Journal of Phonetics 73 (2019) 1–7

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility

Timo B. Roettger [a,*], Bodo Winter [b], Harald Baayen [c]

[a] Northwestern University, Department of Linguistics, United States
[b] University of Birmingham, Department of English Language and Applied Linguistics, United Kingdom
[c] University of Tübingen, Department of Linguistics, Germany

ARTICLE INFO

Article history:
Received 6 April 2018
Received in revised form 10 December 2018
Accepted 11 December 2018

Keywords:
Data analysis
Statistics
Reproducibility
Open science
Null hypothesis significance testing
Bayesian modeling

ABSTRACT

This special issue introduces a series of papers that make available new methods to the phonetic and linguistic community and reflect upon existing data analysis practices. In our introduction, we highlight three themes that we consider pressing issues in data analysis and that run across the contributions to this special issue: the difference between exploratory and confirmatory analyses, different approaches to statistical inference, and the analysis of multidimensional multivariate speech data. Moreover, we provide a call for considering the importance of open and reproducible research practices, such as publishing one's data and analysis code. Rather than being dogmatic about particular statistical methods, the pluralism of analysis approaches in linguistics should excite debate and discussion, to which this special issue is an invitation. In addition, the co-existence of multiple ways of analyzing the same data (each with its own advantages and disadvantages and different analysis goals) makes it all the more important for researchers to make their research process open and accessible to other researchers.

Published by Elsevier Ltd.

1. Introduction

The landscape of data analysis in linguistics and other fields is constantly changing. Advances in computational power have made new analytical approaches possible, and the use of open access software such as R (R Core Team, 2013) increases the speed with which new statistical methods are shared both within our field and across disciplines. As accessibility to these methods increases, more and more people within linguistics employ increasingly complex analytical techniques. Parallel to the ever-growing toolkit of statistical methods, there are shifts in methodological traditions and statistical philosophies, with an array of differing views about how data should be analyzed, how it should be reported, and how it should be shared. In sum, the field of data analysis is in flux. Amidst the backdrop of changing practices, it is important to critically assess past practices, to reflect upon present practices, and to look out for what new developments will affect our future practices.

We approach data analysis with George Box's famous quote in mind, "all models are wrong, but some are useful" (Box, 1979, p. 2). This often-repeated quote embodies a fundamental truth about data analysis: We perform analyses to gain a better understanding of our world and the phenomena we investigate. Statistical models are thus supposed to be "useful". However, all models are also necessarily "wrong" to some extent, with each model providing only a snapshot of the underlying complexity of the phenomena to be modeled. Models can be "useful" in different ways and to differing degrees, and models can be more or less "wrong" as well. There is no single model that is the best model and that is equally useful across theories and phenomena. This very fact necessarily creates a *plurality* of analytical approaches, within and across disciplines. Even expert statisticians reach different conclusions when given the same dataset (Silberzahn et al., 2018). Rather than trying to provide gold standards and recipes, we endorse the plurality of approaches and highlight that pluralism calls for comparison, reflection, and a critical discourse about methods. We should not try to elevate any one method to the status of a "best" method or a canonical way of analyzing particular datasets; instead we should discuss the advantages and disadvantages of particular approaches openly.

In line with the idea of plurality, data analysis varies along important dimensions. We would like to highlight a few of these

* Corresponding author.
E-mail address: timo.roettger@uni-koeln.de (T.B. Roettger).

https://doi.org/10.1016/j.wocn.2018.12.001
0095-4470/Published by Elsevier Ltd.

dimensions to not only introduce the contributions to our special issue, but also to review what we conceive as important topics for data analysis in quantitative fields such as phonetics. In the following, we will discuss the distinction between exploratory and confirmatory analysis (Section 2); the differences between null hypothesis significance testing and Bayesian inference (Section 3); and analytical choices surrounding the multidimensionality of phonetic data (Section 4). Beyond reflecting on past and future methods it is also important to think about how data analyses are communicated and shared with the community. To this end, we will discuss the relevance of reproducibility and the benefits of an open and transparent phonetic community (Section 5), exemplified by the contributions in this special issue.

2. Exploratory vs. confirmatory data analysis

It is important to recognize that data analysis includes two stages which are more or less conceptually distinct, although they may overlap to considerable degrees in practice. In an *exploratory* stage, a researcher observes patterns and relationships leading to the generation of new hypotheses as to how these observations can be explained. This stage is a hypothesis-generating process. Many breakthroughs in science originate from the serendipity of researchers observing an unexpected pattern while exploring their data. In a *confirmatory* stage, novel hypotheses as well as hypotheses extending or challenging established theories are then pitted against new data, obtained in, for example, controlled experimental studies. This stage is a hypothesis-testing process. Putting our hypotheses under targeted scrutiny via confirmatory tests helps us to accumulate evidence in order to challenge, substantiate or revise established theories. The revised theories can then be further informed by additional exploration of the available data, leading to an iterative process that alternates between exploration and confirmation. Exploratory and confirmatory research should be considered complementary; both are necessary components of scientific progress. Moreover, both exploratory and confirmatory research should be guided by theory. An exploratory analysis does not have to be exclusively descriptive, but can, and often should be, tied in with specific linguistic theories.

The distinction between confirmation and exploration has large-scale consequences for research in the language sciences. It is important to realize that in an exclusively confirmatory setting, researchers have only one shot (Harrell, 2014), allowing for only a single theoretically motivated model to be fitted to the data. Subsequently, model criticism is carried out to clarify whether the resulting model is actually appropriate for the data. In a genuinely confirmatory analysis, there is no place for repeated modeling during data collection, no place for adding or removing interactions, and no place for including or removing control variables. As soon as a second model is fitted to a given dataset, the analysis is no longer confirmatory, but exploratory (see Baayen, Vasishth, Kliegl, & Bates, 2017, for further discussion).

Unfortunately, when it comes to publishing work, exploration and confirmation are not weighted equally. Confirmatory analyses have a superior status within the academic incentive system, determining the way funding agencies demand what proposals should look like, and shaping how we frame our papers. The prestige of confirmatory statistics is so high that occasionally the review process can force authors to recast the reporting of exploratory analyses in the format of the reporting of confirmatory analyses (see, e.g., Pham & Baayen, 2015, footnote 1). Whether due to publication pressure or not, the results of what has actually been an exploratory analysis are often presented as if they were the results of a confirmatory analysis. The prevalent expectation that the main results of a study should be predicted based on a priori grounds has led to harmful practices for scientific progress (John, Loewenstein, & Prelec, 2012).

Moreover, each analysis is characterized by a “garden of forking paths” (Gelman & Loken, 2013) or what Simmons, Nelson, and Simonsohn (2011) call “researcher degrees of freedom”. Some relevant researcher degrees of freedom for phonetic studies include what phonetic parameters are measured, how they are operationalized, what data is kept and what data is discarded and what additional independent variables are measured (for a discussion of researcher degrees of freedom in phonetics, see Roettger, 2019). This flexibility in conducting studies and analyzing data can, intentionally or unintentionally, lead to harmful practices such hunting for significant *p*-values, also known as *p*-hacking (see also Simmons et al., 2011) or HARKing “Hypothesizing After Results are Known” (e.g., Kerr, 1998).

Rather than discouraging exploratory analyses, they should be encouraged. The complexity of speech naturally means that we do not always have specific directed hypotheses for all aspects of the data. There are many interesting patterns to be discovered, and later confirmed on separate datasets. It is often the exploratory part of the analysis that we can learn the most from, especially with highly multidimensional data (see Section 4). However, while exploration is necessary, it has to be separated from confirmation. Each analysis needs to be clear about where it stands, i.e., the degree to which an analysis is confirmatory or exploratory needs to be explicitly stated. In particular, exploratory studies should be treated as such, rather than being re-framed as the results of a confirmatory analysis. More and more papers in our field acknowledge this important distinction and discuss confirmatory and exploratory analyses in different sections of their manuscripts, with the latter stressing the caveat that any generated hypotheses are waiting to be confirmed on new data (e.g., Baumann & Winter, 2018; Grice, Savino, & Roettger, 2018 for recent examples)

Researchers carrying out exploratory data analysis can to some extent protect themselves and their colleagues against spurious results by setting much more stringent alpha-levels when evaluating whether there are signals in the noise. In exploratory analysis, it is the researcher’s duty to launch adversarial attacks on potential effects, and to then only report those effects which survived such attacks consistently. If a strict null hypothesis significance testing approach is followed, confirmation cannot happen on the same dataset that was previously used as the basis for exploration. To the extent that confirmatory and exploratory analyses may blend into each other in actual practice, the researcher needs to be aware of this and report results accordingly.

3. Inferential frameworks: Frequentist and Bayesian inference

An important aspect of data analysis is making generalizable statements about observations. Inferential statistics is the process of using samples to make "inferences" for parameters of a population of interest. For example, a study may contain a subset of speakers from a linguistic community, and the sample is used to make inferences about all speakers of the language. Or a study may contain a subset of words from the language, and the sample is used to make inferences about all words of the language (see Clark, 1973). In statistics, there are various different approaches to making this inference, including frequentist and Bayesian statistics (e.g. Fisher, 1955; Gigerenzer, Krauss, & Vitouch, 2004; Dienes, 2008; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; McElreath, 2016). Each approach has different analysis goals and makes different assumptions. Our special issue includes several papers that discuss aspects of different inferential frameworks as well as papers that make use of techniques and methods developed within each of these frameworks.

Classical methods for statistical inference (analysis of variance, discriminant analysis) are grounded in the work of Sir Ronald Fisher (1925). These methods, which are widely used in phonetics and many other fields of inquiry, are known as frequentist, as they are grounded in a particular understanding of the concept of probability, namely, the idea that the probability of an event is given by the limit of its relative frequency across a large number of trials. Fisher's method was later combined with Neyman and Pearson's approach to hypothesis testing (1928) to create what is now known as null hypothesis significance testing (NHST) (Gigerenzer et al., 2004; Lindquist, 1940). This framework became an extremely useful tool at a time in which computers did not exist, and has been used ever since across scientific disciplines. In traditional NHST, a researcher starts by assuming a null hypothesis (such as the absence of an effect) and gathers evidence against that initial assumption. The *p*-value measures the incompatibility of the data with the null hypothesis. It is often used as a hard cut-off, where an effect is accepted as "significant" if its associated *p*-value falls below a preset threshold probability. NHST provides a simple and specific decision procedure (using a particular threshold, such as $p < 0.05$) which will assure low error rates in the long run, across a series of repeated experiments.

The practice of NHST has been much criticized by researchers in many different disciplines (Gigerenzer, 2004; Goodman, 1999; Hubbard & Lindsay, 2008; Kline, 2004; Krantz, 1999; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016; Nickerson, 2000; Sterne & Smith, 2001; and many others). These criticisms surround, among other things, the practice of relying on an arbitrarily defined hard-cut threshold for "significance" (rather than taking the continuous strength of evidence into account), the practice of overly emphasizing point estimates (such as means) over interval estimates (such as confidence intervals or credible intervals) (e.g., Cumming, 2012, 2014), and the practice of not incorporating any prior knowledge into one's models and inferences.[1]

Frequentist inference, as introduced by Fisher, differs in many ways from what is now known as Bayesian inference. The field of statistics has a long history of a deep divide between classical frequentist statistics and Bayesian statistics, each camp having its own philosophical foundations and methodological goals (e.g., Fisher, 1955; Gigerenzer et al., 2004; Dienes, 2008; Wagenmakers et al., 2008).

There are different classes of Bayesian models, but one defining feature is that they quantify the degree to which a researcher needs to adjust their beliefs as a function of the researcher's prior beliefs and the data and model at hand. That is, Bayesian inference critically differs from other inferential approaches by incorporating so-called "priors", which are either defined by a priori assumptions about the measurement system, or estimated from previous research. For example, when estimating the difference in duration between two vowel categories, the researcher can incorporate priors which reasonably rule out durational values below zero and above one hour. As opposed to that, in standard frequentist inference, all parameter values are assumed to be equally likely. Bayesian inference makes it also possible for the analyst to include knowledge outside of the present data when modeling new data, e.g. estimated from previous research. The Bayesian paradigm has, intrinsically, a much more cumulative perspective on the gathering of scientific evidence and offers much more sophisticated tools for integrating knowledge across multiple studies.

That the need for clear 'gold standards' is still felt today is exemplified by the paper by Barr, Levy, Scheepers, and Tily (2013) on how to fit mixed models. With the wide spectrum of analytical techniques currently available, which will also increasingly include methods from machine learning, it is not possible nor desirable to enforce rules by means of which significance can be assessed mechanically. A spirit of plurality is needed that creates space for realizing that there are problems and applications that might be handled more easily by either Bayesian or frequentist approaches, and that the analyst is far better off having both tools in their toolbox. In particular, researchers need to be familiar with both approaches, since there is an increasing number of papers in quantitative linguistics that uses Bayesian approaches.

There are three papers in our special issue that focus on the merits and pitfalls of different statistical philosophies (such as NHST versus Bayesian inference). Vasishth et al. (2018, this collection) give an extended overview of the logic and benefits of standard Bayesian analyses and walk the reader through a concrete standard Bayesian analysis of an acoustic study, investigating whether and how voice onset time measurements discriminate different stop series across three different languages. Their paper provides a useful introduction to Bayesian data analysis in linguistics and offers annotated code to facilitate the implementation of Bayesian modeling.

In a second paper, Nicenboim, Roettger, and Vasishth (2018, this collection) investigate the phenomenon of incomplete neutralization of German final devoicing using Bayesian meta-analysis. Incomplete neutralization is a particularly valuable phenomenon to discuss methodologically, because the available evidence has been the subject of heated methodological debates (Fourakis & Iverson, 1984; Port & O'Dell, 1985; Roettger, Winter, Grawunder, Kirby, & Grice, 2014;

[1] We note here that classical 'frequentist' inference is not necessarily or intrinsically focused on hard cut-offs, which is only a particular interpretation of this framework (see Perezgonzalez, 2015).

Winter & Roettger, 2011). According to some researchers, the German voicing contrast is completely devoiced in final position; according to others, the devoicing is phonetically "incomplete". A number of studies in this literature do not allow adequate statistical inferences because the sample sizes are too small, and hence accumulating the evidence *across* studies in a meta-analysis becomes crucial to establish whether incomplete neutralization effects are robust.

The controversial topic of incomplete neutralization is also explored in another paper that addresses issues in statistical inference. Kirby and Sonderegger (2018, this collection) look at the role of sample size in being able to estimate the incomplete neutralization effect accurately. Their numerical simulations suggest that linguists need to pay more attention to statistical power (the probability that a significance test will correctly reject a false null hypothesis) in designing experiments. Small sample sizes come with unrealistic expectations of replicability of the effect direction and magnitude (e.g. Vasishth, Mertzen, Jäger, & Gelman, 2018, for a recent discussion). Besides making important points about experimental design in phonetics, Kirby and Sonderegger (2018, this collection) demonstrate the utility of performing power simulations.

We want to stress here that including papers on either NHST or Bayesian inference is not intended to suggest that analyzing data within either of these frameworks is right or wrong. Given the prevalence of the decision procedure of NHST within phonetics, we think that it is most prudent at this stage to be aware of the opportunities offered by Bayesian and frequentist approaches. Moreover, learning about standard Bayesian methods may also help clarify misunderstandings about NHST (see Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Wagenmakers, Morey, & Lee, 2016; Morey et al., 2016; Nicenboim et al., 2018).

4. Dealing with the multidimensionality of speech communication

Our choice of data analysis varies tremendously as a function of the way phenomena are observed and measured. Depending on how observations are operationalized, certain analytical tools may or may not apply. Speech is inherently multidimensional and varies across time, as is the case with pitch curves, formant trajectories or articulatory gestures. These time-series data can be analyzed as a sequence of static landmarks ("magic moments", Vatikiotis-Bateson, Barbosa, & Best, 2014) or as continuous trajectories, depending on how relevant the dynamic nature of speech behavior is for any given theory (Mücke, Grice, & Cho, 2014).

This special issue includes an introduction to Generalized Additive Models (GAMs), which are an extension of the classical generalized linear model (GLM) that enjoys wide use within phonetics (e.g., multiple regression, logistic regression, linear mixed effects models). Even traditional tests, such as *t*-tests and ANOVAs are approaches that can be re-expressed in a regression framework, in which case they yield equivalent results (if appropriately specified). GAMs extend GLMs with methods for modeling smooth nonlinear functions between a response and one or more predictors (Winter & Wieling, 2016; Wood, 2006). They also offer tools for addressing autocorrelations in the residual error, which are often present in time-series data, i.e. when observations are ordered in time, current observations may depend on previous observations.

Wieling (2018, this collection) introduces GAMs and offers a step-by-step tutorial based on an analysis of articulatory data (for other introductions, see, e.g., Winter & Wieling, 2016, and Baayen et al., 2017). As with any new tool, it is not always clear what the best approach to using this tool is from the outset. This was the case with linear mixed effects models, which incited a prominent debate about what the best random effects structure for the analysis of experimental designs is (see Barr et al., 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The flexibility inherent to statistical modeling is amplified in the case of GAMs, which provide many more options to their users. Wieling discusses some of these options.

An important complementary aspect of multidimensionality is tackled by Tomaschek, Hendrix, and Baayen (2018, this collection). They deal with a common problem in regression analyses (and by extension mixed models, GAMs etc.), namely, the issue of collinearity. When predictor variables in a model are highly correlated, estimates of parameters may become unstable and researchers can easily draw the wrong conclusions based on their data. Collinearity is an important problem that is often overlooked. As stated by Zuur, Leno, and Elphick (2010, p. 9): "If collinearity is ignored, one is likely to end up with a confusing statistical analysis in which nothing is significant, but where dropping one covariate can make the others significant, or even change the sign of estimated parameters." Tomaschek et al. provide a critical discussion of three methods developed specifically for the analysis of data sets with many correlated predictors - regularization with the elastic net, regularization with supervised component regression, and random forests - each of which has its strengths and weaknesses, depending on the goals of the analysis.

Plummer and Reidy (2018, this collection) discuss another issue related to the multidimensionality of phonetic data analysis. They discuss a method for computing low-dimensional representations of speech which centers on the use of Laplacian Eigenmaps to build structures over data points from which low-dimensional representations of speech are learned. This technique enables researchers to reduce the multidimensional acoustic signal to lower dimensionality, which, as they argue, is a better proxy of cognitive and social speech categories.

Another aspect of multidimensionality is tackled by Danner, Barbosa, and Goldstein (2018, this collection), discussing topics related to the non-verbal context in which speech occurs. Speech communication is accompanied by changes in body posture, head position, gaze, facial expressions, and manual gestures (Goldin-Meadow, 2003; Kendon, 2004; McNeill, 1992). Danner et al. invite the reader to rethink how to characterize multimodal speech by applying dynamic approaches already used in speech research to multimodal communication. They discuss both the problem of automatically identifying visual gestures in video images, as well as the problem of correlating a gestural data stream with an acoustic data stream.

The papers discussed so far are either focused on the merits and pitfalls of different statistical philosophies (such as NHST versus Bayesian inference), or they discuss various

new methods that are useful for different phonetic applications. Another strand that runs across the entire special issue is the issue of *reproducibility*. Reflecting on methods does not end with choosing a particular method, but it also includes thinking about how data analyses are communicated and shared with the community.

5. Towards reproducible phonetic sciences

To assess the strength of evidence for a theory, one needs to consider how the data were collected and how they were analyzed. Evaluating the strength of evidence becomes very difficult if part of the research process is not transparent. Reproducible research involves the capacity of other researchers (who have not conducted the original study) to repeat the analysis that is presented in a published study (see Peng, 2011; Munafò et al., 2017). Reproducibility minimally necessitates that both the data (either raw data or data tables) and the analysis code are made available to the community (if this is possible). Following recent calls for more transparent scientific practices (e.g. the Open Science Framework, see Nosek, 2017), we want to reiterate the plea for more reproducibility within phonetics in particular, and within linguistics more generally.

For our field, reproducible research has numerous advantages. First, as mentioned above, even expert data analysts will perform different analyses based on the same dataset (Silberzahn et al., 2018). Naturally, different analysis choices yield different conclusions (Roettger, 2019; Gelman & Loken, 2013; Simmons et al., 2011). McElreath (2016) emphasizes that statistical modeling is subjective, in the sense that it incorporates the researcher's beliefs and assumptions about a study system. Because of its inherent flexibility and subjectivity, the only way to allow evaluation of the process of statistical modeling by outsiders is to make it open.[2] Transparency then allows other researchers to draw their own conclusions based on the same dataset, reanalyze other aspects of them etc.

From a practical stand point, sharing materials, data, and code publicly has several applied advantages (Houtkoop et al., 2018). For example, data sharing has been associated with a citation benefit (Piwowar & Vision, 2013). Moreover, sharing data on online repositories can be a safeguard against 'scooping' (Houtkoop et al., 2018) since a researcher can claim precedence for a dataset or an analysis before a paper is published. In addition, permanently accessible repositories protect against data loss and link rot. Open research practices have furthermore shown to increase visibility, as well as to increase the number of opportunities for funding, jobs, and collaborations (McKiernan et al., 2016). If we make our materials and code available, the next research group (or our own) might have an easier time to replicate our experiment or extend our findings without duplicating efforts. This saves valuable resources and allows for a more rapid advancement of our field.

Publishing the data and code also facilitates knowledge transfer: Other researchers can learn from the ways a particular dataset was analyzed, and how the analysis was implemented in actual software code. It is within the spirit of sharing knowledge and being transparent, that all authors of this special issue make their code and data available on public repositories, allowing the readership of the special issue to readily implement the methods, as well as to actively participate in the discourse that surrounds the methods presented here. Reproducibility runs as a prominent thread through all of the papers in this special issue. All papers in this special issue contain links to publicly available repositories.

Many of the papers are written in a tutorial-like way, inviting the reader to reproduce and extend the offered analyses (Jadoul, Thompson, & de Boer, 2018; Vasishth et al., 2018; Wieling, 2018). For example, Politzer-Ahles and Piccinini (2018, this collection) discuss ways to visualize the results of hierarchical models that allows one to communicate the population-level estimates alongside the random variation associated with crossed random effects. Data visualization is an important aspect of communicating research findings and has been the subject of ongoing debates across scientific fields (e.g., Tufte, 1990; Kosslyn, 2006; Weissgerber, Milic, Winham, & Garovic, 2015). Politzer-Ahles and Piccinini's paper not only serves as a reminder of the importance of data visualization in communicating data and the results of statistical models; the inclusion of their scripts allows other users to apply them to new datasets.

The topic of reproducibility is also a prominent theme for Jadoul et al. (2018, this collection). As argued by many proponents of reproducible research, *all* aspects of the research workflow interact with reproducibility, not just the "final" data analysis stage. For example, in acoustic analyses, there are many degrees of freedom as to what acoustic parameters to extract and how, such as the settings used for the measurements of a particular speaker's fundamental frequency. We usually perform these analyses in available software such as Praat (Boersma & Weenink, 2018). However, data extraction in Praat is usually detached from subsequent statistical analyses. To streamline these processes, automated techniques can be used, for which Jadoul et al. (2018) propose a new toolkit, Parselmouth, which integrates the extraction of Praat-based acoustic analysis into a Python-based workflow. For users of Python, this allows the combination of acoustic and statistical analyses within one and the same script and may make acoustic analysis using Praat functionalities accessible. For those who currently use Praat, Parselmouth may provide a useful alternative to streamline the process of acoustic analysis and integrate it into a more reproducible workflow.

Taken together, the papers in this volume contribute to our mutual resources by introducing new tools, novel ways of analyzing our data, and by critically evaluating past, present and future analytical practices. Because all authors publicly share their materials, data, and code, they significantly contribute to our shared knowledge and facilitate future research. Aiming at increasing reproducibility has not only practical benefits for individual researchers, but it also benefits us as a collective scientific field, enabling us to access new methods and helping us to substantiate our findings.

[2] At present, many of the descriptions of statistical methods found in phonetics papers do not allow reproducing the performed analysis; in some cases, it is not even clear what general analysis was conducted (e.g., *p*-values may be listed without a detailed description of the associated statistical models these values are based on). For example, Winter (2011) tried to assess how often the independence assumption is violated in speech production data and found that many publications in phonetics do not provide enough information to allow such an assessment. This issue, common in all quantitative sciences, prevents the statistically minded readers to reproduce the analysis and does not allow proper evaluation of the presented evidence.

6. Conclusions

To conclude, we want to emphasize the spirit with which this special issue was conceived. As statistics is constantly evolving within and outside of linguistics and phonetics, there is a plurality of different analysis approaches. Many analytical philosophies alongside methodological tools and techniques co-exist alongside each other at any given point. In many ways, this is advantageous, as this creates the opportunity for discovery of new methods, many of which come from other fields, as well as the opportunity for honest discussion of the advantages and disadvantages of existing approaches. We are in no position, and nor is it our intention, to "police" any existing practices, or to provide recipes or guidelines that everybody should adhere to. Any strict rule will prove to be obsolete in the constantly changing landscape of statistical analysis. Instead, we want to invite the community to reflect on existing practices, as well as to look ahead to incorporate new analysis methods. Instead of accepting any of these techniques as absolute, we have to continue the methodological debate as a community. Moreover, by becoming increasingly reproducible, we can ensure that this plurality of methods benefits our common scientific goal, to understand the physical, cognitive, and social aspects of human speech communication.

References

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language, 94*, 206–234.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv*, preprint arXiv:1506.04967.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics, 70*, 20–38.

Boersma, P. and Weenink, D. (2018). Praat: Doing phonetics by computer [Computer program]. Version 6.0.37, retrieved 3 February 2018 from http://www.praat.org/.

Box, G. E. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics, 1*, 201–236.

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*, 335–359.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics, 71*, 268–283.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London: Palgrave Macmillan.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B (Methodological), 17*(1), 69–78.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical proceedings of the Cambridge philosophical society* (Vol. 22, No. 5, pp. 700–725). Cambridge: Cambridge University Press.

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

Gelman, A. and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press.

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine, 130*(12), 995–1004.

Grice, M., Savino, M., & Roettger, T. B. (2018). Word final schwa is driven by intonation – The case of Bari Italian. *The Journal of the Acoustical Society of America, 143*(4), 2474–2486.

Harrell, F. E. (2014). *Regression modeling strategies* as implemented in R package 'rms' version, 3(3). Berlin: Springer.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164.

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V., Nichols, T. E., & Wagenmakers, E. J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*(1), 69–88.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python Interface to Praat. *Journal of Phonetics, 71*, 1–15.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196–217.

Kirby, J., & Sonderegger, M. (2018). Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics, 70*, 70–85.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford: Oxford University Press.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*(448), 1372–1381.

Lindquist, E. F. (1940). *Statistical Analysis in Educational Research*. Boston, MA: Houghton Mifflin.

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

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton: Chapman & Hall/CRC Press.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., & Spies, J. R. (2016). Point of view: How open science helps researchers succeed. *Elife, 5*, e16800.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

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.

Mücke, D., Grice, M., & Cho, T. (2014). More than a magic moment–Paving the way for dynamics of articulation and prosodic structure. *Journal of Phonetics, 44*, 1–7.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika, 20*, 175–240.

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

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Nosek, B. A. (2017). Center for Open Science: Strategic Plan. Open Science Framework. August 1. doi:10.17605/osf.io/x2w9h.

Peng, R. D. (2011). Reproducible research in computational science. *Science, 334*(6060), 1226–1227.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*, 223.

Pham, H., & Baayen, H. (2015). Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition and Neuroscience, 30*(9), 1077–1095.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1* e175.

Plummer, A. R., & Reidy, P. F. (2018). Computing low-dimensional representations of speech from socio-auditory structures for phonetic analyses. *Journal of Phonetics, 71*, 355–375.

Politzer-Ahles, S., & Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics, 70*, 56–69.

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

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

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.

Roettger, T. B. (2019). Researcher degrees of freedom in phonetic sciences. *Journal of the Association for Laboratory Phonology*. Accepted for publication.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356.

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.

Sterne, J. A., & Smith, G. D. (2001). Sifting the evidence—what's wrong with significance tests? *Physical Therapy, 81*(8), 1464–1469.
Tomaschek, F., Hendrix, P., & Baayen, H. (2018). Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics, 71*, 249–267.
Tufte, E. R. (1990). *Envisioning information*. Cheshire: Graphics Press.
Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175.
Vasishth, S., Nicenboim, B., Fangfang, L., Kong, E., Beckman, M. E., & Edwards, J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics, 71*, 147–161.
Vatikiotis-Bateson, E., Barbosa, A. V., & Best, C. T. (2014). Articulatory coordination of two vocal tracts. *Journal of Phonetics, 44*, 167–181.
Wagenmakers, E. J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer.
Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25*(3), 169–176.
Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: time for a new data presentation paradigm. *PLoS Biology, 13*(4) e1002128.
Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116.
Winter, B. (2011). Pseudoreplication in phonetic research. In *Proceedings of the international congress of phonetic science* (pp. 2137–2140).
Winter, B., & Roettger, T. (2011). The nature of incomplete neutralization: Implications for laboratory phonology. *Grazer Linguistische Studien, 76*, 55–74.
Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution, 1*(1), 7–18.
Wood, S. (2006). *Generalized additive models: An introduction with R*. CRC Press.
Zuur, A. F., Leno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14.