. (2008); Barr et al. (2013); Winter (2015); Judd et al. (2017); Jäger et al. (2017, Appendix B); Vasishth & Nicenboim (2016); Vasishth & Gelman (2017).

Simulation-based design analysis requires three components:

1. $\mathcal{M}$, a statistical model either fitted to previous data or specified using known values;
2. D, a dataset of interest;
3. β_x, an effect size for predictor of interest x.[5]

The basic procedure is then as follows:

(a) Use $\mathcal{M}$ to simulate dependent variable values for D given β_x;

(b) Re-fit $\mathcal{M}$ to D;

(c) Repeat steps (a) and (b) many (hundreds or thousands of) times.

[5]Somewhat confusingly, β is typically used as shorthand for both the Type II error rate as well as for regression coefficients. Here we designate regression coefficients as β_x for clarity.

The percentage of time that the fitted model finds a significant effect of x approximates power as the number of simulations increases. We illustrate this procedure in Section 3.2 below.

A number of software packages are now available for simulating power in mixed models, including the `simr` (Green & MacLeod, 2016), `clusterpower` (Reich & Obeng, 2013), and `longpower` (Donohue et al., 2016) packages for `R`. There are also non-open-source packages such as MLPowSim (Browne et al., 2009), which features an option to output `R` code. These packages differ in the types of data and models they allow, with `simr` allowing for the most generality at the time of writing. While we use `simr` functions in the study reported here, we try to emphasize general principles rather than particular software packages.

It is important to note that the general methodology described above requires various choices to be made at the point of implementation. For example, step (a) requires specifying how to incorporate random effects (e.g. conditioning on fitted values, versus simulating new values), and one might want to use an existing dataset D or have a procedure for simulating a dataset of interest. Software and package developers will have necessarily made such decisions, but they may not be appropriate for all use cases. The researcher can either just use the decisions made in a particular implementation (as we do here with `simr`), or make their own decisions by coding simulation-based power analysis from scratch (as in an earlier version of this work: Kirby & Sonderegger, 2018b). Arnold et al. (2011) provide discussion of the general methodology of simulation-based power calculations for mixed models; see also Gelman & Hill (2007, Ch. 20) and Hox (2010, Ch. 12).

2.4.2 Calculating Type M & Type S errors

Type M and Type S errors are less familiar concepts, so there are not many pre-existing tools to calculate them. Gelman & Carlin (2014) provide a calculator for the simple case of a t-test, which can be applied as an approximation to effects in more complex models (like Option 1 for power, above). Otherwise, one must proceed by simulation. Type M and Type S error can be calculated in the same simulation as power simply by keeping track of the percentage of cases where the predicted and true effect size have the same sign, and the mean magnitude of the ratio of predicted to true effect size. This is relatively straightforward using pre-existing simulation power calculation functions; we provide an example in code in the OSF project accompanying this article (Kirby & Sonderegger, 2018a), used in the case study in Section 3.2 below.

3 Case study: Incomplete neutralization

By way of illustration, we take as a case study the issue of the incomplete neutralization 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 traditionally described as 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 long 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.

Here, we will not take a position on whether or not incomplete neutralization is "real", or the theoretical implications of its (non-)existence. Rather, we are interested in IN because this literature contains both studies which find statistically significant evidence for acoustically incomplete neutralization (Mitleb, 1981a,b; Port & O'Dell, 1985; Port & Crawford, 1989; Roettger et al., 2014) alongside those that do not (Fourakis & Iverson, 1984; Jassem & Richter, 1989). Moreover, effects which are found are always with a positive sign, but with differing magnitudes.

What are we to make of this body of findings? We will show that studies showing non-significant effects are primarily low- or medium-powered, while the most highly powered study (Roettger et al., 2014) finds a significant effect. Taken together, however, this body of studies provides good estimates of both the magnitude and the direction of IN effects.

3.1 Data and method

We explored the ramifications of different power regimes by conducting simulations using data from an experiment examining incomplete neutralization of word-final laryngeal contrasts in German (Roettger et al., 2014). The effect of interest is how vowel duration depends on the phonological voicing specification of the following (word-final) consonant. We examine how power and Type M/S errors for the effect of interest depend on aspects of the study (number of subjects, items, and repetitions) by varying these aspects in our simulations.

The case study considered here is Experiment 1 of Roettger et al. (2014).[6] These authors recorded 16 native speakers of German producing singular forms of nonword nominals (e.g. [goːp]) in response to auditory primes containing either a voiced or a voiceless variant (e.g. [goːbə] or [goːpə]). Phonologically, the singular form of each item in a prime pair is expected to be identical. Each speaker produced one repetition of each target item in response to 24 such critical pairs. A linear mixed-effects model was used to estimate the duration of the vowel preceding the stop as a function of the stop's underlying voicing specification, alongside a number of control predictors. The model included by-subject and by-item random intercepts along with random slopes for voicing. The results indicated that speakers produced longer vowels before underlyingly voiced stops, as assessed by a statistically significant difference in the voicing coefficient in a likelihood-ratio test. The magnitude of the voicing effect was estimated to be 8.6 ms (SE = 2.03 ms).

In order to undertake our design analysis, we cannot simply use these numbers, but instead need a range of plausible effect sizes. For German IN, published estimates

[6] We thank Timo Roettger and Bodo Winter for sharing their dataset with us, as well as for permission to make this dataset publicly available.

have ranged from around 4 ms (Port & Crawford, 1989) to over 20 ms (Mitleb, 1981b), though most studies find 4–14 ms. We might also consider estimates from other languages, such as Dutch (3.5 ms: Warner et al., 2004) or Russian (6 ms: Dmitrieva et al., 2010). Together, these estimates provide us with a reasonable range in which to explore the ramifications of effect size on power in a mixed-model setting.

3.2 Simulation procedure

We simulated a range of datasets by varying sample and effect size to explore their effects on power (Table 1).[7] Sample size was varied by altering the number of subjects (n_s), number of items (n_i), and number of repetitions (n_r) over a range of values those found in the IN literature. Effect size (β_x) was varied from 2 to 10 ms. Although previous work suggests that values in the 10-20 ms range are also possible, here we focus on the lower end of the range, in order to more easily illustrate the differences between different power regimes, and because published effect sizes are likely to be inflated (Ioannidis, 2008; Button et al., 2013; Szucs & Ioannidis, 2017). In addition, while varying residual variance and amount of variability among subjects and items also affect power, here we elected to hold these factors constant in the interest of expositional clarity.

Table 1: Parameters swept in simulation study.

parameter	*range*	*step*
number of subjects (n_s)	6-26	4
number of items (n_i)	10-30	5
number of repetitions (n_r)	1-6	1
true effect size (β_x)	2-10	1

For a given set of parameter values (β_x, n_s, n_i, n_r), a single simulation run was performed by the general procedure described in Section 2.4.1. $\mathcal{M}$ was taken to be the linear mixed-effects model fitted to the original Roettger et al. data, with effect size of following consonant voicing replaced by β_x. In each simulation run, D was constructed as follows:

1. Choose a set of subjects:
    - If $n_s < 16$ (the number of subjects in the original dataset), sample (without replacement) a random subset of subjects.
    - If $n_s \geq 16$, concatenate subjects from the original dataset to make a list of size n_s.
2. Choose a set of items in a similar fashion to the set of subjects (using n_i instead of n_s), relative to the 24 items in the original dataset.

[7] For details of how these factors affect Type I error, see Barr et al. (2013); Matuschek et al. (2017); Winter (2015).

3. Combine the chosen subjects and items into subject/item pairs, and concatenate the subsets of the dataset corresponding to each subject/item pair, making a temporary dataset.

4. Concatenate this temporary dataset n_r times, to make D. This corresponds to a design where all subjects produce all items, possibly multiple times.

Next, $\mathcal{M}$ was used to simulate dependent variables for D given β_x, and $\mathcal{M}$ is re-fit to D, using functions from the `simr` package (Green & MacLeod, 2016); see the OSF project associated with this article for code and details (Kirby & Sonderegger, 2018a).[8] We then assess the significance of the consonant voicing term in the resulting model (p_{cons}) using a likelihood ratio test.[9] We store p_{cons}, as well as the estimated effect size $\hat{\beta}_x$ and the true effect size (β_x). By performing n_{sims} runs for each set of parameter values (β_x, n_s, n_i, n_r) we obtain estimates of power (the fraction of runs where $p_{cons} < 0.05$), Type M error (mean value of $|\hat{\beta}_x/\beta_x|$), and Type S error (the fraction of runs where sign($\hat{\beta}_x$) $\neq$ sign(β_x)) for these parameter values. We can also estimate 95% confidence intervals for each quantity, showing the uncertainty resulting from only using $n_{sims} < \infty$ runs.

We performed simulations with $n_{sims} = 1500$, which appears sufficient to estimate power and Type M error reasonably accurately, but Type S error would be better estimated with higher sample sizes (see Figures 4–7 below, especially the bottom row of Figure 5).

3.3 Results

Before proceeding, we must emphasize that our discussion is in an important sense narrowly applicable to the Roettger et al. (2014) data, because power and Type M/S error estimates are highly dependent on particular properties of the dataset (see Section 3.5.2). As such, Figure 1 does not present accurate power estimates for a given number of subjects/items/repetitions for an arbitrary study. Our discussion also mostly considers the effect of overall sample size on power and Type M/S error, without distinguishing between the different effects of subjects, items and repetitions, but as a general rule in phonetic experiments it is more important to have additional items than additional repetitions (Winter, 2015).

[8] The technically-minded reader may wonder why our simulation procedure includes the step of constructing a new dataset by choosing random subsets of subjects and items. This is slightly more complex than the simple cases presented in most simulation-based power tutorials (e.g. the `simr` vignette), where new subjects and items are generated using the random-effect parameters of $\mathcal{M}$ rather than explicitly constructing a new dataset. The reason is that the Roettger et al. dataset contains covariates—independent variables other than the term of interest (voicing)—so re-fitting $\mathcal{M}$ requires choosing covariate values for each row of the new dataset D. One can either simulate covariate values (as in MLPowSim: Browne et al., 2009), or simply duplicate the covariate values for each subject-item pair. While both approaches have advantages, our code uses the latter. An important consequence is that in steps (1)–(2) above, we only use subsets of subjects and items such that the same model used to fit the original dataset can be fitted. For example, if the first random subset of items only contains items with bilabial place of articulation, we sample a new subset, since the original model fits a place of articulation control variable with three values (bilabial, alveolar, velar).

[9] For more on how model comparison strategies impact power, see Kirby & Sonderegger (2018b).

3.3.1 Power

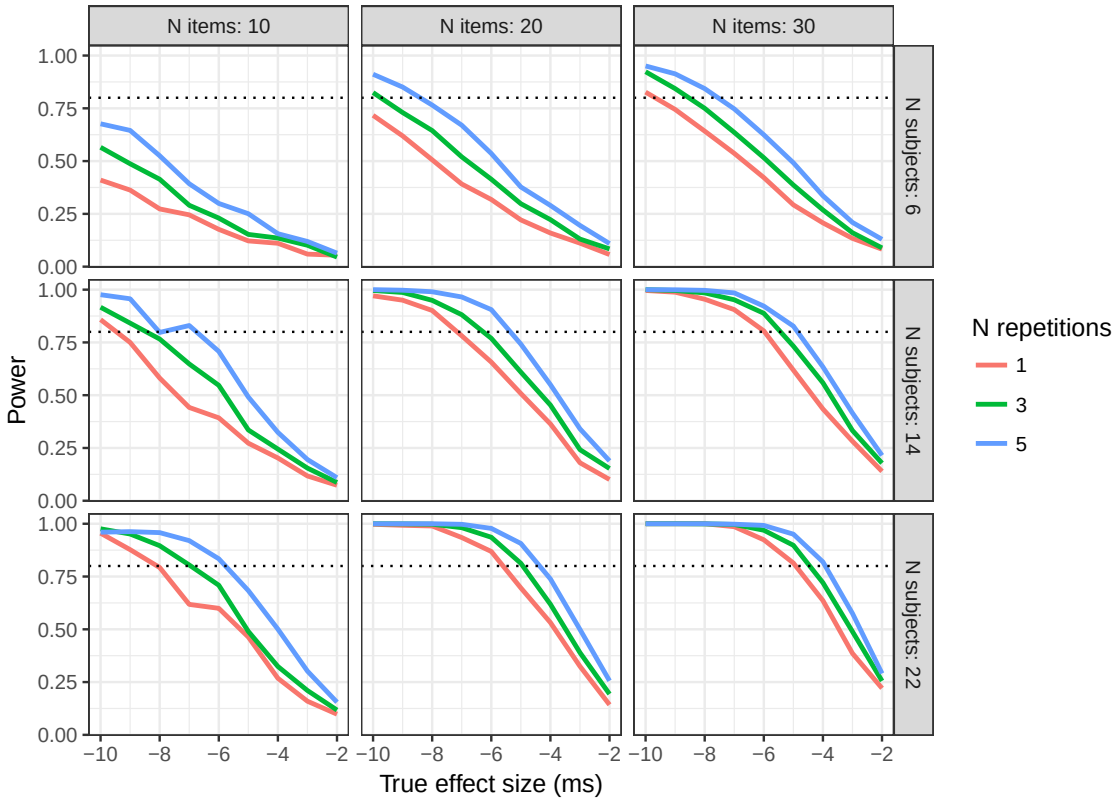

Figure 1: Power curves based on simulation ($n_{sims} = 1500$) from the Roettger et al. (2014) data, for a subset of parameter values used in the full set of simulations. Rows show simulated number of subjects; columns show number of repetitions of each item per subject; line color shows the number of repetitions of each item per subject. The dashed line indicates power of 0.8.

Figure 1 shows the results of these simulations, for an illustrative subset of parameter values: how power (on the y-axis) varies as a function of the true effect size β_x (on the x-axis) for different sample sizes. (Figure 8 in the Supplementary Materials shows the full set of results.) Each curve in the figure represents a different study design, with different choices for the number of subjects, items, and repetitions. This type of plot 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 as expected: as the sample size and the true size of the effect increase, so too does power.

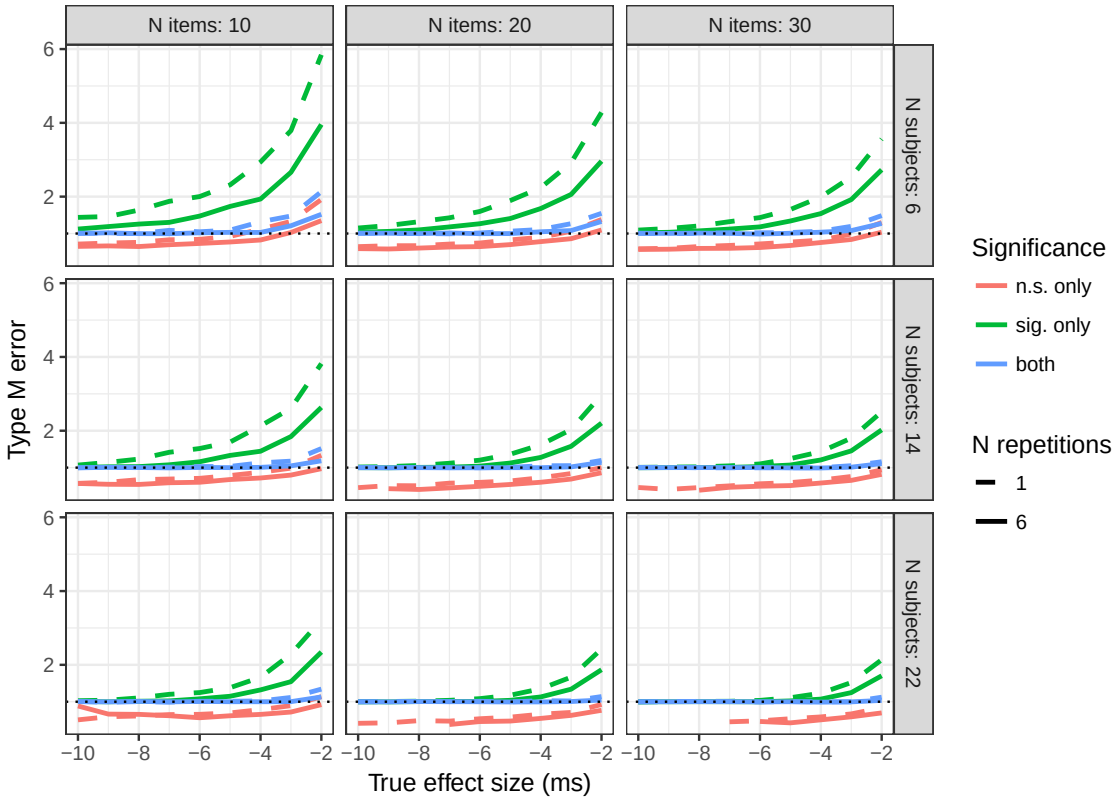

Figure 2: Type M error curves based on simulation ($n_{sims} = 1500$) from the Roettger et al. (2014) data, for a subset of parameter values used in the full set of simulations. Rows show number of subjects and columns show number of items, while line type shows the number of repetitions of each item per subject. Different line colors show the expectation of the absolute value of the estimated effect size divided by the true effect size for runs statistically significantly different from zero, not significantly different from zero, and across all runs. The dotted horizontal line at 1 shows optimal Type M error.

3.3.2 Type M error

Figure 2 shows how Type M error varies as a function of the true effect size β_x for the same data set, again for a subset of parameter values. (Figure 9 in the Supplementary Materials shows Type M error for the full set of simulations.)

Type M error is shown conditioned on significance (i.e. calculated only for runs with significant or non-significant effects), and unconditioned. We focus here on the conditioned results, as any single study to be interpreted by a researcher falls into one of these cases. For any set of parameters, Type M error is lower for non-significant results than for significant results, and thus Type M error unconditioned on significance lies

between the two.

For studies with high power, Type M error is relatively low for significant effects (i.e. the exaggeration ratio ≈ 1), and below 1 for non-significant effects. Intuitively, this is because when power is high a significant result is likely to be "correct", while for a non-significant result to obtain, the effect size must have been underestimated by chance. As power decreases (smaller n_s, n_i, n_r, or β_x), effect size magnitude tends to be overestimated for significant effects, especially if they are small. Intuitively, this is because when power is low, significant results often come from obtaining an inflated effect size that is large enough to cross the α threshold. Interestingly, when Type M error is calculated over all runs, it is usually near 1. This means that on average, estimated effect size is not overly inflated if both non-significant and significant results are considered (except when power is very low).

3.3.3 Type S error

Figure 3 shows the mean Type S error rate as a function of the true effect size β_x for the same data set, again with only a subset of parameter values shown (Figure 10 in Supplementary Materials shows Type S error for the full set of simulations). A run was flagged for a Type S error if the parameter estimate was of the incorrect sign, regardless of the magnitude of the error. As for Type M errors, Type S error rates are shown calculated conditioned on significance, and unconditioned. Once again, for high-powered designs Type S error is minimized (close to zero), but as power decreases, the chance of an estimate having the wrong sign increases substantially. Unlike for Type M error, however, this is less of a problem for significant than for non-significant estimates, because Type S error is always lower for significant estimates for any choice of parameters (and thus unconditioned Type S error is between the two). Intuitively, Type S error is lowest for significant results because it is very unlikely to estimate an effect size that is both significant and has the wrong sign, relative to the true effect size.

3.4 Discussion: interpreting individual studies

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 Figures 1–3 with respect to these two questions, we will first consider three regimes in detail, corresponding to low, medium, and high sample sizes. These regimes roughly correspond to three studies from the existing IN literature. We refer to these regimes as "low power", "mid power", and "high power" for exposition, since relative power is the important difference between them. (Instead of sample size, we could have varied the amount of variability in the data, to make the same general points.) However, it should be remembered that the actual power in each regime depends on the true effect size, and thus power in the "high power" regime can be arbitrarily small for small enough true effect size, and so on.

For each regime, we pose the following question. Suppose we replicated the Roettger et al. study with a different sample size; what should we conclude in case of different outcomes (Q1)? Our discussion of these regimes illustrates points 1 and 2 posed in the

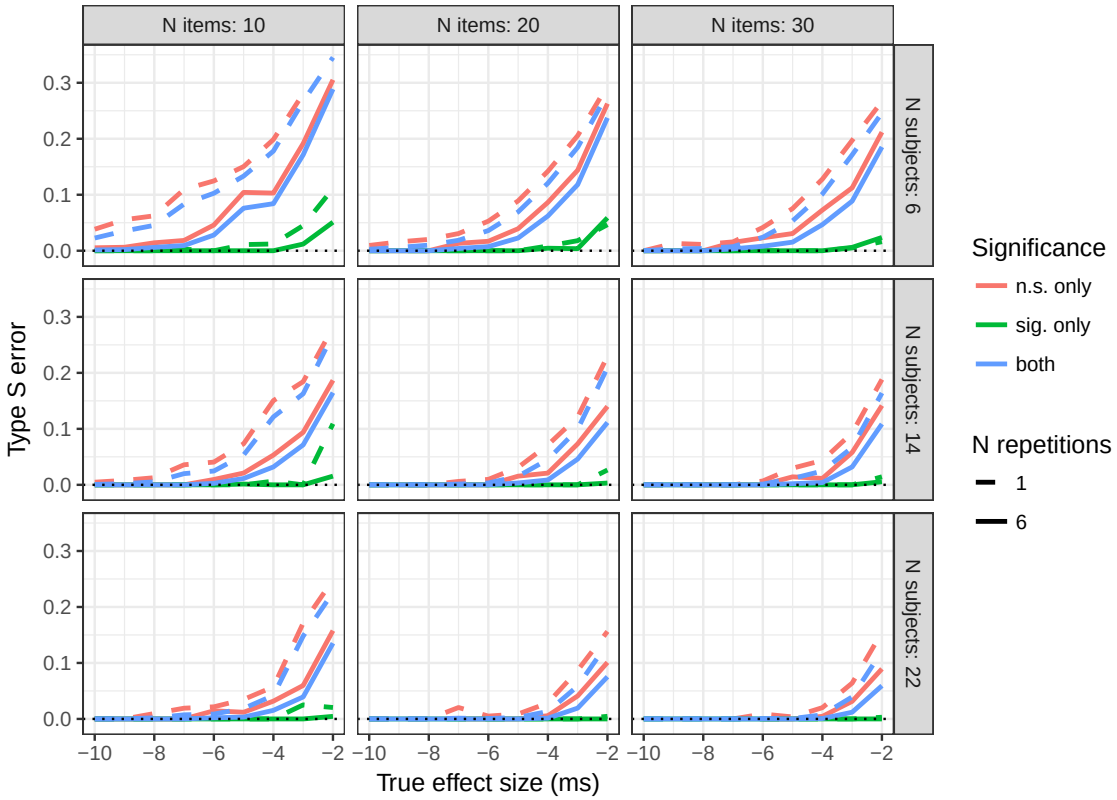

Figure 3: Type S error curves based on simulation ($n_{sims} = 1500$) from the Roettger et al. (2014) data, for a subset of parameter values used in the full set of simulations. Rows show number of subjects and columns show number of items, while line type shows the number of repetitions of each item per subject. Different line colors show the proportion of cases where the signs of the estimated and true effect size differ, for runs statistically significantly different from zero, not significantly different from zero, and across all runs. The dotted horizontal line at 0 shows minimal Type S error.

Introduction: depending on statistical power, a non-significant result can still be informative; and errors in effect size can be substantial even when p-values are low. We then show how non-significant results of an individual study can under some circumstances offer useful information about effect size (point 3 in Introduction), before turning to Q2.

3.4.1 Low-power regime

To illustrate a low-power regime, we select the power curve from simulations with 6 subjects, 10 items, and 1 repetition per item (Figure 4, red line). In this regime, power is always below 50%, far below the 80% cutoff, regardless of effect size (assuming that

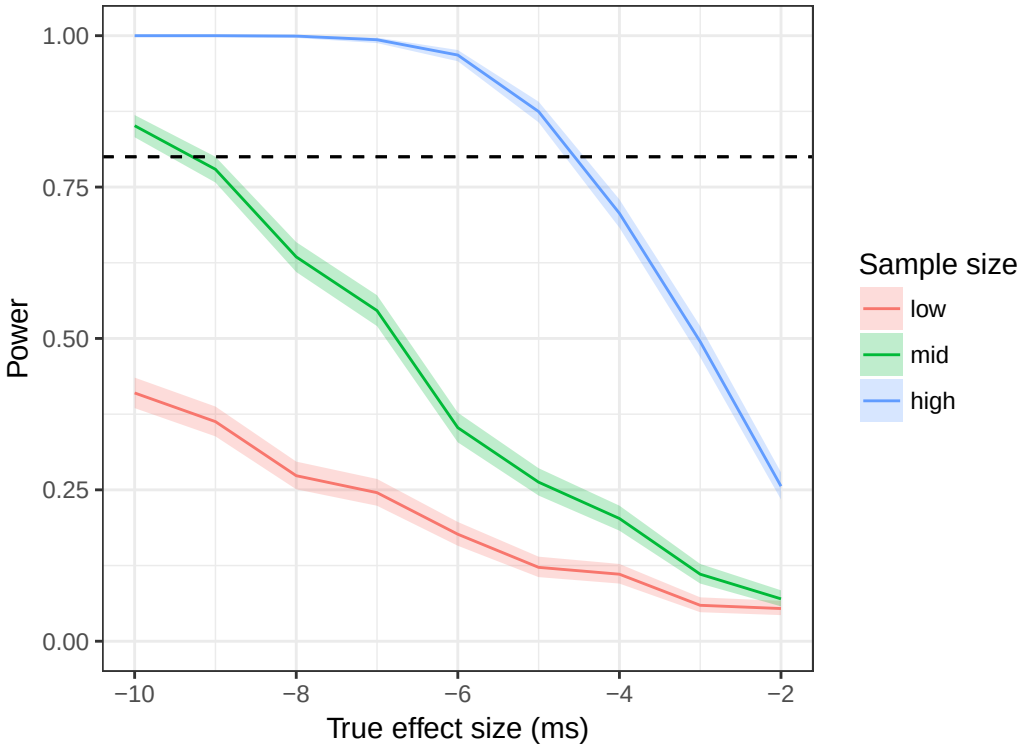

Figure 4: Simulated power calculations based on the Roettger et al. (2014) data for three power regimes: low (6 subjects, 10 items, 1 repetition), medium (10 subjects, 15 items, 1 repetition) and high (18 subjects, 25 items, 5 repetitions). Shading shows 95% confidence intervals. The dashed line shows power of 0.8.

the effect is $\leq$10 ms). However, the logic of interpretation differs depending on the statistical significance of the result:

- If we find a significant result ($p < 0.05$), we may conclude that observing an effect of this magnitude, or larger, is unlikely to have occurred if the true contribution of β_x is in fact zero.
- If we do *not* find a significant result, we should not be surprised, but we cannot interpret this lack of effect as evidence in favor of anything: a non-significant result ($p \geq 0.05$) is likely to occur whether there is in reality a true effect of $\leq$10 ms (low power) or not (high p-value).

In a low-powered study, then, a non-significant result is not informative, in the sense that it cannot be interpreted as refuting the hypothesis that there is an influence of the covariate of interest on the dependent variable.

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

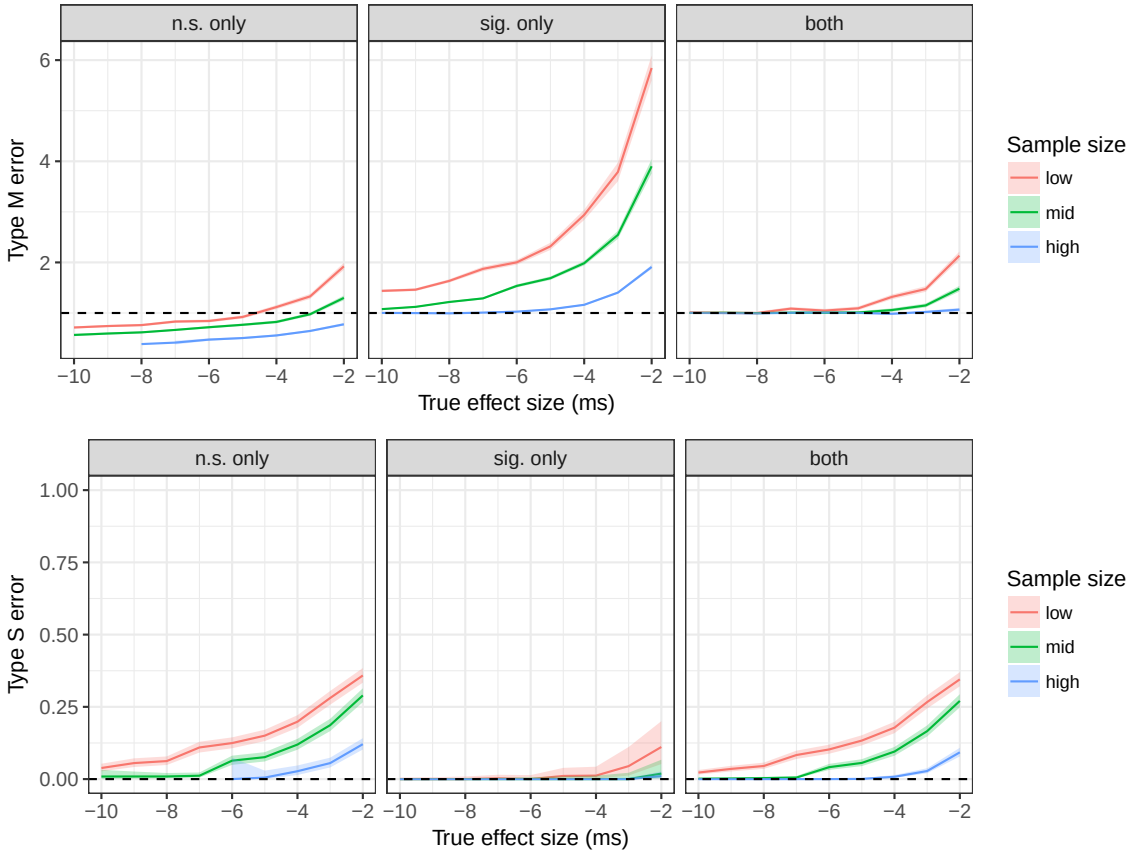

Figure 5: Simulated calculations of Type M error (top row) and Type S error (bottom row) based on the Roettger et al. (2014) data, using only non-significant (left column) or significant (center column) outcomes, or all outcomes (right column), for the same three power regimes as in Figure 4. For Type S error, calculations based on fewer than 25 significant or non-significant outcomes are omitted, as these do not give informative estimates. Note that the optimal values of Type M error is 1 (dashed line), while the optimal value of Type S error is 0 (dashed line) and the maximum is 1. Shading shows 95% pointwise confidence intervals.

for subsets corresponding in our terms to 1–2 items, and none of these tests are significant.[10] Approximate power calculations for these t-tests can be carried out using the information in their Table 2; even assuming a 10 ms 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 of Fourakis & Iverson does not

[10] 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.

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 to our understanding of the phenomenon under study.

However, even if we *do* find a significant effect in a low-power regime, we should not assume that it can automatically be trusted. This is because the likelihood of committing a Type M or Type S error is much higher when power is low, as can be seen in Figure 5 (red lines): statistically significant results from low-powered studies are virtually guaranteed to have an effect size inflated in magnitude, and possibly with the wrong sign, a phenomenon sometimes known as the "winner's curse" (Ioannidis, 2008; Button et al., 2013).

In the IN literature, a possible example of the winner's curse is the study of Mitleb (1981a, pp. 87–89), who found very large differences in preceding vowel duration between German word-final phonemically voiced and voiceless stops compared to other studies: 34 ms for bisyllabic words, and 23 ms for monosyllabic words, with small p-values in both instances ($p < 0.001$). For each word type, Mitleb's design corresponds in our terms to $n_s = 10$, $n_i = 4$, and $n_r = 4$. Although our simulations do not cover this small a number of items, we can see by extrapolating from the second rows of Figs. 2 and 3 that if the true effect size *were* ≤ 10 ms, Type M error for a significant result with this design would be substantially above 1, while Type S error would remain near 0. Thus, Mitleb's finding is consistent with a much smaller true effect size, and the observed effect size is probably exaggerated due to low power.

3.4.2 Mid-power regime

The mid-power regime is illustrated with simulations of 10 subjects and 2 repetitions each of 15 items (Figure 4, green line). In this regime, power is above the 80% mark only for the largest effect sizes considered.

- A significant result can be interpreted as unlikely to have occurred by chance. However, as seen in Figure 5, Type M and Type S errors can still be fairly high, depending on true effect size. For significant findings in particular, the effect sizes from mid-powered studies are almost certain to be inflated on average (although they will usually have the correct sign: see Figure 5 bottom row).

- The interpretation of a null result is still not straightforward under this regime, as it depends heavily on the effect size. If we have reason to believe the true β_x is, say, 10 ms 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 probably 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*.

A possible example from the IN 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, if it is indeed 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.

As seen in Figure 4, in mid-powered regimes power is particularly sensitive to the true effect size. 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.

While a mid-powered study has important drawbacks relative to a high-powered study from the perspective of power (and Type M/S error, discussed below), the mid-power case is important to consider because many published studies in experimental phonetics (like other behavioral sciences) are probably mid-powered: researchers use the smallest sample size that seems reasonably likely to detect an effect, to minimize cost and time.

3.4.3 High-power regime

Finally, we consider a design with 18 speakers, 25 items, and 6 repetitions (i.e. close to the design of Roettger et al., 2014, but with n_i=25 instead of 24, and 6 repetitions instead of 1). As seen by the blue line in Figure 4, power is above the 80% mark for the majority of the range of plausible effect sizes.

- 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 we are also licensed to interpret a *non*-significant result: 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 "in favor of" complete neutralization—at least in the sense of, if there is an effect, we may be confident that it is small.

Also of note is the Type M error rate of high-powered studies. For both significant and non-significant results, Type M error remains relatively near its optimum (=1) when power is high. If both significant and non-significant results are grouped together, the mean degree to which the true effect size is inflated is extremely low, even for very small true effect sizes. This illustrates that high power not only increases our confidence that we haven't accidentally failed to reject the null, but also our confidence in the reasonableness of our effect size estimates. This fact has important ramifications

for our Q2: what we are licensed to conclude from a body of studies (see Section 3.5.1 below).

3.4.4 Robustness of effect size estimates

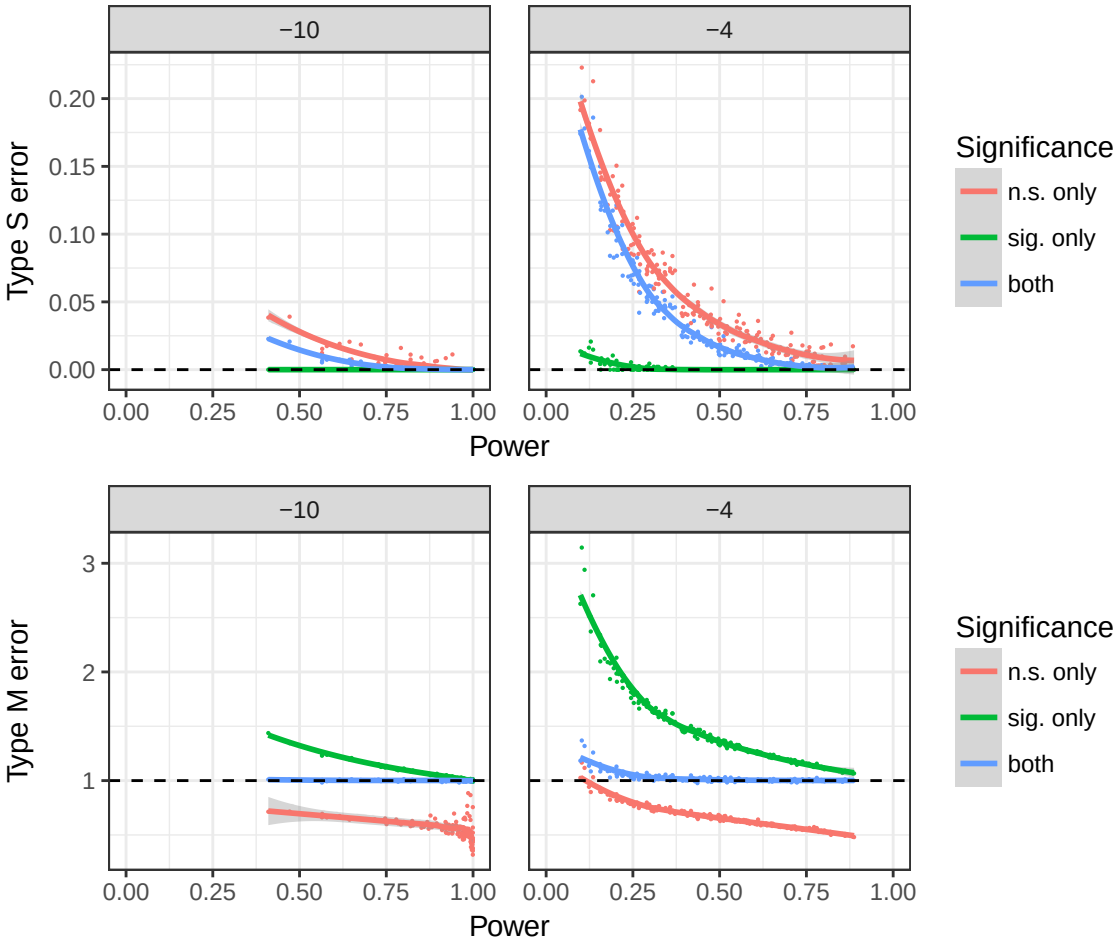

Figure 6: Type S error (top) and Type M error (bottom), as a function of power in simulations using the Roettger et al. (2014) data, using only non-significant (red) or significant (green) outcomes, or all outcomes (blue); for a large effect ($\beta_x = -10$ panel) and a small effect ($\beta_x = -4$ panel). Each point corresponds to one set of values of simulation parameters (Table 1), and lines/shading are LOESS smooths and 95% confidence intervals.

We now turn to point 3 raised in the Introduction: the fact that non-significant results can still give useful information about effect size, depending to a large extent

on power.

Figure 6 shows the relationship between power and Type M and Type S error, both conditional on significance and not (as in Figure 5). These plots show smooths across *all* simulation runs (where one point corresponds to one set of parameters in Table 1) for two effect sizes: a relatively large effect ($\beta_x = -10$ ms) and a smaller effect ($\beta_x = -4$ ms). Thus, different values of power in each panel effectively means different sample sizes. We can use these plots to think about what can be inferred from different studies of the same phenomenon, differing in sample size—and in particular what, if anything, can be inferred from non-significant results.

- For the larger effect, Type S error is effectively zero when power is at least 0.4, regardless of whether the result is statistically significant. When both significant and non-significant results are considered, Type M error is very low (≈ 1); for non-significant results, the magnitude may be underestimated, but probably only by half or less its true size.

- For the smaller effect, Type S error stays low (below about 5%) provided power is at least medium (0.5). Type M error is again very low down to very low power (0.25), if all results are reported. Focusing just on non-significant results, effect magnitude can again be underestimated, but again by only 50% or less.

- Also of note is that as power decreases, Type S error is affected more slowly than Type M error.

These results illustrate that, practically speaking, one can still use non-significant results to say something about (likely) effect size, *as long as power is high enough.* Non-significant results may still give useful information about effect direction, and to some extent effect magnitude—especially when the true effect size is reasonably large. This holds for even medium-powered studies, which will by definition frequently give non-significant results.

This generalization comes with an important caveat. When a result is not significant, the observed effect is (by the definition of significance) also likely to have been observed if the true effect size were *zero*. So, what can be minimally said about medium- to high-powered, non-significant results is that they are consistent with *both* null and non-null effects. What else one might conclude about the true effect size depends on the researcher's *a priori* belief about the true effect size, informed by expectations based on domain knowledge or prior work. (On this latter point, see Section 3.5.3 as well as the paper by Nicenboim, Roettger & Vasishth in this issue.)

This caveat suggests looking more closely at what can be concluded for different non-significant effects: what can be concluded when p fails to cross the pre-specified α threshold, but is low enough for the researcher to attempt interpretation of the result (frequently resulting in turns of phrase such as "marginally significant")? Figure 7 shows the relationship between power and Type M and Type S error, now across *all* simulations (effect size $\in (-2, -10)$), with non-significant results broken down into "marginal" ($0.05 < p < 0.2$) and $p > 0.2$ results.[11] We see that for "marginal" results,

[11] Although we use $p < 0.2$ as a cutoff here, in reality the relationship is gradient.

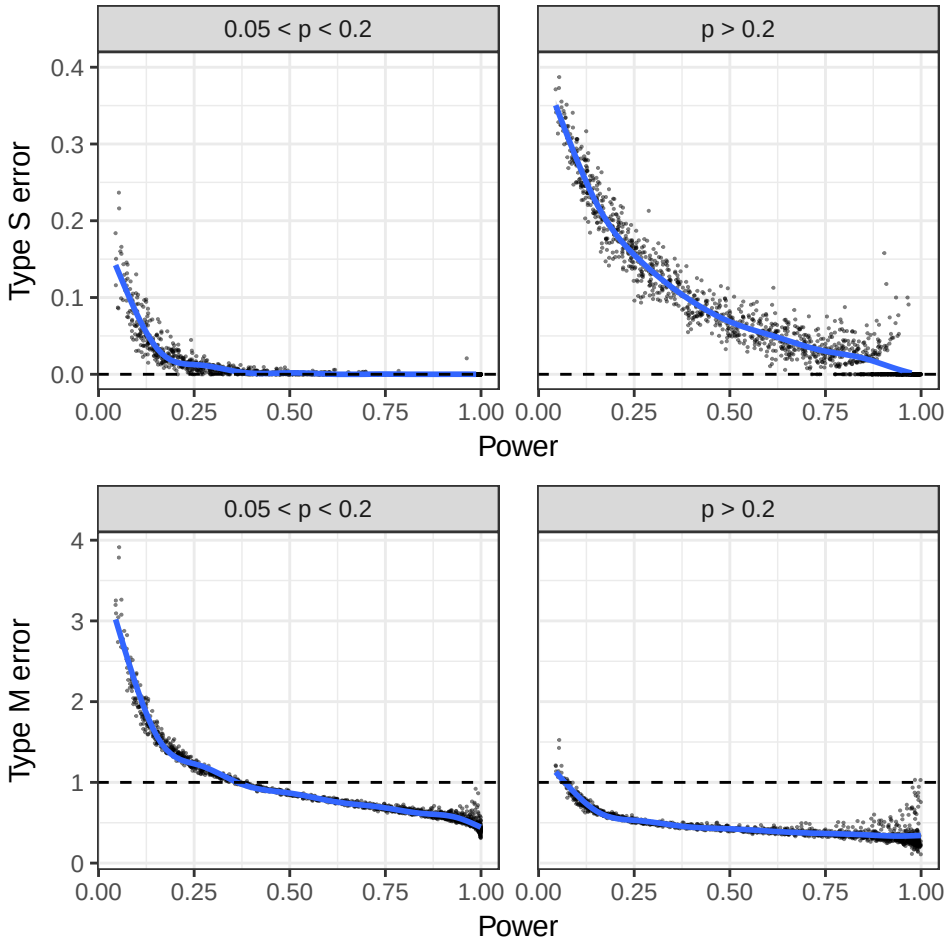

Figure 7: Type S error (top) and Type M error (bottom), as a function of power in our simulations using the Roettger et al. (2014) data, for non-significant effects with $0.05 < p < 0.2$ and with $p > 0.2$. Each point corresponds to one set of values of simulation parameters (Table 1), and lines/shading are LOESS smooths and 95% confidence intervals. (Note that CIs are mostly not visible.)

as long as power is not too low, the effect is almost certain to have the right sign, while the magnitude, though likely underestimated, is not wildly wrong (within <50% of correct value). For $p > 0.2$ results, Type S and Type M error are much worse, and the observed effect is also likely if the true effect size is zero. In this regime, then, nothing can be concluded about true effect size.

3.5 Discussion

3.5.1 Interpreting a body of studies

We now turn to our Q2: what can we conclude from a collection of studies of (more or less) the same phenomenon? Strictly speaking this is the domain of "meta-analysis" (Lipsey & Wilson, 2001; Cumming, 2013), a methodology for pooling the results of previous studies to determine likely parameter values. The paper by Nicenboim, Roettger & Vasishth (this issue) conducts a meta-analysis of 14 previous studies of German incomplete neutralization, and concludes that there is a small but real effect.

However, even in the absence of formal meta-analysis, the considerations of power and effect size discussed above in the context of individual studies can be used to interpret the German incomplete neutralization literature, which is representative of many literatures in phonetics showing "mixed" results (in terms of the significance of a parameter of interest). When we have a mixture of significant and non-significant results, what can we say about likely values of β_x? Do we have evidence for $\beta_x \neq 0$ (IN effect), $\beta_x = 0$ (no IN effect), or truly conflicting evidence?

As the above discussion suggests, the answer depends largely on the power of the studies involved. Consider just the case of whether or not β_x is equal to 0. If all of the studies concerned have high power, then those which find significant results provide us with evidence consistent with $\beta_x \neq 0$, while those that find null results can be interpreted as consistent with $\beta_x \approx 0$. 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 $\beta_x > 0$; the null results are not evidence for or against anything.

Concerning the actual value of β_x, lower-powered studies are more likely to incorrectly estimate its magnitude, and (less often) its sign, while higher-powered studies estimate β_x more accurately. This also plays into how to interpret an ensemble of estimated effect sizes from different studies. For example, finding a mixture of positive and negative β_x in different high-powered studies would give truly conflicting evidence on effect direction, while observing $\beta_x > 0$ in all studies except a couple low-powered ones would be consistent with the true effect in fact being positive.

In the case of the (German) incomplete neutralization literature, previous low- to medium-powered studies have found effects with a consistently positive sign, but differing in magnitude. Those studies showing non-significant effects are low-powered, while the one high-powered study (Roettger et al., 2014) of which we are aware finds a significant effect of 8.6 ms. This is consistent with most studies having sufficient power to correctly detect the sign of the (non-null) effect, but not to accurately estimate its magnitude. Thus, a reasonable interpretation based on the existing literature is not that there is "mixed evidence" for incomplete neutralization, but rather that there is a small, consistently positive effect, probably within the range considered in our simulations (2–10 ms).

3.5.2 Simulated power is conditional on your model and data

It is important to remember that the power estimates derived in the preceding sections are highly dependent on particular properties of the dataset used for the simulations. In the Roettger et al. data, there happens to be very little variation between subjects and items for the particular effect of interest (vowel duration). In the language of mixed models, this means that the random slope variances are quite small. Furthermore, the fact that this a fully crossed design (all subjects produced all experimental items in both conditions) means that subject and item variability in the intercept is not confounded with the effect of interest. As a result, power increases fairly rapidly, and as seen in Figure 1, medium-sized effects can be detected with a relatively small number of subjects and items.

Whether or not this is a typical situation for experimental phonetics is an open question. But the takeaway is that the plots in Section 3.3 should *not* be read as providing estimates for power, Type M and Type S errors for a design of so-and-so many subjects and items; they are *only* valid for this particular dataset. While the analytic approach of Westfall et al. (2014) can provide some more generality on how power is a function of sample size and experimental design, as previously discussed, it does not take covariates into account, which may result in underestimated power. On the other hand, our simulations only studied the power of main effects; even for identical designs, the power to detect interactions may be far lower (Gelman, 2018).

Similarly, in using these results to reason about the body of reported IN studies, we are assuming that the effect sizes and variance components are sufficiently similar to those we have calculated based on the Roettger et al. data that our power calculations are applicable. This may not always be true, for example if the different subject populations of different studies have different degrees of variability (e.g. a mono-dialectal versus multi-dialectal sample). Our general point here is simply to stress that, while the simulation-based approach can be highly accurate for particular designs and datasets, the results do not automatically generalize to other designs and datasets.

3.5.3 Type II error, precision, and uncertainty

The language and logic of Type I and Type II errors are indelibly linked to the Neyman-Pearson decision-theoretic paradigm of hypothesis testing, where hypotheses are accepted or rejected when a test statistic falls below some threshold. As argued by Fisher (1956, Ch. 4) as well as many others, while such a procedure is undoubtably useful in certain settings, it is not clear that it is the appropriate paradigm for experimental scientific inquiry.[12] In this regard, one might wonder why we have devoted an entire paper to the discussion of a concept embedded in such a framework. It is our experience that the Neyman-Pearson hypothesis-testing paradigm, or something like it, is what is familiar to the majority of researchers in phonetics. However, what we are ultimately advocating for is increased attention on the precision and uncertainty of estimates. Power, and Type M and Type S errors, are one way to understand and measure

[12]Even Neyman and Pearson themselves seem to have had their doubts as to whether their framework was well-suited for scientific research: see e.g. the discussion in Hurlbert & Lombardi (2009, p. 319) and references therein.

precision and uncertainty.

In Bayesian approaches to statistical inference, precision and uncertainty take on a more central role. While the core concepts are fairly simple, both the theory and practice of Bayesian statistics can be challenging, and we recognize that not all researchers will have the time or inclination to explore them fully. As we have tried to illustrate here, it is still possible to emphasize the precision and uncertainty of parameter estimates within a (frequentist) hypothesis testing framework, but we encourage interested researchers to explore Bayesian methods as well. McElreath (2015) provides an excellent and accessible introduction to Bayesian statistics, with many examples in `R`. Tutorials on the application of Bayesian techniques from a linguistic perspective include Nicenboim & Vasishth (2016); Sorensen et al. (2016); and Vasishth, Nicenboim, Beckman, Li & Kong (this issue).

4 Conclusions

In this paper, we have emphasized the importance of statistical power in the interpretation of phonetic data. Power calculations, along with the consideration of sign and magnitude errors, can help inform our understanding of both a single study, as well as a body of studies. We have seen that, depending on power, even results which are not statistically significant may nonetheless still be informative, in that they can still provide reasonable estimates of effect sizes, including providing evidence "for the null" in some sense.

At the same time, we have shown how low- and even medium-powered studies can also make substantial errors in estimating the sign and magnitude of effects, even when accompanied by a small p-value. By taking power and effect size errors as well as significance into account, phoneticians are better positioned to reason more carefully about findings of all types, not just those where p happens to be less than 0.05.

In addition, we have provided a concrete example from the literature, including a practical demonstration of how power and design calculations can be performed in the mixed-model setting in which many practicing phoneticians now find themselves working. Together with our accompanying `R` code (Kirby & Sonderegger, 2018a), we hope this helps researchers to perform their own power and design analyses.

However, while simulation is a useful tool, an appreciation of power and effect size errors can inform our reasoning even without it. Consider two studies with the same effect size, but different sample sizes. If the large-sample study finds a significant effect but the small-sample study does not, this is perfectly consistent with there being a true effect. Simply by giving factors such as sample size—one indicator of power—as much weight as we do the p-value in interpreting our model coefficients, we are in a better position to reason about the robustness of our results.

We conclude with three practical recommendations we believe to be beneficial for the phonetic sciences as a whole. Again, these recommendations are in no way novel—see for instance Wilkinson and the ASA Task Force on Statistical Inference (1999)—but we hope that this illustration in a phonetics context will encourage them to be employed more consistently in our field.

1. Report all results We strongly recommend reporting effect size and direction in phonetic studies, *regardless of statistical significance*. Doing so consistently as a field will result in more accurate estimates of the true sizes of effects across studies—as demonstrated by the fact that Type M error *unconditional* on significance remains near 1 (optimal) across most parameter values in our simulations. Researchers are already doing this when they report the full output of their regression analyses (including coefficient estimates and standard errors), but it is still not infrequent to find papers which only report p-values, or which only indicate whether p was at or below some threshold. In isolation, p-values do not communicate scientifically useful information.

Reporting both significant and non-significant results is particularly critical for small effects such as incomplete neutralization, where power of any given study is unlikely to be high. As a result, the effect size of significant results is likely to be inflated (high Type M error). This means that if only statistically significant effects are published and discussed, the problem gets worse, because the effect size across a body of studies will tend to be inflated as a whole (the "file drawer effect": Rosenthal, 1979; Button et al., 2013; Simonsohn et al., 2014). Of course, non-significant findings always have a plausible alternative explanation—this is the definition of significance—so bringing them into interpretation is trickier. Nevertheless, in general our understanding of phenomena such as incomplete neutralization is enriched by having a body of studies to interpret.

However, we emphasize that we are *not* recommending that reviewers and editors simply accept any and all results, or judge them all equally. This leads us to our second recommendation:

2. Conduct high-powered studies Our simulations also underscore the importance of conducting high-powered studies whenever possible, and taking power into account when interpreting statistical analyses. Since in phonetics, we are generally interested in effect sizes and directions, and since data collection and analysis can often by laborious and time-consuming, it is especially important that we be confident of our ability to detect effects of a particular size before we begin data collection. Another way to think of this is to aim to conduct studies with high precision, i.e. where the uncertainty surrounding the size and magnitude of the estimates is minimized (see Section 3.5.3).

For experimental studies, it is often possible to increase power/precision by increasing sample size, both in terms of subjects as well as in terms of items. While we have not conducted a formal survey, our impression is that, when compared to fields such as medicine and ecology, phonetic studies often have very small sample sizes. An opportunistic review of seven papers from the November 2017 issue of this journal containing statistical analyses of acoustic speech production data found the number of participants to range from 11 to 39; within experimental groups, however, sample sizes ranged from 24 to just 6 speakers. The first author has published studies with groups of 6, 10, and 20 participants. Our general impression is that such sample sizes are reasonably representative. As shown in Sections 2.4.1 and 3.4.3, however, even a nominally high-powered study can still produce worryingly uncertain estimates if the true effect is small, and a small effect is highly unlikely to be accurately estimated with just 6 participants per group—at least for effects the size of those considered in our case

study. Similarly, increasing the number of repetitions is generally no replacement for including more unique items (Winter, 2015). It is therefore important to think carefully about the interaction of sample size and known or expected effect size when planning and interpreting studies. For reasonably large effects, 24 participants may be perfectly adequate, but when smaller effects t are of interest, many more participants may be required.

Corpus studies, which are becoming ever more prevalent, present an interesting case. We suspect that power for detecting any given effect is often not high in corpus studies, for several reasons: the large number of predictors being modeled as affecting a dependent variable, high variability in the data (which lowers power), and inherently limited sample size. In this setting, applying commonly recommended analysis techniques (such as fitting the maximal random effect structure) will minimize Type I error, but can dramatically lower power (Matuschek et al., 2017). Null results in this setting will therefore frequently be uninterpretable, but can under some circumstances still give information about effect size.

3. Conduct direct replications Our final recommendation is to encourage direct replications of important studies, and for journals to publish them. This might take the form of a pre-registered study, such as that promoted by the Center for Open Science's *Registered Reports* concept, but a replication could also form a part of a larger study that also includes novel experimental results. As the discussion in Section 3.4 hopefully made clear, there can be substantial sign and magnitude errors in estimating effect size even when p-values are very low. Similarly, the effect size found from a single high-powered study should not be assumed to be accurate or infallible; direct replications should always be regarded as the gold standard.

Acknowledgements

This research was funded in part by SSHRC (#435-2017-092, #430-2014-00018) and FQRSC (#183356) grants to M. Sonderegger. Thanks to David Fleischer for research assistance. We are grateful to editors Bodo Winter, Timo Roettger, and Harald Baayen; two anonymous reviewers; Prof. Shravan Vasishth; and the audience at PaPE 2017 for thoughtful comments and discussion of this work. We alone are responsible for any errors of fact or interpretation. An earlier version of this paper, with reduced scope, was published as Kirby & Sonderegger (2018b).

Supplementary materials

Code and data files for carrying out all analyses in this paper, including worked examples are archived as an Open Science Foundation project at `https://osf.io/e4g5t` (Kirby & Sonderegger, 2018a).

Figures 8, 9, 10 show power, Type M error, and Type S error estimated via simulation with $n_{sims} = 1500$ from the Roettger et al. data, for each combination of parameter values shown in Table 1.

References

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, *11*, 94.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Blevins, J. (1993). Klamath laryngeal phonology. *International Journal of American Linguistics*, *59*, 237–279.

Browne, W. J., Lahi, M. G., & Parker, R. M. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. Bristol, United Kingdom: University of Bristol.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*, 1–20.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munaf, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365376.

Champely, S. (2017). *pwr: Basic Functions for Power Analysis*. R package version 1.2-1.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, *22*, 129–159.

Chow, S.-C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research*. (2nd ed.). CRC Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*, 140216.

Cox, D. R. (1958). *Planning of experiments*. New York: Wiley.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Dmitrieva, O., Jongman, A., & Sereno, J. (2010). Phonological neutralization by native and non-native speakers: The case of Russian final devoicing. *Journal of Phonetics*, *38*, 483–492.

Donohue, M. C., Gamst, A. C., & Edland, S. D. (2016). *longpower: Power and sample size calculators for linear mixed models*. R package version 1.0-16.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.

Fourakis, M., & Iverson, G. K. (1984). On the 'incomplete neutralization' of German final obstruents. *Phonetica*, *41*, 128–143.

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect. http://andrewgelman.com/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/.

Gelman, A., & Carlin, J. (2014). Beyond power calculation: assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*, 641–651.

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

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and bayesian single and multiple comparison procedures. *Computational Statistics*, *15*, 373390.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: what you always wanted to know about significance testing but were afraid to ask. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 392–409). SAGE Publications, Inc.

Green, P., & MacLeod, C. J. (2016). Simr: an r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498.

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, *34*, 489–499.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*, 1–6.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici*, *46*, 311–349.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

Jäger, L., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Jassem, W., & Richter, L. (1989). Neutralization of voicing in Polish obstruents. *Journal of Phonetics*, *17*, 317–325.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

Kessinger, R., & Blumstein, S. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*, 143–168.

Kirby, J., & Sonderegger, M. (2018a). Mixed-effects design analysis for experimental phonetics. Open Science Framework project: `osf.io/e4g5t`.

Kirby, J., & Sonderegger, M. (2018b). Model selection and phonological argumentation. In D. Brentari, & J. Lee (Eds.), *Shaping phonology*. Chicago: University of Chicago Press.

Kline, R. (2013). *Beyond significance testing*. Washington, DC: American Psychological Association.

Lenth, R. V. (2007). *Post hoc power: Tables and commentary*. Technical Report 378 Department of Statistics and Actuarial Science The University of Iowa.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McElreath, R. (2015). *Statistical Rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*, 103115.

Mitleb, F. (1981a). *Segmental and non-segmental structure in phonetics: evidence from foreign accent*. Ph.D. thesis Indiana University, Bloomington.

Mitleb, F. (1981b). Temporal correlates of "voicing" and its neutralization in German. *Research in Phonetics*, *2*, 173–191.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—Part II. *Language and Linguistics Compass*, *10*, 591–613.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*, 11051107.

Nooteboom, S. G., & Doodeman, G. J. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*, 276–287.

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate uses of statistical power analyses. *Communication Methods and Measures*, *1*, 291–299.

Piroth, H. G., & Janker, P. M. (2004). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics*, *32*, 81–109.

Port, R., & Crawford, P. (1989). Pragmatic effects on neutralization rules. *Journal of Phonetics*, *16*, 257–282.

Port, R. F., & O'Dell, M. (1985). Neutralization of syllable-final voicing in German. *Journal of Phonetics*, *13*, 455–471.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Reich, N. G., & Obeng, D. (2013). *clusterpower: Power calculations for cluster-randomized and cluster-randomized crossover trials*. R package version 0.5.

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., & Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *43*.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638641.

Senn, S. J. (2002). Power is indeed irrelevant in interpreting completed studies. *BMJ: British Medical Journal*, *325*, 1304.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.

Snijders, T. A. (2005). Power and sample size in multilevel linear models. In B. Everitt, & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1570–1573). Wiley Online Library volume 3.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, *12*, 175–200.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*, e2000797.

Vasishth, S., & Gelman, A. (2017). The statistical significance filter leads to overconfident expectations of replicability. In *Proceedings of the Cognitive Science Conference*. London.

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

Warner, N., Jongman, A., Sereno, J., & Kemps, R. (2004). Incomplete neutralization and other sub-phonemic durational differences in production and perception: evidence from Dutch. *Journal of Phonetics*, *32*, 251–276.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 220–245.

Wilkinson, L., & The ASA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Winter, B. (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.

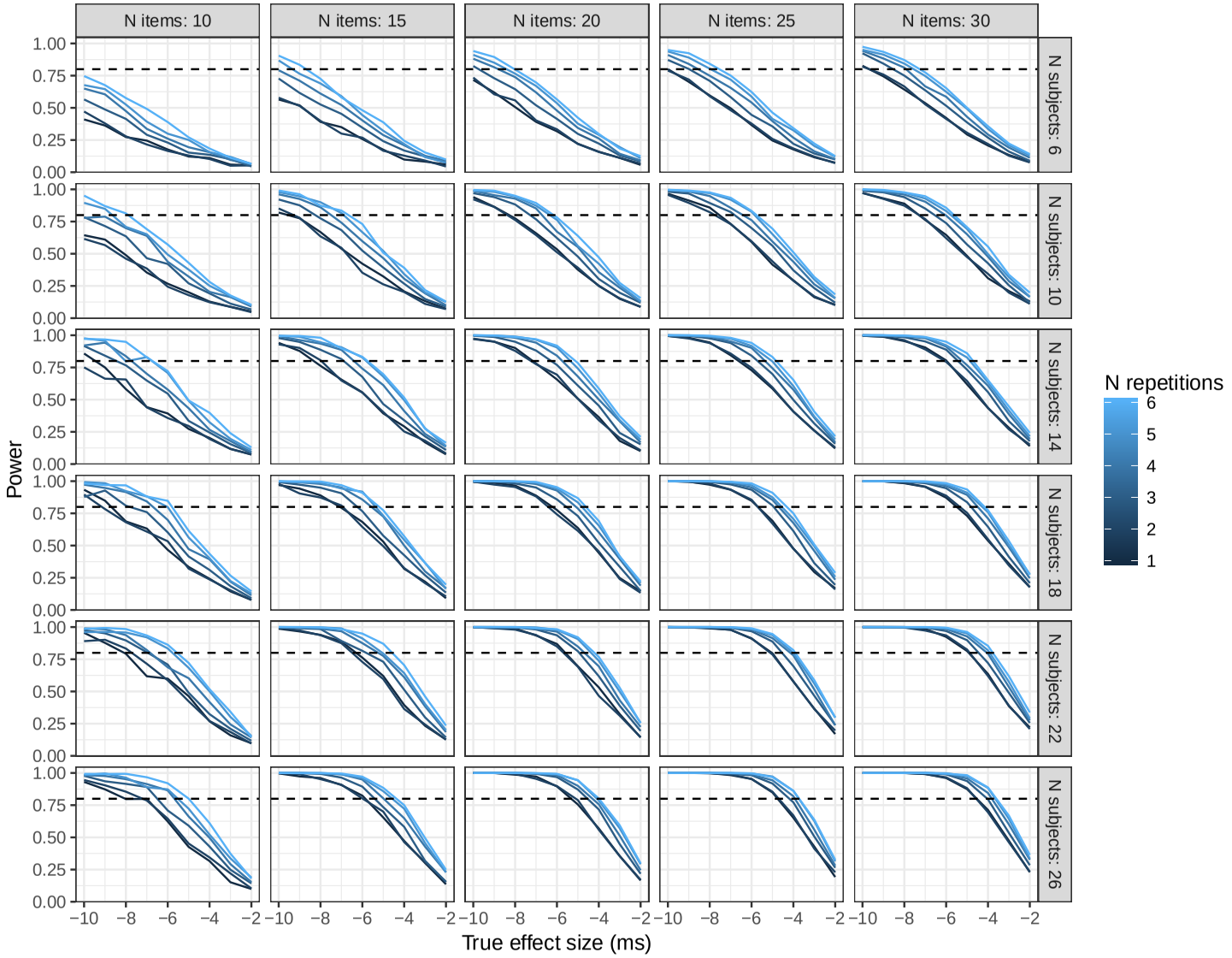

Figure 8: Power curves simulated from the Roettger et al. (2014) data $n_{sims} = 1500$. Rows show simulated number of subjects; columns show number of repetitions of each item per subject. Darker colored lines are for runs with fewer items, lighter colored lines for runs with more items. The dashed line indicates power of 0.8.

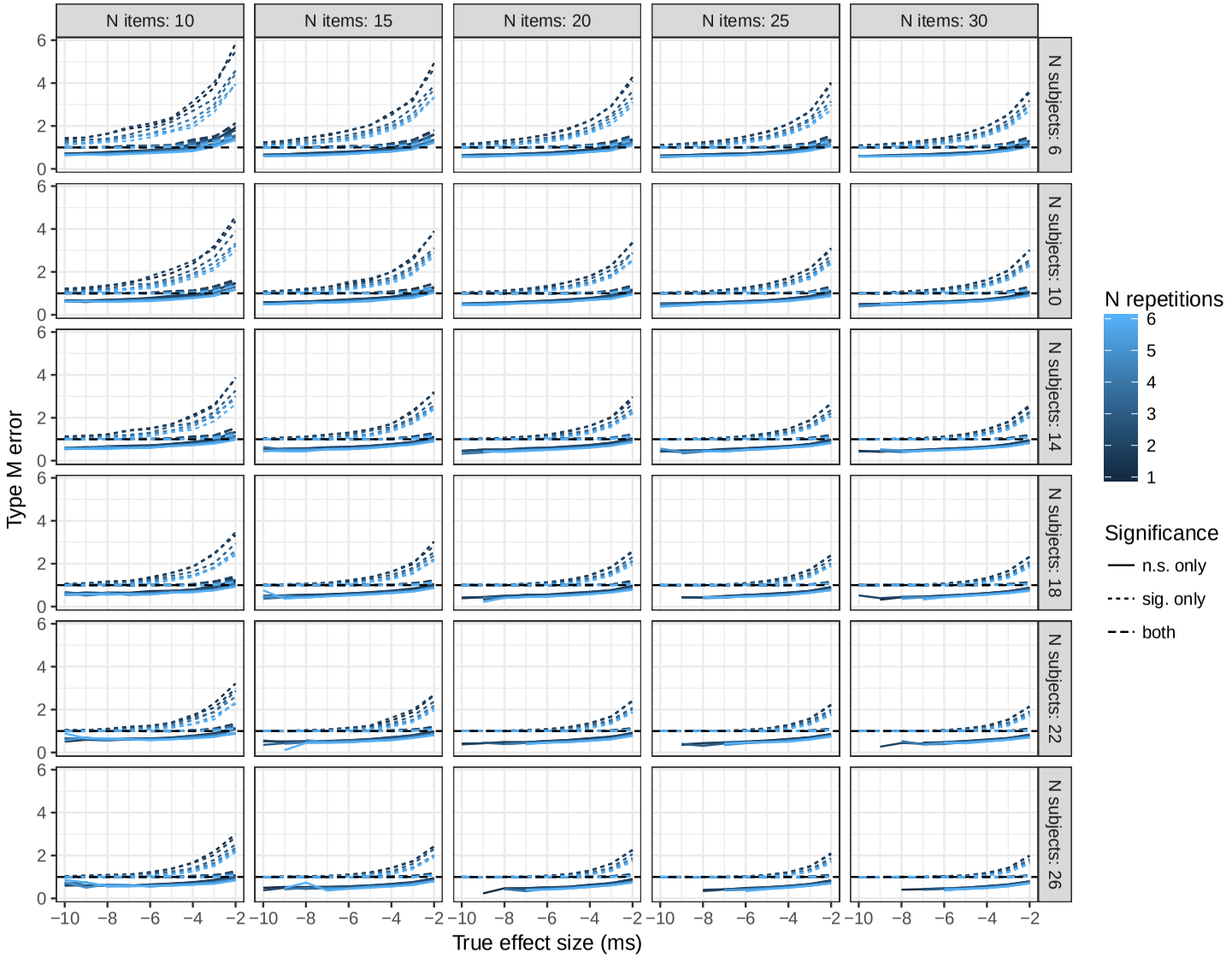

Figure 9: Type M error curves simulated from the Roettger et al. (2014) data ($n_{sims} = 1500$). Rows show number of subjects and columns show number of items, while line color shows the number of repetitions of each item per subject. Different line types show the expectation of the absolute value of the estimated effect size divided by the true effect size for runs statistically significantly different from zero, not significantly different from zero, and across all runs. The horizontal line at 1 shows optimal Type M error.

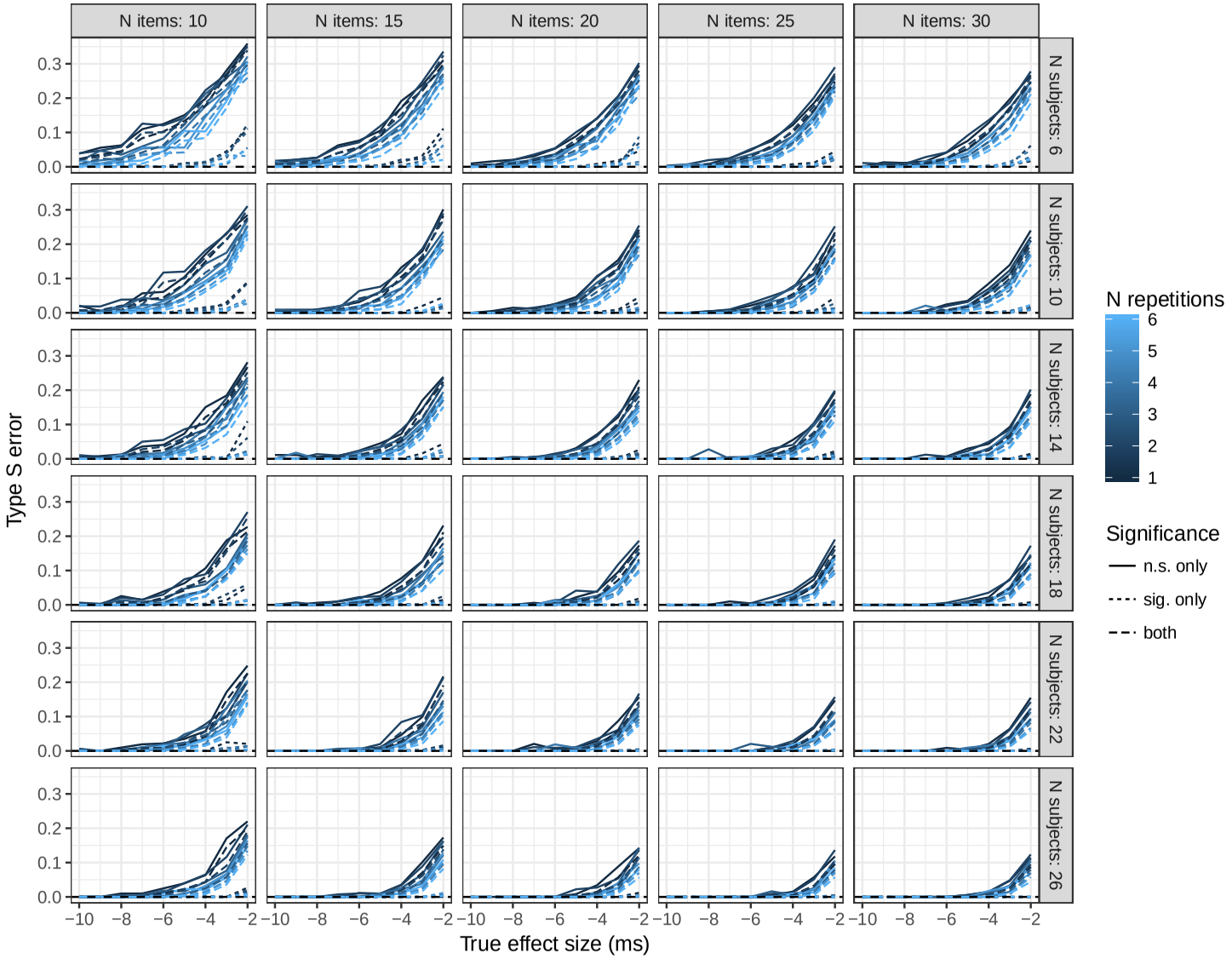

Figure 10: Type S error curves simulated from the Roettger et al. (2014) data ($n_{sims} = 1500$). Rows show number of subjects and columns show number of items, while line color shows the number of repetitions of each item per subject. Different linetypes show the proportion of cases where the signs of the estimated and true effect size differ, for runs statistically significantly different from zero, not significantly different from zero, and across all runs. The horizontal line at 0 shows minimal Type S error.