**Phonological Data & Analysis** Volume 1, Article 4: 1–23 (2019)
https://doi.org/10.3765/pda.v1art4.13

Received 2 October 2018; revised 29 August 2019; accepted 6 September 2019. 

# Toward a replication culture: Speech production research in the classroom

Timo B. Roettger[a] & Dinah Baer-Henney[b*]

[a]Northwestern University – timo.b.roettger@gmail.com
[b]Heinrich-Heine-Universität Düsseldorf – dinah.baer-henney@uni-duesseldorf.de

Our understanding of human sound systems is increasingly shaped by experimental studies. What we can learn from a single study, however, is limited. It is of critical importance to evaluate and substantiate existing findings in the literature by directly replicating published studies. Our publication system, however, does not reward direct replications in the same way as it rewards novel discoveries. Consequently, there is a lack of incentives for researchers to spend resources on conducting replication studies, a situation that is particularly true for speech production experiments, which often require resourceful data collection procedures and recording environments. In order to sidestep this issue, we propose to run direct replication studies with our students in the classroom. This proposal offers an easy and inexpensive way to conduct large-scale replication studies and has valuable pedagogical advantages for our students. To illustrate the feasibility of this approach, we report on two classroom-based replication studies on incomplete neutralization, a speech phenomenon that has sparked many methodological debates in the past. We show that in our classroom studies, we not only replicated incomplete neutralization effects, but our studies yielded effect magnitudes comparable to laboratory experiments and meta analytical estimates. We discuss potential challenges to this approach and outline possible ways to help us substantiate our scientific record.

*Keywords: final devoicing; incomplete neutralization; German; direct replication; pedagogy*

## 1 Introduction

Our understanding of human speech and its cognitive underpinnings is increasingly shaped by experimental data. For example, experimental work has repeatedly demonstrated that putative neutralizations of contrasting speech sounds are incomplete, i.e. there are small but systematic acoustic and articulatory differences between neutralized and non-neutralized speech sounds (e.g. Mitleb 1981; Roettger, Winter, Grawunder, Kirby, & Grice 2014). These findings have led to important discussions about cognitive representations of speech sounds (e.g. Dinnsen & Charles-Luce 1984; Charles-Luce 1985; Port & O'Dell 1985; van Oostendorp 2008) and have served as a springboard for theories about the organization of lexical representations (e.g. Ernestus & Baayen 2006; Goldrick, Folk, & Rapp 2010; Winter & Roettger 2011; Seyfarth, Garrellek, Gillingham, Ackerman, & Malouf 2018).

With a rapidly growing body of experimental evidence, it is of critical importance to evaluate and substantiate existing findings in the literature. This is particularly important because the evidence provided by a single study is limited to the concrete research method and context. Results may depend on

* We would like to thank Julia Muschalik and Charlotte von Kries for their comments on an early draft of this manuscript. We would like to express our gratitude to two anonymous reviewers and the editor Megan Crowhurst for their comments during the review process. All remaining errors are our own.

idiosyncratic properties of the sample, random measurement fluctuations, or human error due to variation in procedures and the experimenters involved.

In order to ensure the reliability of scientific findings, we should thus directly replicate our findings. We define direct replication as "attempting to reproduce a previously observed result with a procedure that provides no a priori reason to expect a different outcome" (Nosek & Errington 2017). In a direct replication, independent researchers follow protocols from the original study using the same or sufficiently similar materials on a new sample.[1] However, running experiments is expensive and time-consuming. This is particularly true for speech production experiments, which often require specific data collection procedures and recording environments. Moreover, our current publication system does not reward direct replications (Koole & Lakens 2012; Makel, Plucker, & Hegarty 2012; Nosek, Spies, & Motyl 2012). The lack of incentives often discourages researchers from investing time and money in direct replications.

The present paper discusses a possible solution to this dilemma. In line with several developments throughout the quantitative sciences, we propose making replication studies an explicit part of the linguistic curriculum. In other words, as we are teaching our students about experimental methods, we should involve them in running replication studies of published findings. While this proposal has been put forward for the psychological sciences (Frank & Saxe 2012; Grahe, Brandt, IJzerman, & Cohoon 2014, Grahe et al. 2018; Hawkins et al. 2018), we believe it is necessary to re-evaluate these suggestions in light of the empirical practices in experimental phonology. Phonological research is often informed by speech production experiments that exhibit their own unique methodological challenges. For the present purpose, we will reiterate these points using the above-mentioned example of incomplete neutralization in German.

The remainder of this paper is organized as follows: In §2 we will discuss the replication crisis within the quantitative sciences, which was mainly triggered by reoccurring failures to replicate 'established' findings within the quantitative sciences (§2.1). Despite the importance of replication studies, very few of them are conducted. We discuss reasons why researchers are reluctant to run replication studies and discuss a practical solution: running replication studies in the classroom (§2.2). To demonstrate the feasibility of this approach for speech production studies, we discuss incomplete neutralization of German final devoicing as a test case (§3). In §4 we describe two experimental attempts to replicate the as of yet largest lab-based experiment on incomplete neutralization in German. In §5 we present the results, and compare them to those of the original study using Bayesian parameter estimation. In §6 we discuss the benefits of replication studies in the classroom, as well as challenges and limitations to the approach. We conclude in §7 that despite a lack of control inherent in students' replication attempts, large-scale speech production studies executed by students can yield valuable data that contribute to the evidence accumulation within our field. This approach is resource-economical and offers feasible ways to replicate empirical evidence from the literature and, at the same time, substantiates the external validity of these findings.

## 2 Background

### 2.1 The 'replication crisis' and why we should be worried

The replication of a scientific finding has been an essential component of the scientific method (e.g. Dunlap 1925; Campbell 1969; Zwaan, Etz, Lucas, & Donnellan 2018). Only by replicating empirical results and evaluating the accumulated evidence can we gain a better understanding of the world. In recent coordinated efforts to replicate published results, the quantitative sciences have uncovered unexpectedly low replicability rates, a state of affairs that has been coined the 'replication crisis'.

For example, the Open Science Collaboration (2015) tried to replicate 100 studies that were published in three high-ranking psychology journals, assessing whether the replications and the original experiments

[1] In a conceptual replication, a different methodology is used in order to test the same hypothesis instead (e.g. by using a different protocol, by using different experimental manipulations, or by using different operationalizations of measures or predictors).

yielded the same result. A widely shared summary of their project was that only 36% of the attempted replications were successful (i.e. finding a significant effect with the same directionality as was found in the original study).

Concerns about the replicability of findings have been raised for many other disciplines including the medical sciences (e.g. Ioannidis 2005), cancer research (Errington et al. 2014), the neurosciences (Wager, Lindquist, Nichols, Kober, & van Snellenberg 2009), economics (Camerer et al. 2016), and genetics (Hewitt 2012).

Most importantly, it is also a very real problem for quantitative linguistics. For example, Stack, James, and Watson (2018) recently failed to replicate a widely cited effect of syntactic adaptation posited by Fine, Jaeger, Farmer, and Qian (2013). After failing to find the original effect in a conceptual replication, they went back and directly replicated the original study with appropriate statistical power. They found no evidence for syntactic adaptation as reported by the original study. Nieuwland et al. (2018) recently tried to replicate a seminal study by DeLong, Urbach, and Kutas (2005), a landmark study in the predictive processing literature. In their preregistered multi-site replication attempt (9 laboratories, 334 subjects), Nieuwland et al. were not able to replicate some of the key findings of the original study.[2] Other prominent cases of failed replications in the language sciences include Chen (2007) failing to replicate findings by Boroditsky, Fuhrman, and McCormick (2001). The original study suggested that Mandarin speakers are more likely to think about time vertically than horizontally. Papesh (2015) failed to replicate the action-sentence compatibility effect (Glenberg & Kaschak 2002), a hallmark of embodied accounts of language comprehension. Westbury (2018) repeatedly failed to replicate his own findings (2005) which originally suggested that language users associate continuant consonants with round shapes, and stop consonants with sharp shapes.

Given these findings, it seems we are facing the beginning of our very own replication crisis. We want to emphasize here that we do not commit to any evaluation of either the original studies or their replication attempts. These studies need to be considered together, evaluating the available evidence as an accumulated whole. We think it is fair to say, however, that these failed replication attempts raise the concern that we cannot take everything in our publication record for granted.

There are many possible reasons why a study cannot be replicated (e.g. Ioannidis 2005; Simmons, Nelson, & Simonsohn 2011; Gelman & Loken 2014; Silberzahn et al. 2017). Idiosyncratic properties of the sample, measurement errors, and the human factor in data collection and analysis can all have a noticeable influence on the interpretation of experimental results. Moreover, statistical issues related to low power (i.e. low probability of a statistical test to reject a false null hypothesis, e.g. for recent discussion of power in speech research see Kirby & Sonderegger 2018; Nicenboim, Roettger, & Vasishth 2018), violation of the independence assumption (Nicenboim & Vasishth 2016; Winter 2011, 2015), exploitation of researcher degrees of freedom (Roettger 2019) and the 'significance' filter (i.e. treating results as publishable because $p < 0.05$ leads to over optimistic expectations of replicability, see Vasishth, Mertzen, Jäger, & Gelman 2018) can lead to biased results. Therefore, it comes as no surprise to us if a large number of experimental linguistic findings may not stand the test of time.

## 2.2 Why don't we try to replicate more?

The aforementioned replication failures have sparked a productive discourse throughout the quantitative sciences and have led to methodological advancements and best practice recommendations. For example, there are several coordinated efforts to directly replicate important findings by multi-site projects such as the ManyBabies project (Frank et al. 2017) or Registered Replication Reports (Simons, Holcombe, &, Spellman 2014). These coordinated efforts can help us put substantive insights onto a firmer empirical

[2] But see Nicenboim, Vasishth, & Rösler (2019) for a recent metanalytical approach suggesting evidence for a small effect across available studies.

footing. However, the logistic and monetary resources required for such large-scale projects are not always available.

The popularity of replication studies is even further diminished because the time and resource investment necessary for such studies are not appropriately rewarded in contemporary academic incentive systems (Koole & Lakens 2012; Makel et al. 2012; Nosek, Spies, & Motyl 2012). Both successful replications (Madden, Easley, & Dunn 1995) and repeated failures to replicate (e.g. Doyen, Klein, Pichon, & Cleeremans 2012) are only rarely published and if they are it is usually in a less prestigious outlet than the original findings. For speech scientists in particular, the often-cited preference for laboratory experiments is an additional factor in the cost-benefit-equation (Xu 2010 for a discussion of why lab speech is more desirable than natural speech elicited in less-controlled environments): There is a common belief that even relatively simple acoustic recordings with speaker populations that are easy to recruit must be conducted under extremely well controlled laboratory conditions (but see Wagner, Trouvain, & Zimmerer 2015 and references therein).

This concern is particularly relevant for the majority of speech production experiments, which, at a minimum, require a recording device and a sufficiently quiet recording environment. These requirements make data collection often resource-intensive (at least in comparison to survey-based research). We usually record speakers in logistically costly laboratory experiments in which factors related to the testing environment are controlled for. For example, we control the recording environment by recording all speakers under sufficiently similar conditions, including conducting the experiment in the same room, by the same experimenter, with the same recording equipment, etc. This obviously poses pragmatic limits on data collection efficiency, feeding into potential reluctance of conducting replication studies. The tension between aiming for a controlled experimental environment and minimizing resource cost is a serious dilemma.

Commonly, this dilemma is tackled by testing only a small number of speakers. For example, Ladefoged (2003) recommends six speakers as a sufficient sample for speech production studies. While certainly not intended as a recommendation based on statistical reasoning, this reflects a common misconception of experimental work in quantitative linguistics. Published studies in linguistics and related areas often have very low statistical power. For example, Kirby and Sonderegger (2018) report simulation studies showing that speech production studies using six speakers may have power as low as 6% (dependent on the effect size), i.e. having a 6% probability that the statistical test will reject a false null hypothesis. While typical subject numbers differ across subdomains of quantitative linguistics, a sample size of six is not uncommon in phonetic experiments (see Roettger & Gordon 2017, on the word stress literature; and Nicenboim et al. 2018, for a more general discussion).

It is important to note that we are neither bound to laboratory experiments nor to small sample sizes. As opposed to rigid laboratory settings, there are alternative data collection approaches that are relatively resource-conserving. These empirical approaches often introduce a less controlled experimental environment, but allow us to gather large data sets quickly and efficiently. Trading control for convenience in data collection and sample size is in line with a noticeable trend toward using crowdsourcing platforms such as Amazon Mechanical Turk (e.g. Mason & Suri 2012). Crowd-sourced experimental data has been repeatedly cross-validated with laboratory experiments (Behrend, Sharek, Meade, & Wiebe 2011; Goodman, Cryder, & Cheema 2012; Shapiro, Chandler, & Mueller 2013) and can thus be considered a valid alternative to laboratory experiments.

However, not all types of empirical investigations can be outsourced to browser-based applications, at least not yet. Speech production studies still necessitate a recording device and (depending on the measure) a relatively quiet recording environment. Even though we are optimistic that we can outsource many empirical efforts to crowdsourcing platforms soon, conducting replication studies on platforms such as Amazon Mechanical Turk still costs a substantial amount of money, making it, again, less attractive for researchers.

We would like to put forward an alternative strategy. We propose to run replication studies conducted by students as class assignments (Frank & Saxe 2012; Grahe et al. 2014, 2018; Hawkins et al. 2018). In other words, while we are teaching our students about experimental methods, we ask them to run replication studies of published findings. Letting students run these studies as class assignments can be a feasible and affordable way to replicate often-cited experimental findings.[3]

Not only would this amplify the number of independent replications, enabling us to put our theories on a firmer empirical footing, but it would also create an educationally rewarding environment for the students to learn about the basic tenets of the scientific method (Frank & Saxe 2012). While some might be concerned that speech production experiments are not suited for this proposal, this paper will demonstrate that the apparent noisiness of speech production data can be overcome by large sample sizes, revealing comparable population estimates and, crucially, comparable or even higher precision (i.e. uncertainty about our estimate of interest) than controlled laboratory studies. We discuss two production experiments that were run in the classroom, investigating a notoriously brittle speech production effect that has caused many methodological and theoretical debates within the last 40 years or so: incomplete neutralization of final devoicing.

## 3 A case study: Incomplete neutralization

Within the last four decades, incomplete neutralization has proven to be a fruitful ground for methodological debates that have advanced methodological rigor and the critical assessment of empirical findings within the phonetic sciences (e.g. Manaster Ramer 1996; Winter & Roettger 2011; Roettger et al. 2014). Recently, incomplete neutralization has served as a prime example to facilitate discussions about statistical concepts such as Type-I and Type-II errors, pseudoreplication, and meta-analysis (Roettger et al. 2014; Kirby & Sonderegger 2018; Nicenboim et al. 2018).

Incomplete neutralization refers to cases in which an assumed phonological neutralization such as final devoicing is phonetically not complete. Cross-linguistically, final obstruent devoicing is a widespread phonological alternation. The textbook example is the German Auslautverhärtung: German contrasts voiced (e.g. /b,d,g/) and voiceless obstruents (e.g. (/p,t,k/) prevocalically but neutralizes the contrast in syllable final or word final position in favor of voiceless obstruents (cf. 1-2):

| | | | |
|---|---|---|---|
| (1) | /d/ in syllable final position | Rad [ʁa:t] | ‘wheel’ |
| | /d/ in syllable initial position: | Räder [ʁɛ:dɐ] | ‘wheels’ |
| (2) | /t/ in syllable final position: | Rat [ʁa:t] | ‘council’ |
| | /t/ in syllable initial position: | Räte [ʁɛ:tə] | ‘councils’ |

In intervocalic position, the obstruent voicing contrast can be manifested by different acoustic dimensions, such as the duration of the preceding vowel, glottal pulsing during the closure, closure duration, or voice onset time (among other cues, e.g. Lisker 1986), with voiced stops being preceded by longer vowels, exhibiting more glottal pulsing during the closure, a shorter closure duration, and shorter (or negative) voice onset time. These acoustic differences are supposedly neutralized in syllable final position, implying that the acoustic form of the alveolar stop in Rad is identical to the alveolar stop in Rat.

This observation was mainly grounded on early ear-phonetic assessments (e.g. Jespersen 1920; Trubetzkoy 1939; Wiese 1996). However, as soon as linguists started testing this assumption experimentally, small but consistent acoustic and/or articulatory differences between words such as Rad and Rat were found. These findings suggest that in German, this neutralization can be considered incomplete (e.g. Dinnsen & Garcia-Zamor 1971; Taylor 1975; Mitleb 1981; Port & O’Dell 1985; Charles-

[3] Note that there are published phonetic studies that have collected data in the class room (e.g. Sluijter, van Heuven, & Pacilly 1997; Iversen, Patel, & Ohgushi 2008). Sluijter et al. (1997) showed that perception data collected in a class room setting yielded comparable results to data collected in the laboratory, demonstrating the general feasibility of data collection in class room settings.

Luce 1985; Port & Crawford 1989; Roettger et al. 2014). While the direction of the measured differences mirrors the non-neutralized contrast (Räder vs. Räte), the magnitude of the effect turned out to be much smaller. For example, in intervocalic position, voiced stops are characterized by large durational differences in the preceding vowel (the /ɛ:/ in Räder vs. the /ɛ:/ in Räte) with reports on vowel duration differences ranging from 20 to 40 ms (see Mitleb 1981; Fuchs 2005; Roettger et al. 2014). When the stop is in syllable final position, however (i.e. the putative neutralized position), the durational difference in the preceding vowel has been reported to be much smaller. For example, Roettger et al. (2014) report a 9 ms difference between devoiced and voiceless stops. These effects, albeit repeatedly found, are most likely below the just noticeable difference threshold (e.g. Kohler 2012) and might not be perceivable by listeners in isolation. This is in line with listeners' poor identification performances in forced-choice tasks (e.g. Roettger et al. 2014), suggesting only limited communicative relevance of these subtle acoustic differences.

The main response to these findings was acknowledgement of the evidence for incomplete neutralization, leading to attempts to implement this phenomenon into formal models of phonological representations (e.g. Dinnsen & Charles-Luce 1984; Charles-Luce 1985; Port & O'Dell 1985; van Oostendorp 2008). More recent accounts of incomplete neutralization are rooted in psycholinguistic models of lexical organization, suggesting that incomplete neutralization and similar phenomena are artifacts of lexical co-activation (e.g. Ernestus & Baayen 2006; Goldrick et al. 2010; Kleber et al. 2010; Winter & Roettger 2011; Roettger et al. 2014; Seyfarth et al. 2018).

In contrast, a substantial number of researchers have remained skeptical regarding incomplete neutralization, mainly concerned that it might be a methodological artifact (Manaster Ramer 1996; Kohler 2007, 2012; Roettger et al. 2014). Some studies even failed to replicate incomplete neutralization (Fourakis & Iverson 1984; Inozuka 1991; Jessen & Ringen 2002; Piroth & Janker 2004, but see Nicenboim et al. 2018 who argue that most of these findings are due to too low statistical power).

The latter position has led to productive methodological debates, drawing attention to the problem of how much a single study can tell us about the world. Nicenboim et al. (2018) took this as a point of departure and gathered all eligible studies on German incomplete neutralization to quantify the overarching evidence and the uncertainty it is associated with. Using a Bayesian random-effects meta-analysis, they showed that across fourteen eligible studies, there was compelling evidence for an incomplete neutralization effect. The meta-analytical estimate for the magnitude of the effect was 10 ms [95% credible interval: 6-16].[4]

Despite its presumed brittleness, incomplete neutralization of final devoicing appears to become more and more substantiated as evidence accumulates. It arguably becomes one of the more robust findings of our field and is one of the few phenomena that have undergone substantial empirical scrutiny. Moreover, the effect has been recently approached via the affordances of meta-analysis, offering a quantitative baseline for future work. Against this background, incomplete neutralization is a promising phenomenon to demonstrate the feasibility of classroom-based replication studies.

## 4 Method

In the following we describe the methodology of two studies that were attempts to replicate the recent incomplete neutralization findings of Roettger et al. (2014, Experiment 1). Since the original data are publicly available (http://osf.io/y7tdq), we can directly compare the estimated effect magnitudes and its precision to the original study. In §4.1 we briefly summarize the experimental design employed in the original study and in §4.2 we discuss the context and methodology of our two classroom studies.

[4] At the time of performing the meta-analysis, the data presented in the present paper were already publicly available. Nicenboim et al. performed their meta-analysis with and without the present data and found similar results.

### 4.1 The experimental design of Roettger et al. (2014)

In the following we briefly summarize relevant dimensions of the original study. For more detailed information about the method, see the original paper. Sixteen native speakers of German participated in the experiment. All were university students in the humanities living in the Cologne area. Most of them grew up in this area and all participants claimed to speak the Standard variety of German spoken in the area. The experiment was conducted in a sound-treated booth in a formal laboratory setting. On each trial, speakers heard a stimulus sentence such as (3) containing a nonce word inflected for plural and were subsequently asked to produce a corresponding sentence such as (4) containing a nonce word inflected for singular. Crucially, the singular form of the nonce contained a stop in word-final position and is thus subject to final devoicing.

(3) Plural stimulus
Aus Dortmund kamen die Drude.
'From Dortmund came the nonce-PL.'

(4) Singular response
Ein Drud wollte nicht mehr.
'One nonce-SG did not want to continue.'

The experimental items consisted of 24 nonce word pairs, varying in terms of vowel quality and place of articulation of the critical stop. The duration of the vowels preceding the final stops were measured and submitted to statistical analysis. The authors report on a significant incomplete neutralization effect, i.e. vowels preceding devoiced stops were on average 8.6 ms longer than vowels preceding voiceless stops. The following two studies took Roettger et al.'s study as a point of departure and attempted to replicate their results.

### 4.2 Study 1 and 2

Study 1 and 2 were conducted by students during the summer terms 2015 and 2017 at the Heinrich-Heine-Universität Düsseldorf. These studies were originally intended with several educational goals in mind: to teach students about relevant and contemporary research questions in phonetics and phonology; to give them first hands-on experience in designing, running, and analyzing experimental data; and to guide their first interpretation of empirical findings in light of the existing literature. These studies lack many aspects of control that we traditionally aim for when we conduct lab-based experiments (see §4.2.5 for a summary). This lack of control might introduce noise. We will discuss this possible source of variability in §6.

During the semester, the lecturer (the second author) introduced the topic of final devoicing in German and its potential incomplete nature. Two opposing proposals on incomplete neutralization were introduced, i.e. the proposal that incomplete neutralization is real and warrants (phonological) explanation (e.g. Port & O'Dell 1985) or the proposal that incomplete neutralization is a methodological artifact (e.g. Fourakis & Iverson 1984). In light of the latter, methodological concerns were discussed without mentioning the experimental studies by Roettger et al. (2014) in class. Guided by the lecturer, a study was planned which was intended to address previous methodological concerns. Unknown to the students, the lecturer secretly proposed the study design used by Roettger et al. As far as we can tell, students were not aware of the original study until the 'reveal' at the end of the semester. In what followed, the lecturer guided students in designing, running, and analyzing the study.

The class spanned 12 weeks (one semester). There was one general lecture each week which gave theoretical and methodological background, and a subsequent sub-seminar, in which students received practical training from more experienced student tutors (graduate students). Different tutors guided the sub-seminar as part of their own teaching training or as student assistants. There were ten different sub-seminars in study 1 and eleven different sub-seminars in study 2, respectively.

### *4.2.1 Experimenters, participants, and procedure*

In Study 1, 79 students recorded data from one speaker each. They were asked to record one native German adult, recruited from their personal environment who was not participating in the class. The procedure resulted in 79 speakers (43 females, 36 males, mean age = 29, ranging from 17 to 62). In Study 2, 65 students collected data from one speaker each, resulting in 65 participants (38 females, 27 males, mean age = 29, ranging from 18 to 52).

Students, functioning as individual experimenters, were undergraduates majoring in general linguistics. They were in their second semester and had already passed an introductory lecture on phonetics. They had no background in experimental research.

### *4.2.2 Design*

Following Roettger et al., students were guided to design a nonce word study in which German speakers would be asked to derive singulars from given nonce word plurals. Generally, speakers heard a stimulus sentence containing a nonce word inflected for plural and were subsequently asked to produce a corresponding sentence containing a nonce word inflected for singular. The singular form of the nonce contained a stop in word-final position.

Study 1 and 2 differed with respect to the context narrative that was spun around the task and presented to the speakers as part of the instructions. In Study 1 speakers were told that Noah was about to rescue more animals, which were still missing just before the departure of the ark. Speakers were instructed to listen to an audio playback telling them which animals were still missing (plural sentence: "Noah wollte schon mit seiner Arche ablegen. Es fehlten nur noch zwei Drude."; English: "Noah wanted to set sail with the ark. Only two nonce-PL were missing."). Participants subsequently had to continue the discourse with the following sentence: "Leider ist ein Drud ausgebüxt." (English: "Sadly, one nonce-SG ran away.")

In Study 2 the story was about "Pakos", a new toy trend among little children. Speakers were instructed to listen to an audio playback in which a sales assistant tells the speaker where to find the specific toy (plural sentence: "In der Spielzeugabteilung gibt es eigene Aufsteller für den neuen Trend. Besonders beliebt sind die Drude." English: 'In the toy department, there are stand-up displays for the new trend. Nonce-PL are particularly popular.'). Participants subsequently had to continue the discourse with following sentence: "Ich habe einen Drud eingepackt." (English: 'I took one nonce-SG.').

The narrative was the same across all sub-seminars within each study, respectively. However, the stimuli differed across sub-seminars (see below). The narrative context alongside the instructions was presented in written form via PowerPoint. After a familiarization phase with six real word examples (i.e. Kühe 'cows', Schafe 'sheeps', Hühner 'chicken', Elefanten 'elephants', Pinguine 'penguins', and Kamele 'camels'), speakers responded to the experimental trials in a self-paced manner. Stimuli were pseudorandomized. Speakers' productions were recorded via Audacity (Team Audacity 2015), Praat (Boersma & Weenink 2015) or other smartphone devices, as the individual experimenters saw fit. There were no restrictions on the type of microphone that students used.

### *4.2.3 Materials*

One hundred and eighty trochaic nonce words of the shape $(C_1)C_2V_1C_3V_2$ served as stimuli for each sub-seminar. $V_2$ was always an unstressed schwa in order to mimic common German plural forms. All stimuli conformed to German phonotactics. In target words, $C_3$ was always occupied by a stop, balanced for place of articulation (labial, coronal, dorsal) and voicing status (voiced, voiceless). This resulted in 90 target pairs. The word onset ($(C_1)C_2$) was variable, with $V_1$ always being a tensed vowel (i.e. one of the following vowels (/i,a,u,o,e/).

Additionally, a list of 90 fillers was prepared. Fillers had the same phonotactic structure as targets. In fillers, $C_3$ was a sonorant (/l, m, n/) or a voiceless fricative (/f, ʃ/). Half of the fillers had a front vowel in $V_1$ position (/y,ʏ,ø,œ,ɛ/) which can be interpreted as an umlaut. Plural forms with umlaut vowels sometimes

do and sometimes do not require a vowel change (e.g. Türme > Turm 'towers/tower' but Hütten > Hütte 'huts/hut'). Following Roettger et al. (2014), this design choice was intended to increase the salience of the fillers and detract attention from the critical manipulation.

All plural sentences were recorded in a sound-treated booth at 48 kHz (Phantom 48V Microphone, Marantz Recorder pmd 570), produced by a different native German speaker for each sub-seminar.

#### *4.2.4 Acoustic analysis*

The students carried out the acoustic analyses. As part of the training, they were taught to identify and annotate segment boundaries for vowels using Praat (Boersma & Weenink 2015/2017). The duration of the vowels preceding the final stops was measured. The following instructions were given to the students: If the sound preceding the vowel/diphthong was a stop, the onset of the vowel was defined as the onset of voicing in cases of voiceless stop or as the end of the burst in cases of voiced stops. A sudden discontinuity in the spectrogram was taken as the onset of vowels following fricatives, nasals, laterals and palatal approximants (e.g. [ʃu:k], [mu:p], [blo:k] or [je:t]). The end of the vowel was defined as the end of the second formant of the vowel, which usually coincided with a sudden drop in amplitude of voicing. The annotated data from each student were then distilled into one data table that was subsequently used for statistical analysis in the class.

#### *4.2.5 Comparison to the original study*

The experimental rationale and most of the design choices are comparable to Roettger et al.'s study. The general elicitation method (auditory plural-to-singular derivation), the structure and segmental make-up of the nonce words and the type of filler items are also comparable to the design choices of the original study. We thus consider the procedure sufficiently close to the original study with no a priori reason to expect a different outcome. The instruction and the accompanying narrative, the actual nonce words, and the speaker voices necessarily differ from the original study (as does, of course, the sample). These differences resemble common differences between original studies and replication attempts.

The original study controlled for several factors that were not controlled for in the classroom studies. Since each student recorded his/her own speaker, there were as many experimenters as there were participants. With each experimenter came a different combination of recording environment, recording device (e.g. laptop or phone), and recording software (e.g. Audacity or Praat). Moreover, conducting the experiment was a homework assignment. Thus, the lecturer was not present. Related to that, our experimenters/annotators had only little experience with conducting experiments, recording, and annotating speech data. All of these factors arguably introduce noise into our dataset.

### 4.3 Statistical analysis

All data tables and R scripts[5] are available here: https://osf.io/9kywf/. We submitted data on vowel duration preceding the critical stops of all three experiments (Roettger et al.'s Experiment 1, Study 1 and Study 2) to Bayesian hierarchical models using the Stan modelling language (Carpenter et al. 2017) and the R package brms (Bürkner 2017). We did not 'clean' the data according to post hoc exclusion criteria in order to avoid exploitation of researcher degrees of freedom (Roettger 2019). We operate within the Bayesian inferential framework (rather than within a frequentist framework) for two reasons:

First, Bayesian methods allow us to directly answer the primary question: How plausible is our hypothesis given the data? We can answer this question by quantifying our uncertainty about the parameters of interest, which frees us from committing to hard cut-off points for statistical significance (such as the arbitrary 0.05 alpha level).

[5] We used the following R packages for data processing and visualization: rstudioapi (Allaire, Wickham, Ushey, & Ritchie 2017), gridExtra (Augui 2017), ggbeeswarm (Clarke & Sherrill-Mix 2017), and tidyverse (Wickham 2017).

Second, it is easier to flexibly define hierarchical models (also known as mixed effects or multilevel models) in the Bayesian framework than in the frequentist framework. Frequentist linear mixed models have become standard in quantitative linguistics and they are commonly fit with the lme4 package (Bates, Mächler, Bolker, & Walker 2015b) in R. However, linear mixed effects models that also include the maximal random effect structure justified by the design (Barr, Levy, Scheepers, & Tily 2013; Bates, Kliegl, Vasishth, & Baayen 2015a; Schielzeth & Forstmeier 2008) tend not to converge or to give unrealistic estimates of the correlations between random effects (Bates et al. 2015b). In contrast, the maximal random effects structure can be fit without problems using Bayesian hierarchical models. Compared to non-maximal models using lme4, our maximal Bayesian models can be considered more conservative.

For Roettger et al.'s study, and both Study 1 and 2, we fit hierarchical regression models to vowel duration predicted by VOICING (voiceless vs. devoiced, dummy coded). The models included maximal random-effect structures, including a random intercept for target words and speakers (nested in sub-seminars for Study 1 and 2), and random slopes allowing the effect of VOICING to vary by target words and speakers (nested in sub-seminars for Study 1 and 2).[6]

For the critical model parameter VOICING, we used regularizing Gaussian priors (e.g. Gelman et al. 2008) ($\mu = 0$, $\sigma = 50$) (i.e. our prior assumption is agnostic as to whether there is an effect of VOICING, thus making our model conservative as to the question of whether neutralization really is incomplete.). We used truncated Student-t priors for all standard deviations ($\mu = 0$, $\sigma = 41$, dfs = 3) and LKJ ($\eta = 1$) priors for all correlation parameters. Four sampling chains with 2,000 iterations were run for each model, with a warm-up period of 1,000 iterations. We report 95% credible intervals (CIs) and the posterior probability that the VOICING coefficient is smaller than zero $Pr(\beta < 0)$. A 95% credible interval demarcates the range of values that comprise 95% of the probability mass of our posterior beliefs. We judge there to be compelling evidence for an effect if 0 is (by a reasonably clear margin) not included in the 95% CI and $Pr(\beta < 0)$ is close to 0.

As of yet, there is no single standard for evaluating replication success (Open Science Collaboration 2015). Here we evaluated the replication of Roettger et al. using the posterior population estimates for the difference between devoiced and voiceless stops. We claim there to be a successful replication *of the general phenomenon of incomplete neutralization*, if the 95% of the probability mass of our posterior interval does not include 0 and $Pr(\beta < 0)$ is close to 0. We claim there to be a successful replication *of the original study*, if the estimated means of our replication attempts fall within the 95% CI of the original estimate (Vasishth et al. 2018).

## 5 Results and discussion

### 5.1 Reanalysis of Roettger et al. (2014) as a baseline

Figure 1 shows the observed differences between devoiced and voiceless stops in Roettger et al.'s Experiment 1. The semitransparent dots are individual aggregates. As the figure clearly indicates, the majority of aggregates for both speakers and items are above zero (dashed line) with most speakers/items showing an incomplete neutralization effect. Our Bayesian inferential assessment speaks to these descriptive observations: Contingent on the data and model, there is compelling evidence for vowels preceding devoiced stops having greater duration than vowels preceding voiceless stops ($\beta = 9.3$, 95%CI = [4.8,13.6]; $Pr(\beta < 0) \approx 0$). Using Bayesian inference, we thus reproduced Roettger et al.'s results and can reiterate that speakers in this study produced larger durations for vowels preceding devoiced stops. These values can now function as a baseline against which we can compare the results of Study 1 and 2. We first

[6] The random variation associated with each speaker also corresponds to variation associated with different experimenters (i.e. each experimenter recorded one speaker), as well as variation associated with a combination of recording device, microphone, and recording environment.

discuss the individual analyses for both Study 1 and 2 individually and then compare them to the original study.

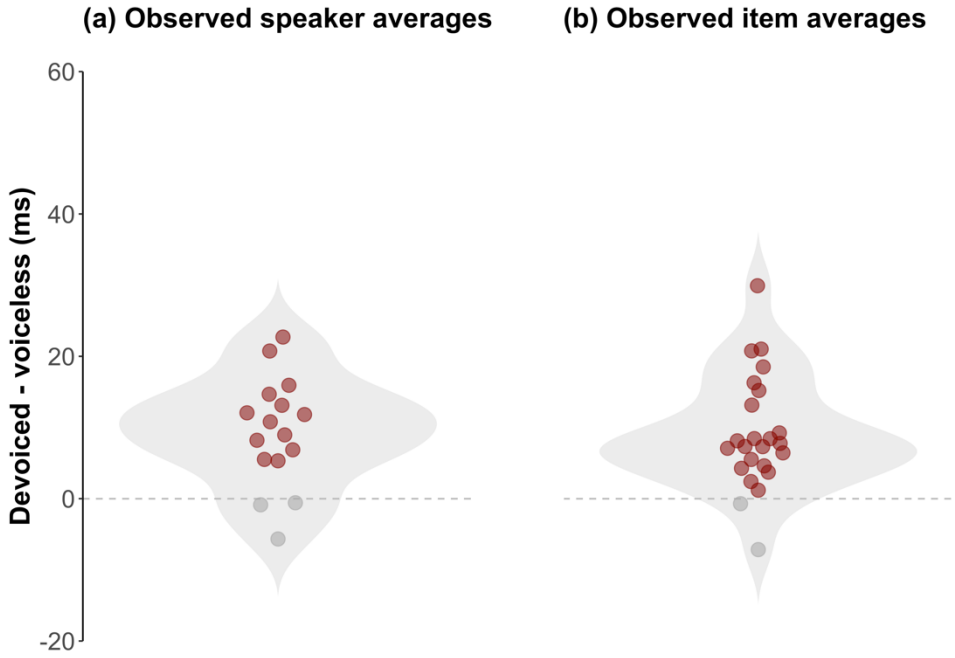


Figure 1: Observed difference between devoiced and voiceless stops in Roettger et al.'s Experiment 1 for individual speakers (left panel) and individual items (right panel). Semitransparent red dots indicate values that are compatible with incomplete neutralization effects, i.e. vowels preceding devoiced stops exhibit greater duration than vowels preceding voiceless stops. The gray density shape in the background represents the probability density along the measurement.

## 5.2 Results of Experiment 1

Figure 2 shows the observed differences between devoiced and voiceless stops in Study 1. Despite a substantial number of data points indicating a reversed effect (see also Roettger et al. 2014), there is clearly a stochastic dominance pointing toward incomplete neutralization (i.e. red dots above zero) for the sample of both speakers and target words.

Our posterior intervals corroborate these descriptive observations: Contingent on the data and model, there is compelling evidence for vowels preceding devoiced stops having greater durations than vowels preceding voiceless stops ($\beta = 7.9$, 95% CI [4.7, 11.1], $Pr(\beta < 0) = 0.0003$), replicating an incomplete neutralization effect.

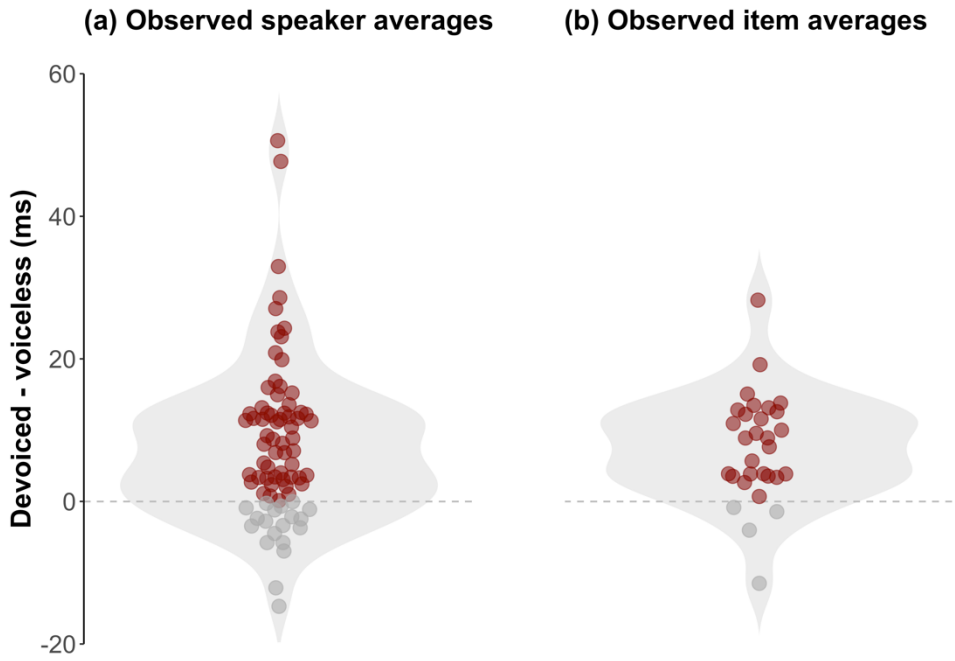


Figure 2: Observed difference between devoiced and voiceless stops in Study 1 for individual speakers (left panel) and individual items (right panel). Semitransparent red dots indicate values that are compatible with incomplete neutralization effects, i.e. vowels preceding devoiced stops exhibit greater duration than vowels preceding voiceless stops. The gray density shape in the background represents the probability density along the measurement.

## 5.3 Results of Experiment 2

Figure 3 shows the observed differences between devoiced and voiceless stops in Study 2. As before, the plot clearly indicates that the majority of aggregates for both speakers and items are above zero (dashed line), with most speakers/items showing an incomplete neutralization effect.

Again, our posterior intervals corroborate these descriptive observations: Based on the data, there is compelling evidence for vowels preceding devoiced stops having greater duration than vowels preceding voiceless stops ($\beta = 9.2$, 95% CI = [5.4, 12.9], $Pr(\beta < 0) \approx 0$). Study 2 has also replicated an incomplete neutralization effect. We proceed by asking how well these two studies replicate the original study by Roettger et al. (2014).

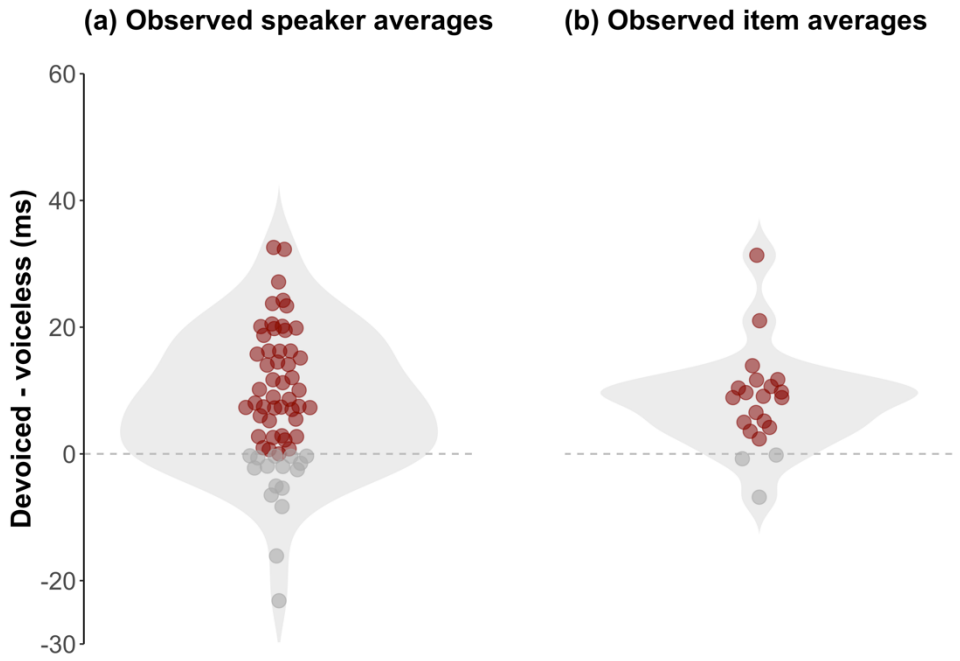


Figure 3: Observed difference between devoiced and voiceless stops in Study 2 for individual speakers (left panel) and individual items (right panel). Semitransparent red dots indicate values that are compatible with incomplete neutralization effects, i.e. vowels preceding devoiced stops exhibit greater duration than vowels preceding voiceless stops. The gray density shape in the background represents the probability density along the measurement.

## 5.4 Comparison to Roettger et al. (2014)

While we have replicated incomplete neutralization of final devoicing with similar numerical differences between devoiced and voiceless stops to other lab-based studies including Roettger et al. (2014), the estimated effect magnitude is very small. This raises the question as to how well our studies replicate Roettger et al.'s results. To quantify the answer to this question, we compare the posterior distributions. As specified above, we claim there to be a successful replication of the original study if the estimated means of our replication attempts fall within the 95% CI of the original estimates (Vasishth et al. 2018).

Figure 4 illustrates a high level of overlap between all three studies. Study 1 suggests a slightly smaller magnitude of the incomplete neutralization effect with higher precision (i.e. uncertainty about our estimate of interest) than both Study 2 and Roettger et al.'s original study. The posterior distribution of Study 2 almost entirely overlaps with the posterior distribution of Roettger et al.'s study, exhibiting a similar location with slightly higher precision. As the point and whiskers below the distributions in Fig. 4 suggest, both the mean of Study 1 and 2 fall within the 95% CI of Roettger et al.'s data.

We can thus conclude that our replications in the classroom were successful. Despite a low level of control over factors commonly controlled for in the laboratory, the classroom replications show an even higher precision than their laboratory counterpart. This increased precision is mainly achieved by a larger sample (16 speakers in Roettger et al. vs. 79/65 speakers in Study 1 and 2). Given the more feasible data collection procedure, such large sample sizes can be easily achieved within the classroom scenario.

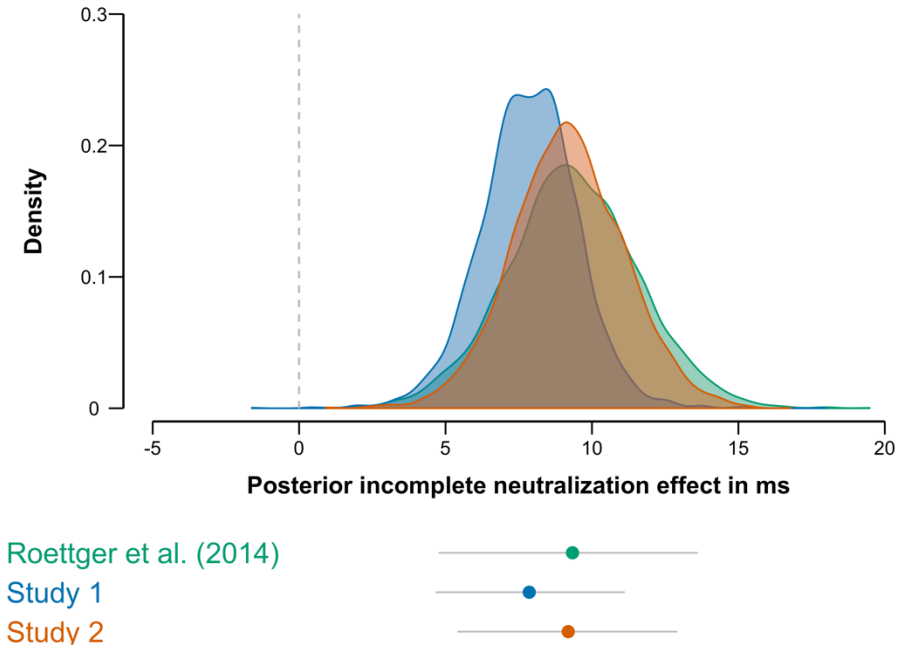


Figure 4: Posterior distributions of the incomplete neutralization effect (duration of vowel preceding devoiced stop - duration of vowel preceding voiceless stop) for all three data sets. The points and whiskers below the density curves indicate the posterior means and 95% credible intervals for all three studies.

## 6 General discussion

### 6.1 Summary

Our studies indicate that final obstruent devoicing in German is incomplete. The vowels preceding 'devoiced' stops, i.e. stops derived from a morphological paradigm that contains voiced stops intervocalically, exhibit a greater duration than those preceding voiceless stops. The design of our studies closely resembled the experimental paradigm used by Roettger et al. (2014), a high-powered laboratory-based study on incomplete neutralization in German.

In addition to simply replicating incomplete neutralization (i.e. merely testing whether the null hypothesis is plausible given the model and the data), we were able to replicate the original magnitude of the effect by Roettger et al. (2014).

It is important to replicate and therefore substantiate the magnitude of an acoustic effect, because there are limits to what acoustic differences listeners can detect reliably. For durational differences for example, it is commonly assumed that listeners can reliably detect differences of 10 ms upwards (e.g. Hirsh 1959; Klatt 1976; Lehiste 1976; Miller 1989; Phillips, Gordon-Salant, Fitzgibbons, & Yeni-Komshian 1994). As our statistical analyses have shown, values above 10 ms are plausible given the data and the model, however values below 10 ms are even more plausible. Nicenboim et al.'s (2018) meta-analysis on incomplete neutralization in German suggests likely values between 6 and 18 ms. However, as the authors discuss in detail, these values might be biased toward higher values due to publication bias (i.e. the tendency to publish positive results more often than null results, e.g. Sterling 1959) and Type-M errors (i.e. an overestimation of effect magnitudes due to small sample sizes, Gelman & Carlin 2014). While we can be more and more

certain that there is a difference in the duration of the vowel preceding devoiced and voiceless stops, the jury is still out on whether the magnitude of the difference is likely to play any role in speech communication (Kohler 2012; Winter & Roettger 2011; Roettger et al. 2014). If these differences are relevant for communication, we should consider them within our models of phonological representations (e.g. Dinnsen & Charles-Luce 1984; Charles-Luce 1985; Port & O'Dell 1985; van Oostendorp 2008). If they are not, incomplete neutralization might be better interpreted as an artifact of lexical organization and spreading activation (e.g. Ernestus & Baayen 2006; Goldrick et al. 2010; Kleber et al. 2010; Winter & Roettger 2011; Roettger et al. 2014; Seyfarth et al. 2018).

Regardless of its interpretation, the present body of evidence invites us to be rather certain about the existence of incomplete neutralization. Given the relatively large number of published studies, its recent meta-analytical assessment (with and without taking the experiments of this paper into account), and the presence of numerous successful replication attempts, including our own, incomplete neutralization of final devoicing has become one of the most robust phenomena within experimental phonology.

## 6.2 Creating a replication landscape

We chose incomplete neutralization for illustrative purposes. Taking this case study as a point of departure, we argue that direct replications of published experiments do not need to be resource-intensive, but can be done in the classroom in collaboration with our students. This concept makes replications feasible in terms of overall cost/benefit ratio. Ultimately, this approach will hopefully lead to more replication studies being conducted. This is important because the recent replication failures across the quantitative sciences demonstrated how misleading published results can be. The awareness of these issues has led to an ongoing and intensive discourse across academic disciplines and has resulted in methodological advancements as well as increased rigor across disciplines. However, there is still much work to be done. One important step to ensure scientific reliability is to conduct direct replication studies and to create an environment in which such replication studies are incentivized.

Although resource-intensive multi-site replication attempts will remain the gold standard (Open Science Collaboration 2015; Frank et al. 2017; Nieuwland et al. 2018), more accessible ways to conduct replication studies have been suggested recently (Frank & Saxe 2012; Grahe et al. 2014, 2018; Hawkins et al. 2018). As we have also argued in this paper, involving our students can be a fruitful and inexpensive way forward for speech production research, yielding educational benefits for students. The fact that the original study was published suggests that the authors asked a relevant research question. Investing time and effort into replicating published results is thus justified and gives the students an opportunity to make a real contribution to the field. The hardest part of research projects for students lacking subject-matter expertise is often finding a research question in the first place, as well as finding a method to answer their question. Taking an original study as a point of departure sidesteps the important conceptual work and allows students to focus solely on methodological considerations. Conducting a replication raises awareness about important details of experimental design and statistical analyses. Trying to reconstruct the original rationale behind methodological decisions offers great insight into the process of quantitative work. After running the experiment and analyses, students will be rewarded with one of two important experiences: Either they are able to substantiate the results of a published study, i.e. they produced research results on par with the scientific record, or they fail to replicate a published study. Either way, they will learn an important lesson about the robustness of published findings and their possible role in advancing the field. Replicating in the classroom thus offers a rich learning experience for our students. At the same time, this approach reduces the resource costs of replication studies.

In addition to finding affordable ways of conducting replication studies, we need to fundamentally reconsider our incentive systems. There are currently few mechanisms in place to encourage replication attempts. For example, Martin and Clarke (2017)'s survey results suggest that in 2015 only 3% of psychology journals explicitly state that they will consider publishing replications. In order to overcome the asymmetry between the cost of direct replication studies and the presently low academic payoff for it,

we as a research community must re-evaluate the value of direct replications. Funding agencies, journals, editors, and reviewers should start valuing direct replication attempts, be it successful replications or replication failures, as much as they value novel findings. For example, we could either dedicate existing journal space to direct replications (e.g. as its own article type) or by creating new journals that are specifically dedicated to replication studies. Alternatively, we could implement Replication Awards by communities, institutes or journals (Gorgolewski, Nichols, Kennedy, Poline, & Poldrack 2018) or encourage Ph.D. students to replicate existing findings as their first step toward approaching any empirical finding ('replication-first rule', Kochari & Ostarek 2018). Another compelling model is the Pottery Barn rule, first proposed by Srivastava (2012) and recently implemented by Royal Society Open Science. The idea behind the Pottery Barn rule is that once a journal has published a study it becomes accountable for this study in that it becomes responsible for publishing direct replications.

As soon as we make publishing replications easier, more researchers will be willing to replicate both their own work and the work of others. Only by replicating empirical results and evaluating the accumulated evidence can we substantiate previous findings and extend their external validity. Thus, the present proposal is only one of many strategies to counter our present crisis in confidence and to offer a solution to the cost-reward asymmetry that is currently in place for replication studies. Our proposed strategy, replications in the classroom, however comes with its own set of limitations and challenges to which we now turn.

### 6.3 Limits and challenges of this approach

Despite methodological debates surrounding the trade-off between experimental control and stylistic variation (Xu 2010; Wagner et al. 2015 and references therein), an increasing number of speech scientist use crowdsourcing platforms to collect linguistic data (Hasegawa-Johnson, Cole, Jyothi, & Varshney 2015). This suggests that researchers are willing to trade an increased level of variability and lack of control with quick and convenient access to large amounts of data. In line with this trend, we propose to give up some level of control for efficient access to large samples. One possible concern with giving up control is whether the data collected in the classroom is of sufficient quality.

Our students were neither experienced experimenters nor trained annotators. Their limited domain expertise might be cause for concern about the validity of the acoustic recordings and annotations. The amount of variability within and across annotators clearly exceeded the more homogeneous patterns found in laboratory studies with experienced annotators. When we compare the credible intervals estimated from the original study with the ones estimated from the replication studies, we can see that replication studies exhibit only slightly greater precision than the original study. This is striking because the classroom studies had a substantially larger number of both speakers and experimental items. Future approaches should attempt to decrease this large variability by for example stricter procedure protocols and by integrating measurement reliability checks in which students negotiate annotation criteria and assess inter-rater agreement across subsets of measurements. Integrating more structured protocols and quality control mechanisms might substantially increase the sensitivity of the study. [7]

We acknowledge that the quality of replications in the classroom will heavily depend on the students' motivation and care as well as on their supervision by the instructors. In light of these degrees of freedom, the instructor ultimately needs to evaluate the quality of the research and, if found suitable, should be the leading communicator of the results. Hawkins et al. (2018) state that over half of the projects conducted in their courses met their quality criterion. Keep in mind that without any post hoc quality control in the present data set, we still produced results compatible with both controlled laboratory experiments and recent meta-

[7] We believe that much of the observed variability is mainly due to lack of experience and is not related to motivation or diligence. The feedback received from the students indicate that they found it highly motivating to know that their work evaluated existing academic findings that have been presented by professional researchers and that have been published in scientific journals.

analytical estimates. It is also noteworthy that concerns related to the research quality can be made for published research in general.

A relative lack of quality and control of classroom studies might become concerning when the classroom study fails to replicate a published finding. It is likely that the reliability of such replication failures will be contested. This might be an important limitation as to what we can learn from classroom replications. In our case study, our data are more variable than laboratory data. We suspect that this variability is due to inexperienced annotators/ experimenters and variable recording conditions. However, this variability is arguably not confounding the investigated relationship between devoiced and voiceless stops. Introduced variability is expected to affect both conditions to the same extent. These concerns, of course, will need to be evaluated on a case-by-case basis, but we do not believe that lack of control can be used as an argument against classroom replications across the board. However, dependent on the statistical framework researchers are operating in, a replication failure might be related to the noisiness of a classroom study. Uncertainty surrounding our interpretation, thus, must be quantified. Bayesian analyses can quantify the evidence for either the null or the alternative hypothesis and can inform future meta analyses in a straightforward way. Consequently, a failed replication attempt can quantify the extent to which we should shift our belief in the reliability of the original study, evaluated against the employed methodologies and the accumulated evidence from both the original and the replication study.

In addition to quality control, there are important limitations of the proposed classroom replications, which are particularly relevant for speech related research. Classroom experiments must stay within the technical abilities of the students, the practical limits of a university class, and the limits of accessible populations (see Franke & Saxe 2012, Hawkins et al. 2018 for similar arguments for psychological studies). There are many experimental procedures that warrant substantially more expertise to execute than speech recordings using a PowerPoint presentation. These limitations also often come with their own restrictions on technical abilities and equipment. While more complex experimental set-ups such as eye tracking and mouse tracking experiments might still be feasible to some extent, neurolinguistic experiments (e.g. using electroencephalography or functional magnetic resonance imaging) and articulatory studies (e.g. using ultrasound or electromagnetic articulography), are often beyond the resources of classroom replications. These techniques not only require access to expensive lab space and equipment, but also require substantial training and technical know-how. Additionally, student replications will be bounded by the temporal structure of a semester and the populations that are accessible to the students. This translates into another area of quantitative linguistics that will not be within the limits of classroom replications: Most field studies that require access to a particular speaker population and/or geographical locations will often not be feasible targets to replicate. We thus need to acknowledge that our proposed strategy will inevitably be biased toward a subset of behavioral research.

In summary, replications in the classroom will be restricted to easy studies that do not require complex experimental set ups, can be executed within a short period of time and be conducted on an accessible population. Again, it is worth iterating that these limitations are probably to some extent present in our published research due to the trade-off between resources and potential academic reward. However, we think it is fair to state that the majority of our theories are founded on rather straightforward speech production and perception studies, most of which may qualify for exactly these types of classroom experiments. We emphasize again that while the proposed method is not a single panacea for the present replication crisis, we firmly believe it is one possible step toward a more substantiated scientific record within experimental linguistics.

### 6.4 Where to go from here

The presented speech production studies were replications of published experiments. Although their protocols were following the original design very closely, they also differed in certain ways. While it can be argued that differences in the material lead to an ecologically more valid extension of the original findings, these differences can also be used to doubt the meaningfulness of the results in relation to the

original study (see Zwaan et al. 2018, for a recent discussion). In the future, it would be desirable to stick even closer to the original study. One could contact the original author and ask for their materials as well as requesting further insights about the procedure that go beyond the reported details from the published manuscript.

We could further improve our protocol by preregistering our replication attempt. A preregistration is a time-stamped document in which researchers specify exactly how they plan to collect their data and how they plan to conduct their confirmatory analyses. Preregistrations can be a powerful tool to reduce exploitation of researcher degrees of freedom (e.g. Simmons et al. 2011; Roettger 2019) because researchers are required to commit to certain decisions prior to observing the data. Additionally, public preregistration can at least help to reduce issues related to publication bias (Sterling 1959), as the number of failed attempts to reject a hypothesis can be tracked transparently. Preregistering our replication attempts (in the classroom) could further improve the quality of our methodology as the instructor can evaluate the written preregistration protocol prior to data collection, suggest changes to the proposed method, and review analysis code for errors.

## 7 Conclusion

The presented results demonstrate the practical possibility of performing replication research on speech phenomena in the classroom. Although associated with a certain lack of control and an increased level of noise in the data, replicating in the classroom offers an easy and inexpensive way to replicate relevant experimental findings in the literature. Beyond the possibility of substantiating our scientific record, this technique has numerous pedagogical advantages for our students.

Still, there are many challenges we need to pay close attention to. We need to find ways to ensure high-quality data. Close supervision of our students by an experienced instructor and preregistration of the methodological protocols can help ensure this quality. Moreover, due to practical considerations, replications in the classroom remain biased toward a subset of linguistic research, restricted to simple studies on accessible population of speakers. Despite these challenges and limitations, we believe that if this approach were implemented more widely across our field, the scientific and educational pay-off would be substantial.

## References

Allaire, J. J., Hadley Wickham, Kevin Ushey & Gary Ritchie. 2017. *rstudioapi: Safely Access the RStudio API*. https://CRAN.R-project.org/package=rstudioapi.

Audacity Team. 2015. *Audacity (R): Free audio editor and recorder [computer application]*. https://audacityteam.org/.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

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

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth & Harald Baayen. 2015. Parsimonious mixed models. https://arxiv.org/abs/1506.04967.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Behrend, Tara S., David J. Sharek, Adam W. Meade & Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43(3). 800–813.

Boersma, Paul & David Weenink. 2015. *Praat: Doing phonetics by computer [Computer program]*.

Boroditsky, Lera. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology* 43(1). 1–22. https://doi.org/10.1006/cogp.2001.0748.

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

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, et al. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351(6280). 1433–1436. https://doi.org/10.1126/science.aaf0918.

Campbell, Donald T. 1969. Reforms as experiments. *American Psychologist* 24(4). 409.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1).

Charles-Luce, Jan. 1985. Word-final devoicing in German and the effects of phonetic and sentential contexts. *The Journal of the Acoustical Society of America* 77. s85. https://doi.org/10.1121/1.2022551.

Chen, Jenn-Yeu. 2007. Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). *Cognition* 104. 427–436. https://doi.org/10.1016/j.cognition.2006.09.012.

Clarke, Erik & Scott Sherrill-Mix. 2017. *Ggbeeswarm: Categorical scatter (violin point) plots*. https://CRAN.R-project.org/package=ggbeeswarm (18 September, 2018).

DeLong, Katherine A., Thomas P. Urbach & Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience* 8(8). 1117–1121.

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

Dinnsen, Daniel A. & Maria Garcia-Zamor. 1971. The three degrees of vowel length in German. *Paper in Linguistics* 4(1). 111–126. https://doi.org/10.1080/08351817109370250.

Doyen, Stéphane, Olivier Klein, Cora-Lise Pichon & Axel Cleeremans. 2012. Behavioral priming: It's all in the mind, but whose mind? *PLOS ONE* 7(1). e29081. https://doi.org/10.1371/journal.pone.0029081.

Dunlap, Knight. 1925. The experimental methods of psychology. *The Pedagogical Seminary and Journal of Genetic Psychology* 32(3). 502–522.

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

Errington, Timothy M., Elizabeth Iorns, William Gunn, Fraser Elisabeth Tan, Joelle Lomax & Brian A. Nosek. 2014. Science forum: An open investigation of the reproducibility of cancer biology research. *eLife* 3. e04333.

Fine, Alex B., T. Florian Jaeger, Thomas A. Farmer & Ting Qian. 2013. Rapid expectation adaptation during syntactic comprehension. *PLOS ONE* 8(10). e77661. https://doi.org/10.1371/journal.pone.0077661.

Fourakis, Marios & Gregory K. Iverson. 1984. On the 'incomplete neutralization' of German final obstruents. *Phonetica* 41(3). 140–149.

Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, et al. 2017. A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy* 22(4). 421–435. https://doi.org/10.1111/infa.12182.

Frank, Michael C. & Rebecca Saxe. 2012. Teaching replication. *Perspectives on Psychological Science* 7(6). 600–604. https://doi.org/10.1177/1745691612460686.

Fuchs, Susanne. 2005. *Articulatory correlates of the voicing contrast in alveolar obstruent production in German* (ZAS Papers in Linguistics 41). Berlin: Zentrum für Allgemeine Sprachwissenschaften.

Gelman, Andrew & John Carlin. 2014. Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9(6). 641–651. https://doi.org/10.1177/1745691614551642.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su & others. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2(4). 1360–1383. https://doi.org/10.1214/08-AOAS191.

Gelman, Andrew & Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University.

Glenberg, Arthur M. & Michael P. Kaschak. 2002. Grounding language in action. *Psychonomic Bulletin & Review* 9. 558–565.

Goldrick, Matthew, Jocelyn R. Folk & Brenda Rapp. 2010. Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language* 62(2). 113–134. https://doi.org/10.1016/j.jml.2009.11.008.

Goodman, Joseph K., Cynthia E. Cryder & Amar Cheema. 2012. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making* 26(3). 213–224. https://doi.org/10.1002/bdm.1753.

Gorgolewski, Krzysztof J., Thomas Nichols, David N. Kennedy, Jean-Baptiste Poline & Russell A. Poldrack. 2018. Making replication prestigious. *Behavioral and Brain Sciences* 41. E131. https://doi.org/10.1017/S0140525X18000663.

Grahe, Jon, Mark Brandt, Hans IJzerman & Johanna Cohoon. 2014. Replication education. *APS Observer* 27(3).

Grahe, Jon, Mark Brandt, Jordan Wagge, Nicole Legate, Bradford Wiggins, Cody Christopherson, Yanna Weisberg, et al. 2018. Collaborative replications and education project (CREP). https://osf.io/wfc6u/ (18 September, 2018).

Hasegawa-Johnson, Mark, Jennifer Cole, Preethi Jyothi & Lav R. Varshney. 2015. Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications. *Laboratory Phonology* 6(3–4). 381–431. https://doi.org/10.1515/lp-2015-0012.

Hawkins, Robert X., Eric N. Smith, Carolyn Au, Juan Miguel Arias, Rhia Catapano, Eric Hermann, Martin Keil, Andrew Lampinen, Sarah Raposo & Jesse Reynolds. 2018. Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science* 1(1). 7–18. https://doi.org/10.1177/2515245917740427.

Hewitt, John K. 2012. Editorial policy on candidate gene association and candidate gene-by-environment interaction studies of complex traits. *Behavior Genetics* 42(1). 1–2. https://doi.org/10.1007/s10519-011-9504-z.

Hirsh, Ira J. 1959. Auditory perception of temporal order. *The Journal of the Acoustical Society of America* 31(6). 759–767. https://doi.org/10.1121/1.1907782.

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

Ioannidis, John PA. 2005. Why most published research findings are false. *PLoS medicine* 2(8). e124.

Iversen, John R., Aniruddh D. Patel & Kengo Ohgushi. 2008. Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America* 124(4). 2263–2271. https://doi.org/10.1121/1.2973189 .

Jespersen, Otto. 1920. *Lehrbuch der Phonetik: Mit 2 Tafeln*. (Trans.) Hermann Davidsen. Leipzig: BG Teubner.

Jessen, Michael & Catherine Ringen. 2002. Laryngeal features in German. *Phonology* 19(2). 189–218. https://doi.org/10.1017/S0952675702004311.

Kirby, James & Morgan Sonderegger. 2018. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* 70. 70–85. https://doi.org/10.1016/j.wocn.2018.05.005.

Klatt, Dennis 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. https://doi.org/10.1121/1.380986.

Kochari, Arnold R. & Markus Ostarek. 2018. Introducing a replication-first rule for Ph.D. projects. *Behavioral and Brain Sciences* 41. E138. https://doi.org/0.1017/S0140525X18000730.

Kohler, Klaus J. 2007. Beyond laboratory phonology – The phonetics of speech communication. In Maria-Josep Solé, Patrice Speeter Beddor & Manjari Ohala (eds.), *Experimental approaches to phonology*. 41–53. New York: Oxford University Press.

Kohler, Klaus J. 2012. Neutralization?! The phonetics–phonology issue in the analysis of word–final obstruent voicing. In Dafydd Gybbon, Daniel Hirst & Nick Campbell (eds.), *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem*. 171–180. Poznań: Polskie Towarzystwo Fonetyczne.

Koole, Sander L. & Daniël Lakens. 2012. Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science* 7(6). 608–614.

Ladefoged, Peter. 2003. *Phonetic data analysis: An introduction to fieldwork and instrumental techniques*. Malden, MA: Blackwell.

Lehiste, Ilse. 1976. Suprasegmental features of speech. In Norman J. Lass (ed.), *Contemporary issues in experimental phonetics*. vol. 225, 225–239. New York: Academic Press.

Lisker, Leigh. 1986. 'Voicing' in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech* 29(1). 3–11. https://doi.org/10.1177/002383098602900102.

Madden, Charles S., Richard W. Easley & Mark G. Dunn. 1995. How journal editors view replication research. *Journal of Advertising* 24(4). 77–87. https://doi.org/10.1080/00913367.1995.10673490.

Makel, Matthew C., Jonathan A. Plucker & Boyd Hegarty. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science* 7(6). 537–542.

Manaster Ramer, Alexis. 1996. A letter from an incompletely neutral phonologist. *Journal of Phonetics* 24(4). 477–489. https://doi.org/10.1006/jpho.1996.0026.

Martin, G. N. & Richard M. Clarke. 2017. Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology* 8. 523. https://doi.org/10.3389/fpsyg.2017.00523.

Mason, Winter & Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1). 1–23. https://doi.org/10.3758/s13428-011-0124-6.

Miller, James D. 1989. Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America* 85(5). 2114–2134. https://doi.org/10.1121/1.397862.

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

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

Nicenboim, Bruno & Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass* 10. 591–613. https://doi.org/10.1111/lnc3.12207.

Nicenboim, Bruno, Shravan Vasishth & Frank Rösler. 2019. Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. https://doi.org/10.31234/osf.io/2atrh.

Nieuwland, Mante S., Stephen Politzer-Ahles, Evelien Heyselaar, Katrien Segaert, Emily Darley, Nina Kazanina, et al. 2018. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife* 7. e33468. https://doi.org/10.7554/eLife.33468.001.

Nosek, Brian A. & Timothy M. Errington. 2017. Reproducibility in Cancer Biology: Making sense of replications. *eLife* 6. e23383. https://doi.org/10.7554/eLife.23383.

Nosek, Brian A., Jeffrey R. Spies & Matt Motyl. 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). https://doi.org/10.1126/science.aac4716.

Papesh, Megan H. 2015. Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General* 144. e116–e141. https://doi.org/10.1037/xge0000125.

Phillips, Susan L., Sandra Gordon-Salant, Peter J. Fitzgibbons & Grace H. Yeni-Komshian. 1994. Auditory duration discrimination in young and elderly listeners with normal hearing. *Journal of the American Academy of Audiology* 5. 210–215.

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

Port, Robert F. & Penny Crawford. 1989. Incomplete neutralization and pragmatics in German. *Journal of Phonetics* 17. 257–282.

Port, Robert F. & Michael L. O'Dell. 1985. Neutralization of syllable-final voicing in German. *Journal of Phonetics* 13(4). 455–471.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1. https://doi.org/10.5334/labphon.147.

Roettger, Timo B., Bodo Winter, Sven Grawunder, James Kirby & Martine Grice. 2014. Assessing incomplete neutralization of final devoicing in German. *Journal of Phonetics* 43. 11–25. https://doi.org/10.1016/j.wocn.2014.01.002.

Roettger, Timo & Matthew Gordon. 2017. Methodological issues in the study of word stress correlates. *Linguistics Vanguard* 3(1). https://doi.org/10.1515/lingvan-2017-0006.

Schielzeth, Holger & Wolfgang Forstmeier. 2008. Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology* 20(2). 416–420. https://doi.org/10.1093/beheco/arn145.

Seyfarth, Scott, Marc Garellek, Gwendolyn Gillingham, Farrell Ackerman & Robert Malouf. 2018. Acoustic differences in morphologically-distinct homophones. *Language, Cognition and Neuroscience* 33(1). 32–49. https://doi.org/10.1080/23273798.2017.1359634.

Shapiro, Danielle N., Jesse Chandler & Pam A. Mueller. 2013. Using mechanical turk to study clinical populations. *Clinical Psychological Science* 1(2). 213–220.

Silberzahn, Raphael, Eric Luis Uhlmann, Dan P Martin, Pasquale Anselmi, Johannes Ullrich, Frederik Aust, Eli C. Awtrey, Štěpán Bahník, Feng Bai & Colin Bannard. 2017. Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.5167/uzh-148470.

Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11). 1359–1366. https://doi.org/10.1177/0956797611417632.

Simons, Daniel J., Alex O. Holcombe & Barbara A. Spellman. 2014. An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspectives on Psychological Science* 9(5). 552–555. https://doi.org/10.1177/1745691614543974.

Sluijter, Agaath M. C., Vincent J. van Heuven & Jos J. A. Pacilly. 1997. Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America* 101(1). 503–513. https://doi.org/10.1121/1.417994.

Srivastava, Sanjay. 2012. A Pottery Barn rule for scientific journals. *The Hardest Science*. https://thehardestscience.com/2012/09/27/a-pottery-barn-rule-for-scientific-journals/ (22 March, 2019).

Stack, Caoimhe M. Harrington, Ariel N. James & Duane G. Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition* 46(6). 864–877. https://doi.org/10.3758/s13421-018-0808-6.

Sterling, Theodore 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. https://doi.org/10.1080/01621459.1959.10501497.

Taylor, Dennis Q. 1975. The inadequacy of bipolarity and distinctive features: The German "voiced/voiceless" consonants. *The Second LACUS Forum*. 107–119.

Trubetzkoy, Nikolaj Sergeevič. 1939. *Grundzüge der Phonologie*. Vol. 7. Prague: Travaux du Cercle Linguistique de Prague.
van Oostendorp, Marc. 2008. Incomplete devoicing in formal phonology. *Lingua* 118(9). 1362–1374.
Vasishth, Shravan, Daniela Mertzen, Lena A. Jäger & Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language* 103. 151–175.
Vasishth, Shravan & Bruno Nicenboim. 2016. Statistical methods for linguistic research: Foundational ideas - Part I. *Language and Linguistics Compass* 10(8). 349–369. https://doi.org/10.1111/lnc3.12201.
Wager, Tor D., Martin A. Lindquist, Thomas E. Nichols, Hedy Kober & Jared X. Van Snellenberg. 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* 45(1). S210–S221. https://doi.org/10.1016/j.neuroimage.2008.10.061.
Wagner, Petra, Jürgen Trouvain & Frank Zimmerer. 2015. In defense of stylistic diversity in speech research. *Journal of Phonetics* 48. 1–12. https://doi.org/10.1016/j.wocn.2014.11.001.
Westbury, C. 2005. Implicit sound symbolism in lexical access: Evidence from an interference task. *Brain and Language* 93(1). 10–19. https://doi.org/10.1016/j.bandl.2004.07.006.
Westbury, Chris. 2018. Implicit sound symbolism effect in lexical access, revisited: A requiem for the interference task paradigm. *Journal of Articles in Support of the Null Hypothesis* 15(1). 1–12.
Wickham, Hadley. 2017. *Tidyverse: Easily install and load 'tidyverse' packages*. https://CRAN.R-project.org/package=tidyverse.
Wiese, Richard. 1996. *The phonology of German* (The Phonology of the World's Languages). Oxford: Clarendon Press.
Winter, Bodo. 2011. Pseudoreplication in phonetic research. *Proceedings of the International Congress of Phonetic Science*. 2137–2140. Hong Kong.
Winter, Bodo. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow: The University of Glasgow.
Winter, Bodo & Timo Roettger. 2011. The nature of incomplete neutralization in German: Implications for laboratory phonology. *Grazer Linguistische Studien* 76. 55–74.
Xu, Yi. 2010. In defense of lab speech. *Journal of Phonetics* 38(3). 329–336. https://doi.org/10.1016/j.wocn.2010.04.003.
Zwaan, Rolf A., Alexander Etz, Richard E. Lucas & M. Brent Donnellan. 2018. Making replication mainstream. *Behavioral and Brain Sciences* 41. E120. https://doi.org/10.1017/S0140525X17001972.

Timo B. Roettger
Department of Linguistics
Northwestern University
2016 Sheridan Rd
Evanston, IL 60208, USA
timo.b.roettger@gmail.com

Dinah Baer-Henney
Institut für Sprache und Information
Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1
D-40225, Düsseldorf, Germany
dinah.baer-henney@uni-duesseldorf.de