Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German

Bruno Nicenboim
University of Potsdam

Timo B. Roettger
University of Cologne

Shravan Vasishth
University of Potsdam

DRAFT

Abstract

It is common practice in statistical analyses of phonetic data to draw conclusions based on statistical significance alone. Using *incomplete neutralization* of final devoicing in German as a case study, we illustrate the problems with this approach. If researchers find a significant acoustic difference between voiceless and devoiced obstruents, they conclude that neutralization is incomplete; and if they find no significant difference, they conclude that neutralization is complete. However, such strong claims regarding the existence or absence of an effect based on significant results alone can be misleading. Instead, the totality of evidence should be brought to bear on the question. Towards this end, we synthesize the evidence from 14 studies on incomplete neutralization in German using a Bayesian random-effects meta-analysis. Our meta-analysis provides evidence in favor of incomplete neutralization. We conclude with some suggestions for improving the quality of future research on phonetic phenomena: ensure that sample sizes allow for high-precision estimates of the effect; avoid the temptation to deploy researcher degrees of freedom when analyzing data; focus on estimates of the parameter of interest and the uncertainty about that parameter; attempt to replicate effects found; and seek to make both the data and analysis available publicly.

1 Introduction

Theories of speech communication and its cognitive underpinnings are increasingly shaped by experimental data and quantitative analyses. Ideally, our theories progressively grow and change with *accumulating* empirical evidence. The evidence provided by a single study, however, is limited to the method applied and the sample tested. Its results are prone to random statistical fluctuations and its interpretation dependent on methodological and

analytical choices. To assess the evidence that a single study can provide, we need a good understanding of statistical theory and inference. There are several specific aspects of statistical analysis, which despite having received little attention in our field, researchers need to be aware of when carrying out statistical inference. Although not unique to our field, some of these aspects of our analyses are particularly relevant for common empirical practices in phonetics.

Beyond statistical assessments of a single study, we can assess the robustness of a phenomenon by synthesizing evidence across many studies. One technique that allows us to synthesize evidence is the meta-analysis: A meta-analysis is a quantitative summary of the results of multiple studies. Here, we apply this technique to a representative phenomenon from the speech production literature which has already fueled fruitful discussions surrounding methodological and analytical practices in phonetics in the past: incomplete neutralization of final devoicing.

1.1 Final devoicing and incomplete neutralization

Final devoicing is a common phonological alternation in the world's languages. For example, languages such as Catalan, Dutch, Polish, Russian, Turkish, and most prominently, German contrast voiced obstruents intervocalically but neutralize the contrast syllable or word finally in favor of voiceless obstruents (cf. 1-2).

(1) Rad [ʁaːt] 'wheel'; Räder [ʁɛːdɐ] 'wheels'

(2) Rat [ʁaːt] 'council'; Räte [ʁɛːtə] 'councils'

In intervocalic position, the voicing contrast of oral stops can be manifested by different acoustic dimensions, such as the preceding vowel duration, glottal pulsing during the closure, closure duration, and voice onset time (inter alia Lisker, 1986), with voiced stops exhibiting longer preceding vowels, more glottal pulsing during the closure, a shorter closure duration, and shorter (or negative) voice onset time. The term neutralization has implicitly implied that the acoustic form of the alveolar stop in Rad [ʁaːt] 'wheel' is identical to the alveolar stop in Rat [ʁaːt] 'council', resonating with ear-phonetic assessments of traditional linguistic descriptions (Jespersen, 1920; Trubetzkoy, 1939; Wiese, 1996). However, numerous experimental studies have argued that there are small acoustic and/or articulatory differences between words such as Rad and Rat, suggesting that in German this neutralization is in fact incomplete (Dinnsen & Garcia-Zamor, 1971; Taylor, 1975; Mitleb, 1981; Port & O'Dell, 1985; Charles-Luce, 1985; Port & Crawford, 1989; Greisbach, 2001; Fuchs, 2005; Smith, Hayes-Harb, Bruss, & Harker, 2009; Grawunder, 2014; Roettger, Winter, Grawunder, Kirby, & Grice, 2014). Importantly, the direction of the difference resembles the non-neutralized contrast, for example, vowels preceding voiceless stops tend to be shorter than vowels preceding devoiced stops. The magnitude of the difference, however, is

For partial support of this research, we thank the Volkswagen Foundation through grant 89 953 to Shravan Vasishth and the Deutsche Forschungsgemeinschaft for two grants: VA 482/8-1 to Shravan Vasishth and Frank Roesler, which funded Bruno Nicenboim, and the Q project, PIs Shravan Vasishth and Ralf Engbert, in the Sonderforschungsbereich 1287, Limits of Variability in Language. We thank Susanne Fuchs and Sven Grawunder for sharing their data with us and allowing us to make them available. We also thank Paul Bürkner for the support on the R package *brms*. The complete code, data, and supplementary material for this paper are available at https://osf.io/g5ndw/

much smaller. For example, Port and Crawford (1989) report a vowel duration difference of approximately 1–6 ms between devoiced and voiceless stops in German, while Warner, Jongman, Sereno, and Kemps (2004) report a difference of 3.5 ms in Dutch (in comparison to substantially larger vowel duration differences found in non-neutralized contexts in German ranging from 24–41 ms; see Mitleb, 1981; Fuchs, 2005; Roettger et al., 2014). Beyond subtle differences in production, these acoustic differences can be perceptually recovered by listeners with above-chance accuracy (e.g., Port & O'Dell, 1985; Port & Crawford, 1989; Kleber, John, & Harrington, 2010; Roettger et al., 2014).

Many scholars have acknowledged the evidence for incomplete neutralization and proposed several ways to implement this phenomenon in formal models of phonological representations (e.g., Dinnsen & Charles-Luce, 1984; Charles-Luce, 1985; Port & O'Dell, 1985; Van Oostendorp, 2008). These formal accounts challenged several assumptions of contemporary phonological models, leading Port and Crawford (1989, pp. 10–15) to claim that incomplete neutralization poses "a threat to phonological theory" (see also Port & Leary, 2005). More recent accounts to incomplete neutralization are rooted in psycholinguistic models of lexical organization, suggesting that incomplete neutralization is an artifact of lexical co-activation (Ernestus & Baayen, 2006; Kleber et al., 2010; Winter & Roettger, 2011; Roettger et al., 2014).

Others scholars have remained skeptical regarding incomplete neutralization, crucially fueled by a few studies that did not find evidence for it (Fourakis & Iverson, 1984; Inozuka, 1991; Jessen & Ringen, 2002; Piroth & Janker, 2004). Studies on incomplete neutralization have also attracted serious criticism on methodological grounds (Manaster-Ramer, 1996; Kohler, 2012; Roettger et al., 2014), leading some researchers to disregard it as a methodological artifact (e.g., Kohler, 2007, 2012). For example, it has been argued that incomplete neutralization is an orthographically induced contrast, where speakers are thought to perform an "artificial" hypercorrection based on the written language (e.g., Fourakis & Iverson, 1984; Manaster-Ramer, 1996). This concern has been tackled by more recent studies, showing that incomplete neutralization is also obtained when participants do not encounter orthographic input (e.g., Roettger et al., 2014).

It has also been argued that early studies on incomplete neutralization have recorded German-speaking populations with high proficiency in English, which is a potential problem because English preserves the final voicing contrast (e.g., bad vs. bat, bed vs. bet) (Kohler, 2007; Winter & Roettger, 2011). However, many later studies used German speakers living in Germany and report similar effect sizes (Grawunder, 2014; Roettger et al., 2014).

It is safe to say that incomplete neutralization is a polarizing phonetic phenomena. One camp of scholars interpret the available evidence in favor of incomplete neutralization, with important implications for models of speech production and linguistic representations, while others interpret the available evidence as either insufficient or pointing towards incomplete neutralization being a methodological artifact. The latter position has led to productive methodological debates, not only raising awareness for important aspects of experimental design, but also drawing attention to important conceptual issues regarding statistical inference beyond the observed data.

Incomplete neutralization is a prime example to discuss statistical misinterpretations due to several reasons. First, incomplete neutralization effects have been reported to be rather small, making an accurate estimate of the effect particularly important for scientific

conclusions. Second, incomplete neutralization studies commonly use multiple acoustic and/or articulatory measures to test one alternative hypothesis such as devoiced stops are different from voiceless stops and the results from statistical tests are not corrected for multiple comparisons (using, for example, the Bonferroni correction). And third, the incomplete neutralization literature has a history of publishing null results, which led to several (conceptual) replication attempts.

All in all, the literature on incomplete neutralization is a representative area of phonetic research which has already been a fertile source of methodological debates. We aim at continuing this tradition and use incomplete neutralization to discuss important aspects of statistical analyses and misconceptions that need to be taken into account when drawing inferences that go beyond the observed data. It is important to emphasize that incomplete neutralization only serves as a representative example for common practices in phonetic research. Both the misconceptions we discuss and potential strategies to avoid potential analytical pitfalls generalize towards other areas of phonetics as well as empirical sciences in general. We further use the available evidence in the literature to assess the robustness of the phenomenon via a meta-analysis, a powerful statistical procedure for combining data from multiple studies. Our meta-analysis suggests that (i) incomplete neutralization is robust across the available data in the literature, (ii) there is insufficient evidence supporting the claim that previously mentioned potential confounds cause the incomplete neutralization, and (iii) some of the often cited earlier studies did not have sufficient evidence to conclude whether neutralization is or is not complete.

The paper is organized as follows. In Section 2, we discuss common statistical misconceptions related to phonetic research in general, and incomplete neutralization in particular. Next, in Section 3, we motivate the meta-analysis as a way to synthesize empirical evidence. Section 4 describes the selection process and inclusion criteria employed for selecting the studies that were included in the meta-analysis, and discusses how we obtained and distilled the data from the literature, including relevant analytical decisions. Also presented here is the Bayesian random-effects meta-analysis used to synthesize the evidence from the available data. In Section 5, we present the results of our analysis and discuss potential caveats. Finally, in Section 6 we relate our findings to both the above discussed misconceptions in interpreting statistical inference related to common practices in phonetic research. We further use our findings as a motivation for proposing suggestions for future phonetic research.

2 Common statistical misconceptions

In the incomplete neutralization literature (as in many other areas), conclusions regarding the existence or absence of the effect have been drawn depending on the results being statistically significant or not, that is, whether p-values were lower or not than a threshold (i.e., the α value), which is traditionally set at 0.05.

Strong claims regarding the existence or absence of an effect based on significant results alone are misleading on several grounds. First, p-values are often misinterpreted (among others, Lecoutre, Poitevineau, & Lecoutre, 2003) leading to several misconceptions regarding what a p-value can and cannot tell us (Vasishth & Nicenboim, 2016). Second, a significant p-value at the conventional Type I error rate (i.e., the probability of incorrectly rejecting the null when it is true) of 5% may not be a convincing rejection of the null hy-

pothesis. This is because the probability of an incorrect rejection of a true null hypothesis (a "false positive") is often inflated due to incorrect practices that we detail below. Third, non-significant p-values may not be informative regarding the absence of an effect. The experimental phonetic literature shows sample sizes (which are a function of the total number of participants, items, and repetitions that are analyzed in a model) and experimental effects that are often very small. This often leads to a large Type II error rate (i.e., the probability of incorrectly failing to reject the null), making it difficult to know whether a non-significant result is due to the true absence of an effect or due to low power. Finally, statistically significant results from low-powered experiments are guaranteed to yield overestimates of effects; this can lead to overconfident beliefs about replicability (Vasishth, Mertzen, Jäger, & Gelman, 2018).

In this section, we point out common misinterpretations of significant and non-significant results in the context of phonetics in general, and the incomplete neutralization literature in particular. The problems we discuss are rooted in some misunderstandings about what we are allowed to do and infer under the null hypothesis testing (NHST) framework (i.e., the use of p-values) –the most common use of Neyman-Pearson frequentist statistics– which is standardly used in linguistics and psychological sciences. Although none of our observations are novel (for a book-length treatment, see Chambers, 2017), it is important to discuss them within the specific context of experimental phonetics.

2.1 Common problems with significant findings

2.1.1 Misinterpretations of statistically significant p-values. The way that p-values are used in fields like phonetics, psycholinguistics, and psychology is that when the p-value falls below a specific threshold (usually 0.05), we reject a null hypothesis (typically, the hypothesis that there is no effect). Often, if the p-value is greater than 0.05, we end up "accepting" the null as true. Both these conclusions are problematic.

Strong claims, e.g., about the existence of incomplete neutralization, that are based on a significant result are an incorrect use of the frequentist framework. A p-value below 0.05 (a "significant" result) only allows us to reject the null hypothesis (here, that the neutralization of the final voicing contrast is phonetically complete) and does not furnish any information about the specific favored alternative. This is because rejecting a null hypothesis that a parameter (i.e., an unknown value that needs to be estimated, in this case the difference in vowel duration) is zero leaves open all possible non-zero values as candidates for such a parameter. Furthermore, no absolute certainty is afforded by the p-value from a single experiment, no matter how low it is. This is because a p-value is uniformly distributed when the null hypothesis is in fact true. That is, if there truly is no effect (i.e, the null hypothesis is true), any p-value between 0 and 1 is equally likely to be the result of a statistical test (with 5% of the p-values being under 0.05, 10% of the p-values being under 0.1, and so forth). Based on a single p-value that is less than 0.05 (no matter how low it is), it is impossible to distinguish between two possible scenarios: (a) the null hypothesis is false and that is why we obtained a low p-value, or (b) the null hypothesis is true and we happened to get a low p-value by chance. The statement from the American Statistical Association (Wasserstein & Lazar, 2016) provides a detailed discussion on several widely agreed upon principles underlying the proper use and interpretation of the p-value,

among them, the statement is clear in that "by itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis".

2.1.2 Low power increases Type S and M errors. Published studies in linguistics and related areas often have very low statistical power, i.e., low probability that the statistical test will reject a false null hypothesis (equivalently, high Type II error; power is 1-Type II error). For example, in their recent review on sentence processing, Jäger, Engelmann, and Vasishth (2017) show (their Appendix B) that for typical sample sizes in reading time studies, power may be as low as 6-20%. That is, if there is a true effect, these studies have a 6-20% chance of finding it due to too small sample sizes. Similarly, Kirby and Sonderegger (this issue) report simulation studies showing that incomplete neutralization studies using six speakers have a power of approximately 6-50%. While typical subject numbers differ across subdomains of phonetics, six participants is not an uncommon sample size in phonetic experiments: Within the incomplete neutralization literature on German, Fuchs (2005) had three speakers, Fourakis and Iverson (1984) had four speakers, Charles-Luce (1985) and Port and Crawford (1989) had five speakers, and Piroth and Janker (2004) had six speakers.

One might think that the only implication is that many non-significant and inconclusive results will be found. However, as Gelman and Carlin (2014) point out, another surprising consequence of low power is that significant results will have exaggerated effects. Some examples from published data are discussed in Jäger et al. (2017); studies with very low power can have effects that may be as much as 5-7 times larger than the true effect; another set of examples is discussed in Vasishth et al. (2018). These errors of overestimation are called Type M(agnitude) errors.

In the context of phonetics, consider Port and Crawford (1989). They report a vowel duration difference of approximately 1–6 ms. If, for the sake of argument, incomplete neutralization was real and the true effect size was around 1–6 ms, low powered studies would lead to extremely exaggerated effects of over 20 ms. Mitleb (1981) reports on a vowel duration difference of 23 ms. Fuchs (2005, figure 4.29, page 142) reports on a vowel duration difference of around 30 ms. These numerically large effects could be accurate if power were high; but they could simply be due to Type M error being high. As in any other empirical science, Type M errors are relevant for phonetic research. The magnitude of an acoustic effect has direct implications for interpreting its potential practical relevance. The human ear has certain thresholds of what constitutes a least-perceptible difference (e.g., Huggins, 1972). If an acoustic effect is observed, it might be perceivable or not depending on its magnitude. In fact, Kohler (2012) has argued that incomplete neutralization effects commonly reported on in the literature cannot have any perceptual relevance and should thus be discarded as a genuine phonological phenomenon (see also Roettger et al., 2014). This is in line with often cited just noticeable differences for vowel duration range between 10 and 25 ms (e.g., Klatt, 1976).

A second bad consequence of low power is Type S(ign) error; because the magnitude can be exaggerated in low-power settings, the sign of the effect can also flip. If the true effect is positive in sign, a low power experiment may well find an effect that is negative in sign. Thus, even a study that exhibits effect sizes pointing in the opposite direction, i.e., longer vowels preceding voiceless stops, is not entirely surprising if it is underpowered, and thus should not be overinterpreted. For a more detailed discussion of power, Type S, and

Type M errors related to experimental phonetics in general and incomplete neutralization in particular, consult Kirby and Sonderegger (this issue).

Coupled with publication bias, i.e., journals tending to favor results which are significant, the field can gradually fall into the collective illusion that an effect is large and robust; because exaggerated effects from significant studies tend to be seen as newsworthy and get published, we would see only the overestimated effects and not the unpublished studies that failed to reach the 0.05 threshold with the p-value. This point is discussed further in Vasishth et al. (2018).

2.1.3 Inflation of Type I error. Moreover, recent replication attempts in different disciplines (e.g., Open Science Collaboration, 2015; Begley & Ioannidis, 2015) show that the false positive rate (Type I error rate) may be much higher than 5%. We discuss here two main problematic practices that are particularly relevant to the analysis of phonetic data: (i) issues with the way the data are (un)aggregated for analysis, and (ii) the multiple comparisons problem (for a general discussion of problematic practices in linguistics and psycholinguistics, see Vasishth & Nicenboim, 2016).

The first practice that inflates the number of significant results has to do with the way that phonetic data are sometimes pooled. One problem arises when unaggregated data is analyzed with methods such as ANOVA and t-tests without paying attention to the assumptions underlying these tests. This problem, also known as pseudoreplication, arises because multiple samples from one participant or item are treated erroneously as independent data points in the statistical analyses (Hurlbert, 1984). To illustrate this, imagine that in an experiment, four participants read aloud ten words ending with 'd' and ten words ending with 't'. Thus the forty elicited words of each condition are not independent samples, since we expect commonalities between the words produced by each speaker. If we ignore this, and we compare the forty words in the 'd' condition with the forty words in the 't' condition using, for example, a t-test, we will artificially inflate the degrees of freedom of the statistical test to 78 (informally, this is the number of values in the final calculation of a statistic that are free to vary). This will in turn lead to an artificial decrease in the variance of the estimates (i.e., the estimates will seem artificially precise) and thus to incorrect significant results (for more examples, see Winter, 2011). Interestingly, this problem has already been pointed out by Charles-Luce (1985, p. 318), who notes that earlier studies exhibited inflated degrees of freedom. However, pseudoreplications are a problem in many recent studies as well (Greisbach, 2001; Fuchs, 2005; Piroth & Janker, 2004). For example, Piroth and Janker (2004) present an experiment with six speakers, but the degrees of freedom (≈ 1400) are greatly inflated. This problem seems to be prevalent in the analysis of phonetic data (Winter, 2011; but not only, see also: Freeberg & Lucas, 2009; Lazic, 2010). Simulations show that in some situations, this can inflate the Type I error to almost 40% (Winter, 2011).

Aggregating data by participants and by items and doing separate analysis for participants and items solves the problem of pseudoreplication. However, this also reduces the sources of variance (through aggregation). For example, Vasishth, Chen, Li, and Guo (2013) discuss the re-analysis of a published paper where by-participants and by-items F-scores from a repeated measures ANOVA showed significant effects, while a linear mixed model on unaggregated data, simultaneously taking both sources of variance into account, failed to do so. Analyses on aggregated data are especially problematic when a p-value is

below 0.05 only for the by-participants (or the by-items) analysis, and this is reported and used to argue for a significant result. In addition, Vasishth, Beckman, Nicenboim, Li, and Kong (this issue) show how aggregating voice onset time (VOT) and vowel durations (as a proxy for speech rate) shows a strong effect of vowel duration on VOT. However, this changes once one takes the uncertainty of the means into account.

In addition, for the aggregation by subjects, there is a conceptual problem: the lack of generalizability over items. While it is common practice to draw inferences about a speaker/listener population from a sample (that is, to infer about the totality of the speakers and listeners based on the subset that participated in an experiment), it is less common to draw inferences about the speech material. A claim such as "the final devoicing contrast of German is incomplete" needs to be based not only on participant-based analyses (e.g., aggregated over all stimuli), but also on items-based analyses (e.g., aggregated over all participants; see Clark, 1973). Incomplete neutralization assertions are claims not only about a population of speakers, but also about the language they speak, thus about a population of linguistic items in the lexicon.

The second reason for an inflation in the number of significant results is the multiple comparisons problem. It is not uncommon to fit statistical models for several acoustic measures. Without a statistical correction such as the Bonferroni correction, this practice increases the chances of finding a false positive. If n independent comparisons are performed, the false positive rate would be $1-(1-.05)^n$ instead of 0.05; four comparisons, for example, will produce a false positive rate of approximately 19%. If we want to keep the false positive at 5%, we should use the Bonferroni correction which implies testing each individual hypothesis at a significance level of 0.0125 (.05 divided by four) instead of .05.

Multiple testing problems surface in most studies on incomplete neutralization (and phonetics in general) because in these studies, multiple tests are conducted for multiple different dependent measures. In fact, except for Roettger et al. (2014), all studies on incomplete neutralization in German have tested several acoustic measures and did not correct for this type of multiple testing.

One might argue that corrections for multiple testing are not reasonable for phonetic studies on the grounds that acoustic / articulatory measures that are used to study speech phenomenon are often correlated, and so corrections such as the Bonferroni correction might be too conservative. However, as von der Malsburg and Angele (2017) showed, correlated measures in eyetracking (reading studies) lead to Type I error inflation that is nearly as high as in independent multiple tests. Thus, a multiple comparisons correction is necessary even with correlated measures in order to obtain the conventional Type I error.

A related problem of multiple comparisons that has received less attention in linguistic research is based on analytical decisions that the researcher faces before he or she presents the statistical significant results. This is generally known as researchers degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) or the garden of forking paths (Gelman & Loken, 2014). Both terms roughly refer to all the decisions regarding the data analysis that researchers face: the choice of the statistical test (t-test, ANOVA, mixed model), which covariates or measures to include, decisions on what constitutes an outlier observation, and even decisions that could have been taken, if the data would have been different (for an example, see Vasishth & Nicenboim, 2016). While fitting many models to a dataset is certainly a component of the data analysis process, the problem arises when researchers choose to

present only the models with statistically significant results (or the ones without) ignoring the alternative analyses. Gelman and Loken (2014) point out that given multiple ways one could analyze the data, once we start looking hard enough, it is almost always possible to find a significant effect. Researcher degrees of freedom can be especially problematic when a seminal paper shows a significant effect that then cannot be replicated (e.g., Vasishth et al., 2018). A failure to replicate may lead to researchers doing new studies on the topic to look hard enough until something is significant and the seminal paper is at least conceptually replicated. This perpetuates the cycle of significant results arrived at through exercising researcher degrees of freedom.

The issue of researcher degrees of freedom is prevalent in phonetic research. For example, several studies on incomplete neutralization have included different co-variates such as the prosodic position of the word (e.g., Charles-Luce, 1985; Port & Crawford, 1989; Jessen & Ringen, 2002; Piroth & Janker, 2004) or the elicitation method (e.g. Port & Crawford, 1989). Moreover, speech production data is prone to a lot of variation: Speakers sometimes mispronounce speech material, produce hesitations, or produce different prosodic realization of the same speech material. They also exhibit variation in their pronunciation of the segments under scrutiny, such as producing a stop with or without a release. What data to include and not to include is up to the researcher and introduces further degrees of freedom, which —no matter how well they are justified— can increase the chance of finding a significant result. For example, Piroth and Janker (2004) excluded all unreleased stops for their entire analysis, although measures such as preceding vowel duration can be reliably measured even without the consonantal release.

2.2 Common problems with non-significant findings

The incomplete neutralization literature is one of the few areas in phonetics that has a rich history of publishing null results. As with significant results, non-significant results are also commonly misinterpreted. A common mistake is to interpret non-significant findings as evidence for the absence of an effect. However, a p-value is a conditional probability: The probability of getting a statistic as extreme or more extreme as the one we obtained, conditional on the null hypothesis being true. A conditional probability is not reversible, and a large p-value does not tell us that there is a large probability of the null being true, *conditional* on the extreme statistic that we obtained.[1] Except in high power experiments (Hoenig & Heisey, 2001), a p-value greater than .05 can only tell us that we failed to reject the null hypothesis. Given the small sample sizes and small effects in the experimental phonetic literature, a likely explanation for non-significant results is low power (i.e., a low probability of correctly rejecting the null).

Studies on incomplete neutralization reporting null results have made their claims based on very small sample sizes (e.g., Fourakis & Iverson, 1984; Inozuka, 1991; Jessen & Ringen, 2002; Piroth & Janker, 2004). Their null results may thus well be due to low statistical power. This would not be the first time this has happened with respect to incomplete neutralization. For Dutch final devoicing, while Baumann (1995) and Jongman,

[1] Dienes (2011) illustrates this with a very colorful example: While the probability of dying conditional on that a shark has bitten one's head clean off is actually one, the reverse is close to zero. Since people are not usually eaten by sharks, given that one is dead the probability that a shark has bitten one's head clean off is very small.

Sereno, Raaijmakers, and Lahiri (1992) failed to find significant incomplete neutralization effects, Warner et al. (2004) did, indeed, find significant effects based on a larger speaker sample. Of course, the above discussion does not imply that there is no way to argue in favor of evidence for the null hypothesis; we come back to this issue in the general discussion section.

This problem of low power is further exacerbated by subsetting the data or performing nested comparisons. Subsetting and analyzing independently the items or participants decreases the sample size (and therefore power) even further. This has been the case for example in Piroth and Janker (2004) and Fuchs (2005). They subsetted the speech material and ran separate comparisons for individual speakers. A similar situation arises if the difference between two means d_1 in one experiment is significant and the difference between two means d_2 in another independent experiment is not significant. One cannot then argue that the difference between d_1 and d_2 is meaningful (i.e., statistically significant) without testing for an interaction. Echoing an example from Gelman and Hill (2007), if $d_1 = 10$ with $SE_1 = 4$, and $d_2 = 5$ with $SE_2 = 10$, the difference between the two comparisons yields a mean difference $d_1 = 10 - 5 = 4$ with a standard error of $\sqrt{SE_1^2 + SE_2^2} = 11$, which is not significant. For more discussion of this point, see Gelman and Stern (2006), and Nieuwenhuis, Forstmann, and Wagenmakers (2011). This can be related to the paper by Fourakis and Iverson (1984), the most often cited study claiming to have shown the null with respect to German voicing neutralization. They ran two different experiments, one of which they interpreted as showing the null, and one of which they interpreted as showing an incomplete neutralization effect comparable to Port and O'Dell (1985). Without showing that there is a significant interaction between experiments and the obtained effect, their comparison is statistically not meaningful.

3 Synthesizing empirical evidence with a meta-analysis

Given the arguments above, a single study, whether providing a significant result or not, cannot tell us much about a phenomenon. Literature reviews are very helpful here, but the conventional approach in linguistics and the psychological sciences involves counting the number of significant and non-significant effects across studies, and using a majority vote approach to making a binary decision as to whether an effect is present or not. For example, Phillips, Wagers, and Lau (2011) take a voting-based approach to summarize the literature on retrieval effects in the processing of reflexives. The evidence is summarized (p. 156) by classifying each published claim into falling into one or the other bin without regard to the magnitude or uncertainty of the estimate, and the majority vote from the literature is taken as the conclusion: "Thus, most evidence suggests that the processing of simple argument reflexives in English is insensitive to structurally inappropriate antecedents, indicating that the parser engages a retrieval process that selectively targets the subject of the current clause."

As an example from phonetics, in the neutralization case, twelve out of the fourteen studies we consider in this paper reported significant results in the original analyses (see Table 1); the conventional approach would be to simply conclude that the effect is therefore present. No attention is paid to the magnitude and uncertainty of the estimate in each study. A study with a 50 ms effect and a standard error of 25 has the same meaning as

a study with a 20 ms effect with a standard error of 5. As mentioned above, low power studies coupled with publication bias may well result in exaggerated effects which may not reflect the truth. Therefore, a more reasonable approach—widely used in medical statistics (Higgins & Green, 2011)—is to derive a quantitative estimate of the effect from available studies. A meta-analysis can allow us to quantitatively summarize the results of multiple studies by estimating the underlying effect of interest from these studies. In essence, each study is weighted by the precision of the estimate; studies with large standard errors play a smaller role in determining the overall effect, and studies with small standard errors have more influence. The overall effect estimated from a meta-analysis is thus analogous to a weighted mean of the individual studies, weighted by their precision.

An interesting aspect of a meta-analysis is that it allows us to take all the relevant quantitative evidence available into account (see the Study selection section). While intuitively it makes sense that a scientific conclusion should be based *quantitatively* on a body of work, meta-analyses are still not common in linguistics and phonetics (but see, for example, Vasishth et al., 2013; Mahowald, James, Futrell, & Gibson, 2016; Jäger et al., 2017).

A meta-analysis, however, can be problematic if it is suspected that a field suffers from publication bias, that is, if only statistically significant results are published (see e.g., Sterling, 1959; Rosenthal, 1979; Fanelli, 2011). As mentioned above, one major adverse consequence of publication bias is that published effects tend to have exaggerated effect sizes that arise from low power studies (or Type M errors; Gelman & Carlin, 2014); studies with smaller (but more realistic) effect sizes may never be published because they are not significant (Hedges, 1984; Ioannidis, 2008). Any meta-analysis that depends on studies with exaggerated effects will of course overestimate the effect (Simonsohn, Nelson, & Simmons, 2014). While there are tools to address the problem of publication bias in meta-analyses (see, for example, Moreno et al., 2009; Simonsohn et al., 2014; McShane, Böckenholt, & Hansen, 2016), the case of the incomplete neutralization literature is one of the few areas in phonetics that has a history of publishing non-significant results. Of course, we do not doubt that publication bias exists here too; it follows that any meta-analysis will yield biased estimates. Despite this problem, the meta-analysis is an improvement over the voting system that is commonly used to decide if an effect is seen in the literature; it sets the focus on the best estimate we have, along with the uncertainty of our estimate. Ignoring the magnitude and uncertainty of the estimate can lead to overoptimistic beliefs about the existence of an effect (Vasishth et al., 2018).

Another practical problem with conducting a meta-analysis is that published studies often fail to report estimates and/or standard errors (or any measure of dispersion), or lack enough information to deduce this information. When these statistics are provided, they are often based on inappropriate statistical analyses. The ideal solution is to analyze the raw data; but these are usually not available.[2] However, as we discuss in the Methods section, for the incomplete neutralization literature, when raw data were not available, in many cases, tables with some type of summaries were provided. As we present in detail later, Bayesian models can be used to reconstruct the plausible values of the individual estimates based on the summaries provided in the papers. Once estimates with their measures of

[2]Websites such as the Open Science Framework (http://osf.io/) are becoming increasingly popular for archiving data (and e-prints). The Journal of Phonetics also strongly encourages authors to deposit data and code with their article submissions —this is not yet standard.

dispersion were obtained, we use a Bayesian random-effects meta-analysis (Sutton, Welton, & Cooper, 2012) to synthesize the evidence for incomplete neutralization.

4 Methods

4.1 Eligibility criteria and study selection

The experiments included in the Bayesian meta-analysis are summarized in Table 1. This list of studies was generated as follows: We first generated a list of potentially relevant items to be included in our meta-analysis using the google scholar search engine, with the search terms 'incomplete neutralization' and 'German'. This search was carried out in June 2017. We inspected the first 100 results. Ten additional items were included based on recommendations and by checking references of included papers. We checked the abstracts of the remaining papers and identified 19 items for full-text inspection according to the following selection criteria (see also the related PRISMA checklist, Liberati et al., 2009, available at https://osf.io/wjpbg/)

We screened the 19 studies and selected fourteen studies based on the following criteria: (i) acoustic correlate, (ii) recoverability of effect, (iii) elicitation and prosodic context, and (iv) the sampled population.

(i) Acoustic correlate. We included all experiments that investigated the acoustic correlates of voicing in syllable-final position in German. Since there are many acoustic correlates that are potentially co-varying with the paradigmatic voicing status of a stop (e.g., Keating, 1984) across different studies, numerous phonetic properties have been found to distinguish voiceless from devoiced stops in domain-final position. These include the duration of the preceding vowel, the closure duration, the duration of the "voicing-into-the-closure", as well as the burst and aspiration durations (among others). Across different studies on German final devoicing, the duration of the preceding vowel has been shown to be the most reliable correlate of obstruent "voicing" in final position and also the acoustic correlate that was most often measured in the incomplete neutralization literature. Thus, in the present study we shall focus on this acoustic parameter. We look at preceding vowel duration for final stops only, excluding measurements of vowel duration preceding fricatives, because only a subset of studies have looked at acoustic correlates of final devoicing in fricatives. Note that one study (Piroth & Janker, 2004) included in our meta-analysis did not allow us to separate vowel measurements preceding stops and fricatives because data are presented as pooled. Sometimes, vowel duration was measured in combination with other segments (the onset or parts of the rhyme). Given the assumption that other segments are not systematically co-varying with voicing, we make the simplifying assumption that this inclusion does not confound the analysis. Applying the above criteria led us to exclude two studies that did not measure preceding vowel duration (Taylor, 1975; Jessen & Ringen, 2002).

(ii) Recoverability of effect. We included all speech production experiments that measured the acoustic dimension specified above and provided sufficient information to recover at least an estimate of the effect (vowel duration difference between devoiced and voiceless stops) and a measure of dispersion (e.g., standard error). Some studies that examined incomplete neutralization using pre-stop vowel duration were excluded because they did not provide enough information for an extraction of these estimates. These are

Dinnsen and Garcia-Zamor (1971), Inozuka (1991), Piroth, Schiefer, Janker, and Johne (1991). For the details about the calculation of the estimates from the published studies, see Section 4.3 and the online supplementary material (https://osf.io/3qmf5/).

(iii) Elicitation and prosodic context. We included all speech production experiments that measured the acoustic dimension specified above, excluding speech perception experiments on the perceptual recovery of investigated effects. Within these criteria, we included production experiments that used different elicitation tasks ranging from reading word lists, sentence lists, repeating auditorily presented stimuli, deriving word forms from auditorily presented paradigmatic neighbors, up to dictating contrasting words to the experimenter. Moreover, studies differed regarding the embedding of the target words in their prosodic environment, including words in isolation and words embedded into utterances in phrase-medial or phrase-final position.

(iv) Sampled population. We restricted the review to experiments with linguistically unimpaired, native, adult participants. This included populations living abroad (e.g., students in the United States, Mitleb, 1981; Fourakis & Iverson, 1984; Port & O'Dell, 1985; Smith et al., 2009; as well as German speakers of different dialects, Piroth & Janker, 2004; Fuchs, 2005; Grawunder, 2014).

The final sample consisted of fourteen studies from eight journal papers, three books/theses, and one unpublished report (all the data are available in https://osf.io/4c25h/).

4.2 Analysis

To extract the estimates from each individual study and to run the meta-analysis, we used a Bayesian data-analyis approach implemented in the probabilistic programming language *Stan* (version 2.16.2 Stan Development Team, 2017) using the model wrapper package *brms* (version 2.1.0 Bürkner, 2017) in *R* (version 3.4.0 R Core Team, 2017). The *brms* package allows the specification of models using a formula syntax which is similar to the popular *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). One major reason that Bayesian methods never caught on in the psychological sciences and related areas is that until recently, it was difficult, if not impossible, to fit complex Bayesian models. This was due to the computational difficulties involved; complex Bayesian models use sophisticated sampling algorithms to compute the distributions of the parameters. However, these computational problems have largely been resolved as far as linguistics and psychology are concerned. As a consequence, in the last few years, there has been a strong move towards Bayesian modeling in these and other areas.

The Bayesian approach is quite different in its goals from the Neyman-Pearson frequentist method we standardly use in linguistics and the psychological sciences. The central goal in Bayesian data analysis is to quantify the uncertainty about a particular parameter of interest, given the data. For example, the question about neutralization can be seen as a question about the sign and magnitude of the effect in a particular statistical model. Given a particular data-set, the Bayesian approach provides a distribution of plausible values representing this effect. This information is of much more direct relevance than null hypothesis significance tests, which answers a question that we don't actually want the answer to (is the null false?), and which relies on the imagined (and usually unrealistic) properties of data that we *didn't* collect. Another important motivation for using the Bayesian approach

is that it is easy to fit complex models that reflect the data-generation process more accurately than the canned models commonly used in the frequentist framework. Notice that in order to fit a Bayesian model, we need to specify prior distributions over the different parameters of our models. These distributions express our initial state of knowledge. In all our models, we use regularizing or weakly informative priors. These priors give some minimal amount of information and have the objective of yielding more stable inferences in comparison with maximum likelihood estimation or Bayesian inference with flat ("uninformative") priors (Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2013; Gelman, Jakulin, Pittau, & Su, 2008; Gelman, Simpson, & Betancourt, 2017). Nicenboim and Vasishth (2016) and Vasishth et al. (this issue) discuss the Bayesian approach in detail in the context of linguistic and phonetic research.

As outcomes of the analyses, we summarize the posterior distributions of non-standardized differences in milliseconds in the following way: (i) 95% credible intervals, and (ii) the posterior probability of the estimate being positive given the data ($P(\beta > 0)$). 95% credible intervals demarcate the range within which we can be certain with probability 0.95 that the parameter (which represents here the difference between the means of two conditions) lies, given the data at hand and our model (see, for example, Jaynes & Kempthorne, 1976; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). Posterior probabilities tell us the probability that the parameter has a value greater than zero (given the data and model); note that these probabilities are not frequentist p-values. Note also that there is no notion of Type I or II error in Bayesian statistics because the inference does not depend on hypothetical repetitions of the experiment; the data are evaluated on their own merits, and no supposition is made about the replicability of the effect.

4.3 Estimates of the individual studies

We extracted the posterior distribution of the difference in duration between vowels preceding a (partially) devoiced consonant vs. vowels preceding a voiceless consonant by reanalyzing the data when possible. In Table 1, we present the means, 95% credible intervals, and the posterior probability that the difference between conditions is positive for the studies of the meta-analysis. Notice that the evidence provided by our estimates do not necessarily match the authors' conclusions; see Table 1. The studies that we included in the meta-analysis had different types of analyses (t-tests, ANOVAs, linear mixed models, etc.), and the information they provided was quite variable; we calculated the estimates in the following manner.

For the main effect of vowel length, we always coded the stimuli with a final devoiced consonant (e.g., *Rad*) with 0.5 and the stimuli with a final voiceless consonant (e.g., *Rat*) with -0.5. This means that the estimate of the effect, $\hat{\beta}$, represents the difference between the two conditions. We never subsetted the data of the individual studies, and instead when it was possible, we added random effects for each sub-study or condition: elicitation method, population, material, and prosodic position.

When raw data were available, we used Bayesian linear mixed models with the maximal random effects structure and weakly informative regularizing priors. This was the case for Fuchs (2005), Grawunder (2014), Experiments 1 and 2 of Roettger et al. (2014), and Experiments 1 and 2 of Baer-Henney and Roettger (2017).

When raw data could not be obtained, we used the information provided in the publications.[3] Some studies presented data that were already summarized at some level (some combination of by-items, by-participants and/or by-repetitions); this was the case for Mitleb (1981), Fourakis and Iverson (1984), Charles-Luce (1985), Port and O'Dell (1985), Port and Crawford (1989), Greisbach (2001), and Piroth and Janker (2004). If we would fit linear mixed models directly to the means provided by the summaries, we would ignore the true variability of the responses, and we would thus overestimate our precision of the estimates. However, except for Charles-Luce (1985), all the summaries included not only means but also standard deviations, allowing us to estimate standard errors. In those cases, it was possible to use Bayesian measurement error models to take into account the original variability in the responses. The idea behind this class of models is that instead of fitting our linear mixed model to the observations, we fit it to a distribution of possible values given the means and the standard errors provided. The intuition behind this is that with large standard errors, a large range of observations is plausible and we take into account this by increasing the uncertainty in the final estimate. This means that a "regular" linear mixed model is a special case of a measurement error model, where the standard error is exactly zero (see also Chapter 14 of McElreath, 2015). The models are detailed at the OSF repository (available from https://osf.io/g5ndw/).

In the single case where a summary of the aggregated data was provided without standard deviations (Charles-Luce, 1985), we fitted the aggregated data to a linear mixed model. This means that the posterior distribution of this estimate might be artificially "tight", or in other words, we might be overestimating the certainty over the range of plausible values.

When no data were available (original data or a summary), as was the case for Smith et al. (2009), we used the mean estimate of the differences between conditions provided, and we calculated the standard error from the F-value provided.[4] However, given that the data were aggregated before performing an ANOVA, the standard error might be underestimated. For Smith et al. (2009), we report an approximate 95% credible interval in Table 1; the interval is assumed to be 2 times the standard error.

4.4 Bayesian meta-analysis

The logic of a meta-analysis assumes that there is a unique underlying effect (i.e., a difference between voiceless and devoiced consonants) to be estimated from all the studies. However, it is possible to add random effects to a meta-analysis. This assumes that there might be heterogeneity in the different studies, and allows for each individual study to be adjusted based on its observed effect (in this case, the posterior distributions of each study).

[3] The second author extracted the data from the publications and the first author checked the extracted data. We contacted three authors for further information, of which two (Fuchs, 2005; Grawunder, 2014) were able to share their raw data with us.

[4] This can be calculated in the following way. Assuming that the sample mean is $\bar{\mu}$, and the reported F-score is F, the t-score can be computed by taking the square root of F, because $t^2 = F$. Then, we simply solve for SE using the equation:

$$t = \frac{\bar{\mu} - 0}{SE} \Longrightarrow SE = \frac{\bar{\mu}}{t}$$

study	conclusion	vowel dur.	$\hat{\beta}$ (ms)	95% CrI	$P(\beta > 0)$
Mitleb (1981)	✓	*	12	$[-42, 60]$	0.76
Fourakis & Iverson (1984)	✗	*/-	12	$[-73, 94]$	0.69
Port & O'Dell (1985)	✓	*	18	$[3, 33]$	0.99
Charles-Luce (1985)	✓	-	-3	$[-88, 82]$	0.45
Port & Crawford (1989)	✓	-	3	$[-82, 90]$	0.53
Greisbach (2001)	✓	-	1	$[-116, 117]$	0.51
Piroth & Janker (2004)	✗	-	8	$[-16, 34]$	0.88
Fuchs (2005)	✓	*	32	$[-14, 66]$	0.96
Smith et al. (2009)	✓	*	13	$[1, 25]$†	-
Roettger et al. (2014) Exp 1	✓	*	9	$[4, 13]$	≈ 1
Roettger et al. (2014) Exp 2	✓	*	6	$[3, 9]$	≈ 1
Grawunder (2014)	✓	*	18	$[13, 23]$	≈ 1
Baer-Henney & Roettger (2017) Exp 1	✓	*	8	$[5, 10]$	≈ 1
Baer-Henney & Roettger (2017) Exp 2	✓	*	9	$[6, 12]$	≈ 1

Table 1

Summary of the studies of the meta-analysis. The column conclusion *indicates whether the authors concluded that there is incomplete neutralization (✓) or not (✗) based on a significant result in any phonetic measure; notice that this is based on the original analysis. The column* vowel dur. *indicates whether the authors found a significant difference (*) or not (-) in vowel duration; */- indicates that one experiment yielded a significant difference and the second one did not, see the discussion section. The symbol β refers to the effect of interest, namely, the difference in vowel duration between the devoiced and voiceless consonants (e.g.,* Rad *vs* Rat*). Table shows the mean of the posterior distribution $\hat{\beta}$, 95% credible intervals, and the posterior probability of the effect being positive (i.e., a positive difference between the vowel duration). †: This is a confidence rather than a credible interval; see Section 4.3.*

study	location of the participants	method of elicitation
Mitleb (1981)	English-speaking country	reading
Fourakis & Iverson (1984)	English-speaking country	no reading/reading
Charles-Luce (1985)	German-speaking country	reading
Port & O'Dell (1985)	English-speaking country	reading
Port & Crawford (1989)	German-speaking country	reading/no reading
Greisbach (2001)	German-speaking country	reading
Piroth & Janker (2004)	German-speaking country	reading
Fuchs (2005)	German-speaking country	reading
Smith et al. (2009)	English-speaking country	reading
Grawunder (2014)	German-speaking country	no reading
Roettger et al. (2014) Exp 1	German-speaking country	no reading
Roettger et al. (2014) Exp 2	German-speaking country	no reading
Baer-Henney & Roettger (2017) Exp 1	German-speaking country	no reading
Baer-Henney & Roettger (2017) Exp 2	German-speaking country	no reading

Table 2
Summary of the studies location of the participants for the different studies, and the method(s) of elicitation used.

Such random-effects meta-analyses can be fit in a frequentist framework too. However, we fit a Bayesian meta-analysis because of the many advantages it affords over a frequentist one. First, the overall estimate of the effect and its uncertainty interval has a clear and intuitive interpretation: we can quantify the range over which we are 95% certain that the true value of the parameter lies, given the data and the model. The frequentist confidence interval does not have this interpretation (Morey et al., 2016). Second, due to the fact that Bayesian models involve regularizing priors, even when data are sparse, the model can generate posterior distributions for the parameters of interest. For an example demonstrating a failure of a frequentist model to estimate parameters in a linear mixed model, and the effect of the regularizing prior, see Vasishth et al. (this issue). Finally, posterior distributions allow us to quantify the probability of the parameter of interest being positive or negative, given the data and the model; this is not possible to do in a frequentist framework.

We carried out two different Bayesian random-effects meta-analyses of the studies presented in Table 1. The objective of the first one was to quantify the evidence for (or against) incomplete neutralization. However, given that experiments on incomplete neutralization have been criticized on methodological grounds (see section 1), we did a second exploratory meta-analysis where we added the location of the population (Germany or Austria, coded as -0.5 vs. United States, coded as 0.5) and the elicitation method (reading, coded as 0.5 vs. any other method, coded as -0.5) as covariates; see Table 2. See the OSF repository (https://osf.io/g5ndw/) for the models specification.

5 Results and discussion

5.1 Main results

The first meta-analysis with no covariates shows a very clear effect of incomplete neutralization; $\hat{\beta} = 10$ ms, 95% credible interval $= [6, 16]$, $P(\beta > 0) \approx 1$. Figure 1 shows the

95% credible intervals of the meta-analytic estimate, and of the non-pooled and partially-pooled estimates of the original studies, that is, the 95% credible intervals estimated either without taking into account the other studies or as part of the random-effects meta-analysis. This incomplete neutralization effect is substantially smaller than acoustic effects observed in non-neutralized positions: Mitleb (1981) report 31–41 ms vowel duration differences between voiced and voiceless stops in non-neutralized, contexts; Fuchs (2005) reports 24–41 ms differences; Roettger et al. (2014) report 28 ms.

The second meta-analysis suggests that adding covariates increases the estimate of the main effect only slightly, and it still shows a very clear effect of incomplete neutralization; $\hat{\beta} = 12$ ms, 95% credible interval $= [7, 18]$, $P(\beta > 0) \approx 1$; see Figure 2(a). This analysis shows no evidence for location of the studied population affecting the results and very weak evidence for reading increasing the effect of incomplete neutralization in comparison with non-reading methods. As Figures 2(b) and (c) show, the posterior distributions are very wide. For the location of the studied population affecting the results (a positive estimate indicates longer vowel durations for participants in English speaking countries): $\hat{\beta} = 0$ ms, 95% credible interval $= [-19, 19]$, $P(\beta > 0) \approx 0.5$, and for the elicitation method affecting the results (a positive estimate indicates longer vowel durations due to reading method): $\hat{\beta} = 6$ ms, 95% credible interval $= [-13, 24]$, $P(\beta > 0) \approx 0.73$.

5.2 Account of possible biases

A clear result of the meta-analysis is that it supports incomplete neutralization in German. However, there are several potential concerns with the meta-analysis which we will address below: the meta-analytic estimate might be biased due to (i) potential confounds in the individual studies, (ii) publication bias, or (iii) individual studies that might not be representative.

5.2.1 Potential confounds in the individual studies. It has been argued that acoustic differences are greater in tasks with orthographic input than without orthographic input (Ernestus & Baayen, 2006; Warner et al., 2004; Warner, Good, Jongman, & Sereno, 2006; Kharlamov, 2014) and that hypercorrection based on the written language may be triggering incomplete neutralization (Fourakis & Iverson, 1984). Since some of the studies (or conditions) included in the meta-analysis used reading as a method of elicitation (see Table 2), the meta-analytic estimate might be an artifact of these studies. In addition, it has been argued that incomplete neutralization might be the result of the influence of English in German speakers living in English speaking countries (Kohler, 2007; Winter & Roettger, 2011) and several studies included in the meta-analysis were based on German speakers in English speaking countries (see Table 2). However, we ran a second meta-analysis in which we included method of elicitation and the location of the studied population as covariates, and we found only very weak evidence of incomplete neutralization being affected by them (see Figure 2). In fact, this meta-analysis including the covariates showed a slightly stronger effect of incomplete neutralization.

5.2.2 Publication bias. As we mentioned before, if only studies with significant results are published, we would see only overestimated effects that would bias our meta-analysis. While we have argued that this might not be the case for the incomplete neutralization literature, a look at Table 1 reveals that all but two of the studies in the

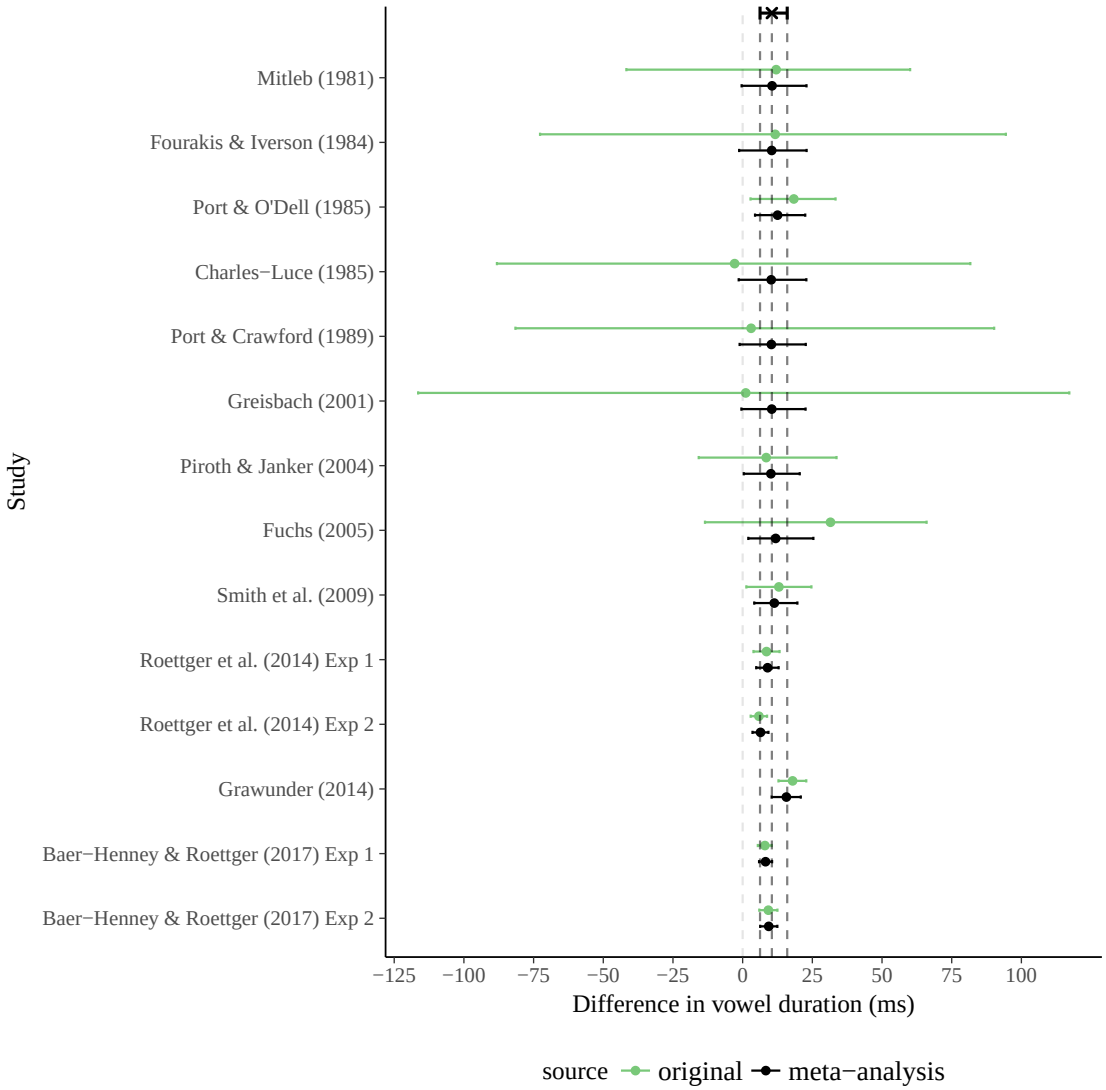

Figure 1. Forest plot of the estimates of the difference in vowel duration; a positive difference indicates evidence for incomplete neutralization. Horizontal lines represent 95% credible intervals. The cross at the top of the plot represents the meta-analytic estimate, green circles are the estimates reconstructed from the original studies, and black circles are the shrinkage estimates of the individual studies delivered by the random-effects meta-analysis.

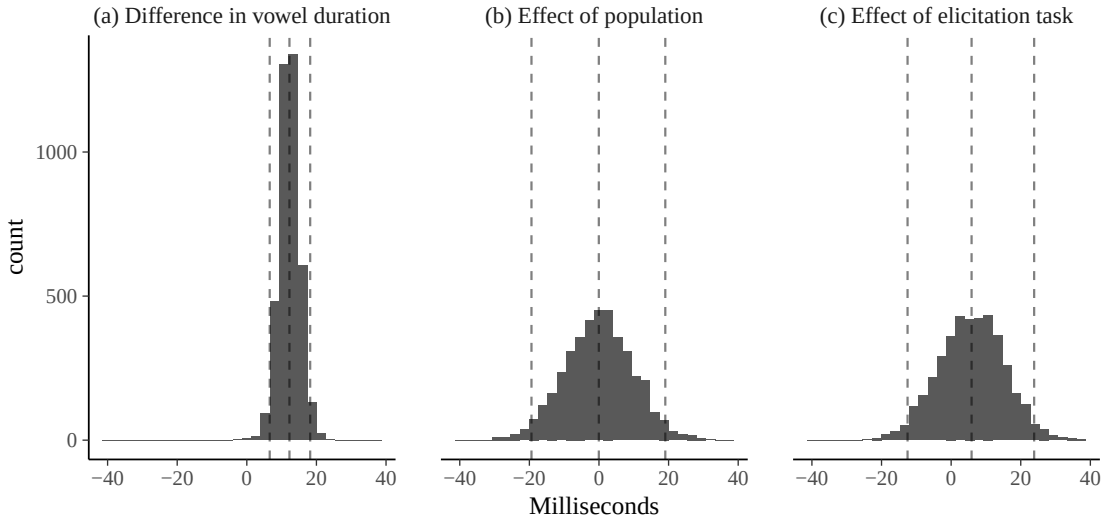

Figure 2. Histograms showing the posterior distributions for (a) the difference in vowel duration in a meta-analysis with covariates, the influence of both (b) the elicitation task and (c) the location of the population on the vowel duration difference. Positive values in the difference in vowel duration indicate evidence for incomplete neutralization, positive values in the covariates indicate evidence for reading increasing the effect of incomplete neutralization in comparison with other elicitation tasks, and for English-speaking countries increasing the effect of incomplete neutralization in comparison with German-speaking countries respectively. The vertical dashed lines indicate the means and 95% credible intervals.

meta-analysis concluded that there is incomplete neutralization based on significant results. However, this is ameliorated by two characteristics of the studies: First, in four of the fourteen studies which reported significant incomplete neutralization effects, there was no significant result for preceding vowel duration. In light of potentially finding incomplete neutralization effects for several different acoustic measures, researchers are more likely to report a null result for one dependent variable when another dependent variable shows a significant effect. Second, in some cases, even when the study argued for incomplete neutralization based on a significant result (in some of the acoustic measures originally examined), the estimates that we re-calculated for the difference in vowel duration do not necessarily match the original conclusion.

In addition, it is possible to examine the extent of publication bias using a graphical approach, namely a funnel plot (Egger, Smith, Schneider, & Minder, 1997; Light & Pillemer, 1984). We plotted the estimates of the individual studies in a funnel plot in Figure 3. This funnel plot shows the precision (inverse of the square of the standard deviation of the posterior distribution or standard error) against the difference between vowel duration observed in each study; a positive difference indicates evidence for incomplete neutralization. Note that low precision entails low power studies, which are shown at the bottom of the precision axis (y-axis), while higher power studies appear higher up. A gap in a funnel plot around the estimates close to zero can be explained by publication bias, especially when the

funnel plot is not symmetric. In the absence of publication bias, we would expect that the estimates of the means would be spread evenly around the meta-analytic estimate, with low power studies showing a larger spread and higher power studies being progressively more clustered near the meta-analytic estimate. While the funnel plot shown in Figure 3 is not completely symmetric (see the next paragraph), it does not seem to show strong indications of publication bias.

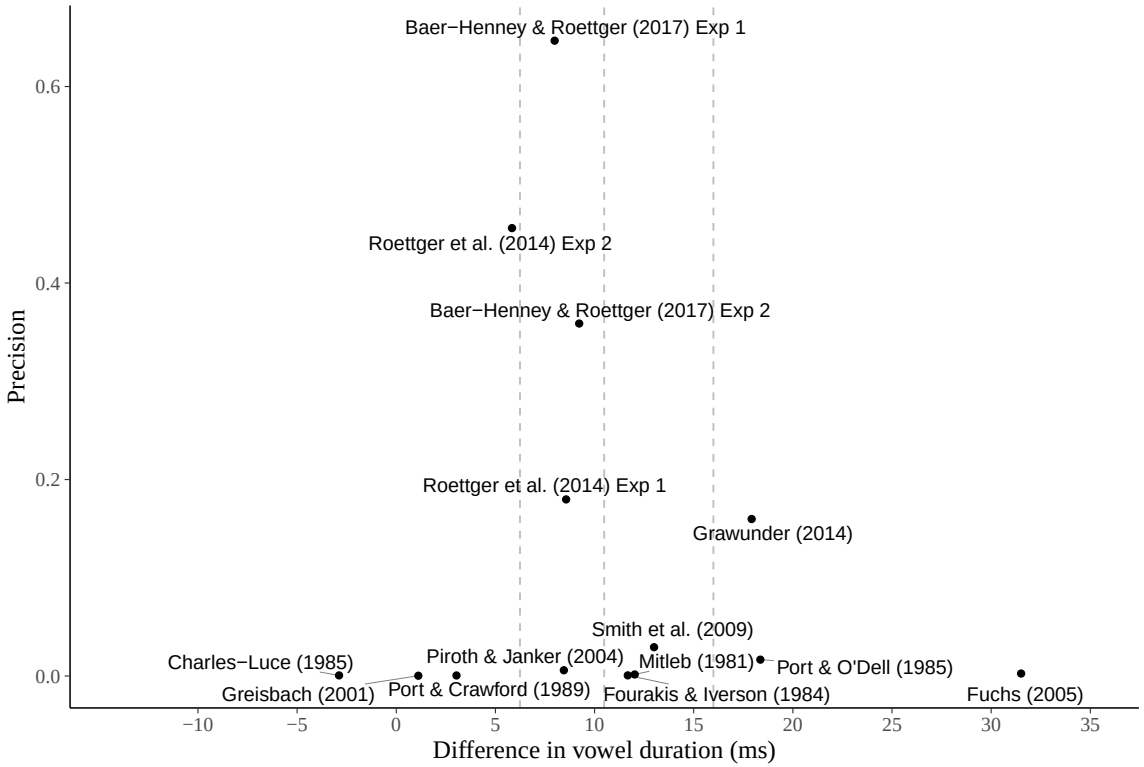

Figure 3. Funnel plot with all the studies included in the meta-analysis. The points represent the difference between vowel duration estimated from the individual studies, a positive difference indicates incomplete neutralization (see Section 4.3). The precision value (y-axis) was calculated as 1/posterior distribution SD^2, except for Smith, Hayes-Harb, Bruss, and Harker (2009) where it was calculated as $1/SE^2$. The vertical dashed lines indicate the meta-analytic estimate ($\hat{\beta}$) and its 95% credible interval.

5.2.3 Individual studies that might not be representative. It is also possible that certain individual studies might have a strong influence in the meta-analytic estimate. The funnel plot in Figure 3 suggests that Fuchs (2005) might be showing an exaggerated effect and biasing the meta-analytic estimate. A meta-analysis excluding this study still provides evidence for incomplete neutralization, with a funnel plot that is more symmetric; Figure 4(a). The magnitude of the meta-analytic estimate remains virtually unchanged; $\hat{\beta} = 10$ ms, 95% credible interval $= [6, 15]$, $P(\beta > 0) \approx 1$, while the original meta-analytic estimate including this study is $\hat{\beta} = 10$ ms, 95% credible interval $= [6, 16]$, $P(\beta > 0) \approx 1$.

A further concern is with Baer-Henney and Roettger (2017), which has not been

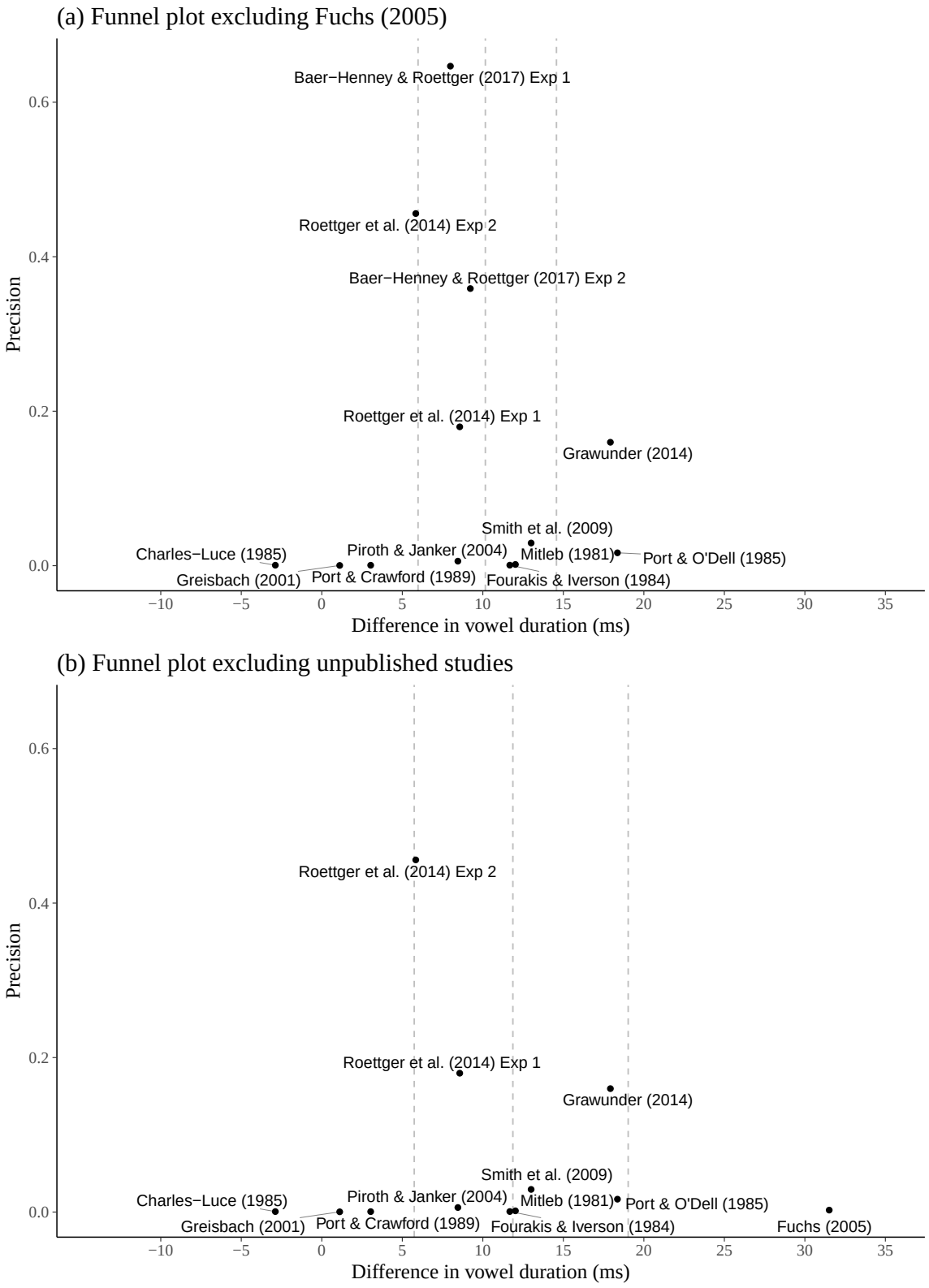

Figure 4. Funnel plots (a) excluding Fuchs (2005), and (b) Baer-Henney and Roettger (2017).

published yet. It could be argued that since the study was not peer-reviewed, the data should not be included. For this reason we also ran another meta-analysis excluding these studies. The new meta-analytic estimate is slightly larger: $\hat{\beta} = 12$ ms, 95% credible interval $= [6, 19]$, $P(\beta > 0) \approx 1$; see funnel plot in Figure 4(b).

6 General discussion

A substantial number of experiments conducted over the last three decades have reported subtle acoustic differences between elements in a phonologically neutralizing context. Although the first seminal papers on this family of phenomena were conducted on final devoicing in German (Mitleb, 1981; Port & O'Dell, 1985), such findings have been advanced for other languages as well. However, the results of many of these studies have been called into question on methodological grounds and there have been several studies that aimed at arguing for the null, i.e., that there is no incomplete neutralization. In this paper, we performed a meta-analysis on fourteen studies on German final devoicing in order to quantitatively synthesize the evidence for incomplete neutralization. Focusing on the vowel duration preceding the obstruent as a cue to voicing, we find an estimated difference of $\hat{\beta} = 10$ ms, 95% credible interval $= [6, 16]$ between vowels preceding devoiced stops and vowels preceding voiceless stops. Our analysis suggests that, given the available evidence, neutralization of German final stops is incomplete.

While the meta-analysis suggests that there is evidence in favor of incomplete neutralization, the case is by no means closed. Given that the current meta-analysis was based on only fourteen studies and that the only two covariates we investigated did not seem to have much of an influence on neutralization, future work can still inform new meta-analyses that build on the present one. These new meta-analyses could yield a more precise estimate of the effect of incomplete neutralization and assess how it is influenced by different factors.

Beyond the aim at synthesizing the available evidence for a particular phonetic phenomenon, the present paper has emphasized the importance of meta-analyses for the phonetic sciences (and the sciences in general), a method for accumulating evidence that is rarely used in our field (but see e.g. Maryn, Roy, De Bodt, Van Cauwenberge, & Corthals, 2009; Tsuji & Cristia, 2014). Science is a cumulative enterprise: As we have discussed in the introduction, what we can learn from a single study in isolation is limited. This is not to say, however, that all studies are equally informative regarding the phenomenon under investigation. For example, the estimates based on some of the seminal papers on incomplete neutralization have such a low precision that, taken in isolation, their informativity is very limited regarding the existence or absence of incomplete neutralization. This issue becomes clearer when we consider our reconstructed estimates based on Mitleb (1981), Fourakis and Iverson (1984), Charles-Luce (1985), and Port and Crawford (1989) in Figure 1. The 95% credible intervals cover a large range of values: from large negative to large positive differences in vowel durations. These results are consistent with complete neutralization, incomplete neutralization, and also with reversed incomplete neutralization, i.e., shorter vowel duration for devoiced stops. Given the large range of possible differences, the results are also compatible with implausibly large effects: Based on the possible values, the acoustic difference could even be so large that they should be ear-phonetically assessable. Such an assumption is obviously at odds with both ear-phonetic assessments of traditional

linguistic descriptions (Jespersen, 1920; Trubetzkoy, 1939; Wiese, 1996) and native speaker intuitions.

These reconstructed estimates are very inaccurate for the following reason: Since the original estimates could not be used —they were estimated by either analyzing each item individually or aggregating by items or by participants, and in some cases with pseudoreplication— we had to reconstruct the estimates. This led us to use measurement error models to fully take into account the data available in the summaries provided in the papers. Due to this imprecision in our estimates, our results might be more conservative than if we had the complete datasets. However, even with the data, the situation would not improve much given the small number of observations in these studies. The small number of observations together with the small effect of incomplete neutralization leads to unreliable estimates which appear at the bottom of the funnel plot in Figure 3. Given the low power of these studies and the possibility of Type-M(agnitude) and Type-S(ign) errors (see Kirby & Sonderegger, this issue), it is possible that the results of these studies have only limited informativity. If the original data were available, it may well have been possible to obtain more precise estimates of the effects. Given the high uncertainty of the reconstructed estimates of Mitleb (1981), Fourakis and Iverson (1984), Charles-Luce (1985), and Port and Crawford (1989), removing these studies from the meta-analysis has a only a very small effect on the meta-analytic estimate for the difference in vowel duration. The new estimate, $\hat{\beta} = 10$ ms, 95% credible interval $= [6, 16]$, $P(\beta > 0) \approx 1$, is virtually identical to the estimate that includes all the studies: The small differences between the estimates are after the decimal point.

Figure 1 shows that the situation has improved in the past fifteen years (at least for the incomplete neutralization literature), in the sense that it is possible to get more precise estimates from the individual studies. This is mainly due to larger sample sizes. However, some of the statistical pitfalls such as pseudoreplication, multiple comparisons, and analyses pooling at an inadequate level are still present in many of the current publications. In addition, in some cases, there is not enough information in the papers to assess the quality of the statistical analysis.

Given that a meta-analysis is composed of individual studies, and as researchers we want to maximize what we can learn from the studies we run, we would like to make several suggestions for the design of future studies in the phonetic sciences. We focus on the following: (i) adequate sample size; (ii) account of multiple comparisons (disclosed or not); (iii) adequate analysis (i.e., answering the research question); (iv) replicability, and (v) reproducibility.

(i) Adequate sample size. No matter how sophisticated the statistical analysis that we employ, with a sample that is not large enough, there is not much that can be learned from a single study. In the frequentist framework, a sample that is too small leads to low power, and to Type-S and M errors (see for an extensive discussion Kirby & Sonderegger, this issue); in the Bayesian framework, it leads to posterior distributions that are wide and uninformative. One solution for this problem is to simply increase the sample size by increasing the number of participants, items, and/or repetitions. The amount of variation among participants, items, or repetitions can suggest which is more efficient to increase. As a rule of thumb, participants show more variation than items, and items, in turn, show more variation than repetitions. This suggests that it will be more efficient to increase the

number of participants, then the items, and then the number of repetitions (see also Rouder & Haaf, In press). Increasing the sample size arbitrarily can easily become unnecessarily "expensive". This is a particularly relevant concern for certain phonetic studies. There are many phonetic methods that are logistically complex and/or use invasive techniques such as electromagnetic articulography or laryngoscopy. Data collection and speaker acquisition is costly and very time-consuming. Additionally, some phonetic studies investigate speech phenomena in understudied languages in which the available speaker population might be very limited.

Instead of arbitrarily increasing the sample size, an adequate sample size can be assessed with simulations: First, we define the range of potential effect sizes, which could be based on either a meta-analysis (but notice that this might be an overestimation) or, could be derived from a computational model or from theory. Second, we generate hundreds of fake datasets based on the assumed effect size(s) (and other assumed characteristics that we know from typical experiments: intercept, standard deviation, variation among participants and items). Finally, we fit statistical models (e.g., linear mixed models) to the generated datasets with different potential sample sizes until we achieve either the desired power in a frequentist framework, or the desired precision of the 95% credible interval in a Bayesian framework. For an example of such power analyses for phonetic research, see Kirby and Sonderegger (this issue). An alternative Bayesian approach is to pre-define a desired precision (inverse of the variance) of the estimate of a parameter, and then run the experiment until that precision is reached. For an example implementing this, see Vasishth et al. (2018).

(ii) Account of multiple comparisons. The problem of multiple comparison is relevant both for when the researcher analyzes multiple (acoustic) measures and for when the researcher has several alternatives for the analysis (Simmons et al., 2011; Gelman & Loken, 2014). Regarding the case when the researcher analyzes multiple measures, corrections such as the Bonferroni correction can be used to correct the α-level to a more conservative threshold and counteract the increase of Type I error. Multiple testing problems are very common in phonetic studies in general because, usually, multiple tests are conducted for multiple different acoustic parameters. However, as for example in the case of incomplete neutralization, the research hypothesis is usually globally defined, i.e., any acoustic measure that significantly distinguishes voiceless from devoiced stops should lead to the rejection of the null hypothesis that neutralization of the final voicing contrast is complete. Thus, any additional acoustic measure that is tested increases the probability of finding a spurious significant result. This is a classic example where correction of the α-level is needed. However, such a correction is seldom done in phonetic research. In fact, except for Roettger et al. (2014), all studies on incomplete neutralization have tested several acoustic parameters and none of them corrected for this type of multiple testing.

A less explored solution is to build a single hierarchical model that accounts for the relationship between the acoustic measures (Gelman, Hill, & Yajima, 2012). However, building such a model is not always trivial, since it entails spelling out precisely how the different measures are (or could be) related to each other (e.g., some are biomechanically or mathematically related, others are not).

Regarding researcher degrees of freedom, this is problematic regardless of whether researchers "p-hack", that is try a number of different analyses until they find a significant

result, or they just explore their data. However, since it is not possible to know ahead of time for which measure an effect will appear, what will be the right transformation of the dependent variable, and so forth, each new model is a new comparison that inflates the Type-I error (De Groot, 1956,2014). Several possible solutions are reviewed in Vasishth and Nicenboim (2016); in addition, Simmons et al. (2011) provide some guidelines for both authors and reviewers. When new data can be easily gathered, an attractive solution is to treat studies as exploratory until confirmed with new data (Tukey, 1977; De Groot, 1956,2014). Once an analysis regarding measures, transformations, covariates, outliers, and so forth is decided, a second confirmatory study identical to the first one can be run. This can be done either with a preregistered replication (Nosek, Spies, & Motyl, 2012) or by gathering more data so that the full dataset could be divided into two (e.g., Nicenboim, Vasishth, Engelmann, & Suckow, 2018). We acknowledge that new data cannot always be easily gathered; however, if all data and code associated with a published paper are released, other researchers can evaluate by themselves the robustness of the presented findings. Platforms such as the Open Science Framework (http://osf.io/) can be useful for this purpose.

(iii) Adequate analysis. While we expect that the statistical analysis should be able to answer our research question, this is not always the case. Issues such as pseudoreplication (i.e., treating all the observations as independent), or aggregation either by participants or by items are examples of decisions made by the researcher that lead to invalid conclusions. This is straightforwardly solved by using frequentist or Bayesian (generalized) linear mixed models (Pinheiro & Bates, 2000; Gelman & Hill, 2007), which have become standard tools that can take into account sources of variance from participants and items simultaneously. An orthogonal problem is to try to argue for the absence of an effect using null hypothesis testing (NHST). This is a problem because NHST can only reject the null or fail to do so, but it generally cannot find support for the null. However, both the frequentist and Bayesian frameworks can address this issue. From the frequentist perspective, one can reverse the null and alternative hypothesis with the equivalence testing approach (Stegner, Bostrom, & Greenfield, 1996) and argue for the null hypothesis. From the Bayesian perspective, one can use Bayes factors (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; and see the review in Nicenboim & Vasishth, 2016), or establish a region of practical equivalence (ROPE) around the null value which is assumed to be practically equivalent to the null effect (Kruschke, Aguinis, & Joo, 2012). However, all these methods, frequentist or Bayesian, require the researcher to make a commitment as to the range of values that count as representing the null or the smallest meaningful effect size. In the case of investigating the communicative function of an acoustic difference, one could for example define the range of values representing the null based on the just noticable difference (Huggins, 1972).

(iv) Replicability. A single study in isolation cannot furnish any information about the *replicability* of any novel result we find. While there is value in conceptual replications (i.e., testing the underlying hypothesis of an experiment using different methods), only a direct or "exact" replication (i.e., repeating an experiment using the same methods) can convincingly establish the robustness of our findings. The idea behind a direct replication is very simple: Any researcher should in principle be able to obtain the original result if they repeat the experiment using the same method and materials, provided that power is sufficiently high (see also Simons, 2014). When logistically feasible, we should attempt to

report direct replications of our findings or, better yet, coordinate direct replications with different laboratories. Only direct replications can verify (or falsify) the predictions of our theories.

(v) Reproducibility. It is very important that published results are reproducible. By reproducible we mean that the reader should be able to take the authors' data, and to regenerate the findings reported in the paper. This is important for several reasons. First, the reader can explore aspects of the data that may not have been discussed in the published paper. Second, future generations can build on previous work to incrementally synthesize the acquisition of knowledge about a topic. Putting this suggestion into the context of the present paper, available data and scripts could have not only allowed us to estimate the effects for each individual study more accurately, but also speed up our analysis. One important tool for facilitating reproducibility is literate programming: the use of tools like *RMarkdown* and *knitr* (Xie, 2014; Xie, 2015; Xie, 2017) to produce documented code that can be released with a published paper and is available permanently in a repository.

7 Concluding remarks

Since the amount of information provided by a single study is limited, a scientific conclusion should be based on the totality of the evidence available. Using incomplete neutralization in German as a case study, we showed how quantitative evidence in the phonetic sciences can be synthesized from several studies. Our meta-analysis provides evidence in favor of incomplete neutralization, and shows that there is insufficient evidence supporting the claim that the most remarked confounds such as orthography and location of the population cause incomplete neutralization. In addition, we showed that some of the often cited earlier studies were not entirely adequate to address whether neutralization is or is not complete. These findings have led us to propose several suggestions for improving the quality of future research on phonetic phenomena. When conducting experimental studies, we should ensure that our sample sizes allow for higher-precision estimates of the effect; we should avoid the temptation to deploy researcher degrees of freedom when analyzing data; we should focus on estimates of the parameter of interest and the uncertainty about that parameter by using adequate analyses for our data; and we should allow other researchers to regenerate our results by making scripts and data publicly available.

Within the last four decades or so, incomplete neutralization has turned out to be a fruitful ground for methodological debates that advanced methodological rigor and the critical assessment of empirical findings within the phonetic sciences tremendously. We hope that the present paper continues this tradition and helps phonetics to grow further as an empirical science.

References

Baer-Henney, D. & Roettger, T. B. (2017). Control vs. power in phonetic research - the case of incomplete neutralization. Unpublished, retrieved from https://osf.io/9kywf/.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Baumann, M. (1995). *The production of syllables in connected speech* (Doctoral dissertation, University of Potsdam, University of Nijmegen).

Begley, C. G. & Ioannidis, J. P. A. (2015). Reproducibility in science. *Circulation research*, *116*(1), 116–126.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Charles-Luce, J. (1985). Word-final devoicing in German and the effects of phonetic and sentential contexts. *Journal of Phonetics*, *13*, 309–324.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2013). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Manuscript submitted for publication*.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

De Groot, A. (1956,2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica*, *148*, 188–194.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290.

Dinnsen, D. A. & Charles-Luce, J. (1984). Phonological neutralization, phonetic implementation and individual differences. *Journal of Phonetics*, *12*(1), 49–60.

Dinnsen, D. A. & Garcia-Zamor, M. (1971). The three degrees of vowel length in German. *Research on Language & Social Interaction*, *4*(1), 111–126.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*(7109), 629–634. doi:10.1136/bmj.315.7109.629. eprint: http://www.bmj.com/content

Ernestus, M. & Baayen, R. H. (2006). The functionality of incomplete neutralization in Dutch: The case of past-tense formation. *Laboratory phonology*, *8*(1), 27–49.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Fourakis, M. & Iverson, G. K. (1984). *Phonetica*, *41*(3), 140–149. doi:10.1159/000261720

Freeberg, T. M. & Lucas, J. R. (2009). Pseudoreplication is (still) a problem. *Journal of Comparative Psychology*, *123*(4), 450–451. doi:10.1037/a0017031

Fuchs, S. (2005). *Articulatory correlates of the voicing contrast in alveolar obstruent production in German*. ZAS Papers in Linguistics. Berlin: Zentrum für Allgemeine Sprachwissenschaften.

Gelman, A. & Carlin, J. (2014). Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

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

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. doi:10.1080/19345747.2011.618213

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, *19*, 555. doi:10.3390/e19100555. arXiv: 1708.07487 [stat.ME]

Gelman, A. & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Grawunder, S. (2014). Wie schaukt a Pruag aos? - Stabile phonetische Unterschiede in Wortformen nach Auslautverhärtung in Tirol. In *Sprechwissenschaft: Bestand, Prognose, Perspektive* (50, pp. 209–220). Hallesche Schriften für Sprechwissenschaft und Phonetik. Frankfurt a. M.: Peter Lang.

Greisbach, R. (2001). *Experimentelle Testmethodik in Phonetik und Phonologie: Untersuchungen zu segmentalen grenzphänomenen im deutschen*. Frankfurt am Main: Lang.

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85.

Higgins, J. P. & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

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), 19–24.

Huggins, A. W. F. (1972). Just noticeable differences for segment duration in natural speech. *The Journal of the Acoustical Society of America*, *51*(4B), 1270–1278.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*(2), 187–211. doi:10.2307/1942661

Inozuka, E. (1991). The realization of the German neutralized word-final plosives /g, k/: An acoustic analysis. *Sophia Linguistica*, *30*, 119–134.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Jäger, L. A., 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.

Jaynes, E. T. & Kempthorne, O. (1976). Confidence intervals vs. Bayesian intervals. In W. L. Harper & C. A. Hooker (Eds.), *Foundations of probability theory, statistical inference, and statistical theories of science* (Vol. 6b, pp. 175–257). The University

of Western Ontario Series in Philosophy of Science. Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-1436-6_6

Jespersen, O. (1920). *Lehrbuch der Phonetik: Mit 2 Tafeln*. Leipzig: BG Teubner.

Jessen, M. & Ringen, C. (2002). Laryngeal features in German. *Phonology*, *19*(2), 189–218.

Jongman, A., Sereno, J. A., Raaijmakers, M., & Lahiri, A. (1992). The phonological representation of [voice] in speech perception. *Language and Speech*, *35*(1-2), 137–152.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language*, 286–319.

Kharlamov, V. (2014). Incomplete neutralization of the voicing contrast in word-final obstruents in Russian: phonological, lexical, and methodological influences. *Journal of Phonetics*, *43*, 47–56.

Kirby, J. & Sonderegger, M. (this issue). Power and effect size considerations in experimental phonetics. *Journal of Phonetics*.

Klatt, D. H. (1976). Linguistic uses of segmental duration in english: acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kleber, F., John, T., & Harrington, J. (2010). The implications for speech perception of incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *38*(2), 185–196.

Kohler, K. J. (2007). Beyond laboratory phonology. In *Beyond laboratory phonology. the phonetics of speech communication* (pp. 41–53). Oxford University Press, USA.

Kohler, K. J. (2012). Neutralization?! the phonetics–phonology issue in the analysis of word–final obstruent voicing. *Gybbon, D./Hirst, D./Campbell, N.(Hg.): Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem. Poznan*, 171–180.

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods*, *15*(4), 722–752. doi:10.1177/1094428112457829

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? *BMC Neuroscience*, *11*(1), 5. doi:10.1186/1471-2202-11-5

Lecoutre, M.-P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology*, *38*(1), 37–45.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., ... Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLOS Medicine*, *6*(7), 1–28. doi:10.1371/journal.pmed.1000100

Light, R. & Pillemer, D. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech*, *29*(1), 3–11.

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

Manaster-Ramer, A. (1996). A letter from an incompletely neutral phonologist. *Journal of Phonetics*, *24*(4), 477–489.

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America*, *126*(5), 2619–2634.

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with R examples*. Chapman and Hall/CRC.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis. *Perspectives on Psychological Science*, *11*(5), 730–749. PMID: 27694467. doi:10 . 1177 / 1745691616662243. eprint: https : / / doi . org / 10 . 1177 / 1745691616662243

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

Moreno, S. G., Sutton, A. J., Ades, A., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*(1), 2. doi:10.1186/1471-2288-9-2

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. doi:10.3758/s13423-015-0947-8

Nicenboim, B. & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas - Part II. *Language and Linguistics Compass*, *10*(11), 591–613. doi:10.1111/lnc3.12207. eprint: https://arxiv.org/abs/1602.00245

Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*. doi:10.17605/OSF.IO/MMR7S

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature neuroscience*, *14*(9), 1105–1107.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716. eprint: http://science.sciencemag.org/content/349/6251/aac4716.full.pdf

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, *37*, 147–180.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer-Verlag.

Piroth, H. G. & Janker, P. M. (2004, January 1). Speaker-dependent differences in voicing and devoicing of German obstruents. *Journal of Phonetics*, *32*(1), 81–109. doi:10.1016/S0095-4470(03)00008-1

Piroth, H. G., Schiefer, L., Janker, P. M., & Johne, B. (1991). Evidence for final devoicing in German? An experimental investigation. In *Proceedings of the international congress of phonetic sciences* (Vol. 12, pp. 138–141).

Port, R. F. & Crawford, P. (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics*, *17*(4), 257–282.

Port, R. F. & Leary, A. P. (2005). Against formal phonology. *Language*, *81*(4), 927–964.

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

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Roettger, T. B., Winter, B., Grawunder, S., Kirby, J., & Grice, M. (2014). Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics*, *43*, 11–25.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Rouder, J. N. & Haaf, J. M. (In press). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, 2515245917745058. doi:10.1177/2515245917745058

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. PMID: 26173243. doi:10.1177/1745691613514755. eprint: https://doi.org/10.1177/1745691613514755

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size. *Perspectives on Psychological Science*, *9*(6), 666–681. PMID: 26186117. doi:10.1177/1745691614553988. eprint: https://doi.org/10.1177/1745691614553988

Smith, B. L., Hayes-Harb, R., Bruss, M., & Harker, A. (2009). Production and perception of voicing and devoicing in similar German and English word pairs by native speakers of German. *Journal of Phonetics*, *37*(3), 257–275. doi:10.1016/j.wocn.2009.03.001

Stan Development Team. (2017). Stan: A C++ library for probability and sampling, version 2.17.0.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning*, *19*(3), 193–198.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, *54*(285), 30–34. doi:10.1080/01621459.1959.10501497. eprint: http://dx.doi.org/10.1080/01621459.1959.10501497

Sutton, A. J., Welton, N. J., & Cooper, N. (2012). *Evidence synthesis for decision making in healthcare*. New York, NY: John Wiley & Sons.

Taylor, D. Q. (1975). The inadequacy of bipolarity and distinctive features: The German "voiced/voiceless" consonants. In *The second lacus forum* (pp. 107–119). Columbia: Hornbeam Press, Inc.

Trubetzkoy, N. (1939). *Grundzüge der Phonologie*. Prag: Travaux du cercle linguistique de Prague.

Tsuji, S. & Cristia, A. (2014). Perceptual attunement in vowels: a meta-analysis. *Developmental psychobiology*, *56*(2), 179–191.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Oostendorp, M. (2008). Incomplete devoicing in formal phonology. *Lingua*, *118*(9), 1362–1374.

Vasishth, S., Beckman, M., Nicenboim, B., Li, F., & Kong, E. J. (this issue). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*.

Vasishth, S., Chen, Z., Li, Q., & Guo, G. (2013). Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, *8*(10), 1–14.

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). *The statistical significance filter leads to overoptimistic expectations of replicability*.

Vasishth, S. & Nicenboim, B. (2016). Statistical methods for linguistic research: Foundational ideas - Part I. *Language and Linguistics Compass*, *10*(8), 349–369. doi:10.1111/lnc3.12201

von der Malsburg, T. & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. doi:http://dx.doi.org/10.1016/j.jml.2016.10.003

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. doi:10.1016/j.cogpsych.2009.12.001

Warner, N., Good, E., Jongman, A., & Sereno, J. (2006). Orthographic vs. morphological incomplete neutralization effects. *Journal of Phonetics*, *34*(2), 285–293.

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*(2), 251–276.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108. eprint: http://dx.doi.org/10.1080/00031305.2016.1154108

Wiese, R. (1996). *The phonology of German*. The phonology of the world's languages. Oxford: Clarendon Press.

Winter, B. (2011). Pseudoreplication in phonetic research. In *Proceedings of the international congress of phonetic science* (pp. 2137–2140). Hong Kong.

Winter, B. & Roettger, T. (2011). The nature of incomplete neutralization in German: Implications for laboratory phonology. *Grazer Linguistische Studien*, *76*, 55–74.

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. ISBN 978-1466561595. Chapman and Hall/CRC.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd). ISBN 978-1498716963. Boca Raton, Florida: Chapman and Hall/CRC.

Xie, Y. (2017). *Knitr: A general-purpose package for dynamic report generation in r*. R package version 1.17.