Open Science in phonetics and phonology*

Stefano Coretta

22/01/2020

1 Introduction

Open Science is a movement that stresses the importance of a more honest and transparent scientific attitude by promoting a series of research principles and by warning from common, although not necessarily intentional, questionable practices and misconceptions. The term Open Science as a whole refers to the fundamental concepts of 'openness, transparency, rigour, reproducibility, replicability, and accumulation of knowledge' (Crüwell et al. 2018:3). The goodness of the latter depends in great part on the reproducibility and replicability of the studies that contribute to knowledge accumulation. While reproducibility and replicability are generally used interchangeably, they refer to two different ideas. A study is *replicable* when researchers can independently run the study on new subjects/data and obtain the same results (in brief, same analysis, different data/researchers). The *reproducibility* of a study is, instead, related to the ability of independent researchers to run the original analysis on the original data and obtain the same results as those presented by the original authors, pending enough information on the analysis procedures is given (in brief, same analysis, same data).

A sense of need for Open Science, now increasingly spreading to different disciplines and enterprises, arose primarily from the ongoing so-called 'replication crisis' (Pashler & Wagenmakers 2012; Schooler 2014), which has attracted the most attention within the circles of medical and psychological sciences. Recent attempts to replicate results from high-impact studies in psychology have demonstrated an alarmingly high rate of failure to replicate. For example, in a replication attempt of 100 psychology studies, only 39% of the original results were rated

*This preprint is a standalone version of a chapter from Stefano Coretta's PhD dissertation (Coretta, Stefano. 2020. Vowel duration and consonant voicing: A production study. University of Manchester dissertation.). Please use the OSF preprint DOI when citing.

by annotators as successfully replicated (Open Science Collaboration 2015). Failure to replicate previous results have been claimed to be a consequence of low statistical power (Button et al. 2013), and of so-called questionable research and measurement practices (Simmons et al. 2011; Morin 2015; Flake & Fried 2019). The following sections discuss these problems in turn.

2 "With great power comes great replicability"

One of the issues that can affect statistical analysis is related to errors in rejecting the null hypothesis.[1] A researcher could falsely reject the null hypothesis when in fact is correct (Type I errors, an effect is found when there is none), or they could falsely fail to reject the null hypothesis when in fact it should have been (Type II errors, an effect is not found when there is one). Type I and Type II errors do occur and cannot be totally prevented. Rather, the aim is to keep their rate of occurrence as low as possible. The generally accepted rates of Type I and Type II errors are 0.05 and 0.2 respectively (usually referred to as the α and β levels). This means that, in a series of imaginary multiple replications of a study, 5% of the times the null hypothesis will be falsely rejected, and 20% of the times will falsely be not rejected. A concept closely related to Type II errors is statistical power, which is the probability of correctly rejecting the null hypothesis when it is false (calculated as $B = 1 - \beta$). In other words, power is the probability of detecting an effect equal or greater than a specified effect size. Given the standard $\beta = 0.2$, an accepted (minimum) power threshold is 80% (which means that an effect equal or greater than a chosen size will be detected 80% of the times).

Two other types of statistical errors are the Type S (sign) and Type M (magnitude) errors (Gelman & Tuerlinckx 2000; Gelman & Carlin 2014). Type S errors refer to the probability of the estimated effect having the wrong sign (for example, finding a positive effect when in reality the effect is negative), while Type M errors correspond to the exaggeration ratio (the ratio between the estimated and the real effect). When the statistical power of a study is low (below 50%), Gelman & Carlin (2014) show that the exaggeration ratio (Type M error) is particularly high (from 2.5 up to 10 times the true effect size). Type S errors (wrong sign) are more common at lower power levels (below 10%), although these can easily arise due to small sample sizes and high variance.

Several researchers have shown that the average statistical power of studies in different disciplines is very low (35% or below) and that the last 50 years did not witness an improvement. Bakker et al. (2012) show that the median statistical power in psychology is 35%, while Button et al. (2013) reports a median of 21%

[1] The quote in the title is from a 2016 twitter status by Nathan C. Hall (https://twitter.com/prof_nch/status/790744443313852417?s=20).

obtained from 48 neuroscience meta-analyses. In Dumas-Mallet et al. (2017), half of the surveyed biomedical studies (N = 660) have power below 20%, while the median ranges between 9% and 30% depending on the subfield. Rossi (1990) and Marszalek et al. (2011) show that from the 70s to date there hasn't been an increase in power and sample sizes. Tressoldi & Giofré (2015) also find that only 2.9% of 853 studies in psychology report a prospective power analysis for sample size determination, i.e. the estimation of the smallest sample size necessary to obtain a certain power level before the experiment is run. In sum, low statistical power (well below the recommended 80% threshold) seems to be the norm.

3 The dark side of research

Questionable research and measurement practices are practices that negatively affect the scientific enterprise, but that are employed (most of the time unintentionally) by a surprisingly high number of researchers (John et al. 2012). Silberzahn et al. (2018) asked 29 teams (61 analysts) to answer the same research question given the same data set, and showed that data analysis can be highly subjective. A total of 21 unique combinations of predictors were used across the 29 teams, leading to diverging results (20 teams obtained a significant result, while 9 did not). At various stages of the study timeline, a researcher can exploit the so-called 'researchers' degrees of freedom' to obtain a significant result (Simmons et al. 2011). The researchers' degrees of freedom create a 'garden of forking paths' (Gelman & Loken 2013), that the researcher can explore until the results are satisfactory (i.e., they lead to high-impact or expected findings).

P-hacking is a general term that refers to the process of choosing and reporting those analyses that change a non-significant *p*-value to a significant one (Simmons et al. 2011; Wagenmakers 2007; Motulsky 2014). *P*-hacking can be achieved by several means, for example by trying different dependent variables, including and/or excluding predictors, selective inclusion/exclusion of subjects and observations, or sequential testing (collecting data until the results are significant). Another common practice is to back-engineer a hypothesis after obtaining unexpected results, also known as Hypothesising After the Results are Known (HARKing, Kerr 1998). Lieber (2009) warns against 'double dipping', or the use of the same data to generate a hypothesis and test it. Morin (2015) and Flake & Fried (2019) more specifically discuss questionable practices related to how research variables are measured and operationalised. The literature reviewed in Flake & Fried (2019) suggests that a very high percentage of published papers contains measures that are created on the fly but lack any reference to reliability tests. Researchers have also been found to manipulate validated scales to obtain desired results.

Cognitive biases and statistical misconceptions can also have a negative impact on research conduct. Wagenmakers et al. (2012) discuss the effects of cognitive biases like the confirmation bias (the tendency to look for facts and interpretations that confirm one's prior conviction, Nickerson 1998) and the hindsight bias (the tendency to find an event less surprising after it has occurred, Roese & Vohs 2012). Greenland (2017) defines further common distortions pertaining to methodological approaches, like statistical reification (interpreting statistical results as reflections of an actual physical reality). Finally, Wagenmakers (2007) and Motulsky (2014) examine mistaken beliefs about the meaning of *p*-values and statistical significance (like interpreting *p*-values as an index to statistical evidence or the idea that *p*-values inform us about the likelihood of the null-hypothesis given the data).

A bias in the observed effects can also arise at the stage of publication. A publication bias has been observed in that significant and novel results are generally favoured over null results or replications (Easterbrook et al. 1991; Ioannidis 2005; Song et al. 2010; Kicinski 2013; Nissen et al. 2016). Rosenthal (1979) called the bias agains publishing null results the 'file drawer' problem. Studies that don't lead to a significant result are stored in a metaphorical file drawer and forgotten. This practice not only can bias meta-analytical effect sizes, but also allows for waste of resources when studies with undisclosed null results are repeatedly performed. The questionable research and measurement practices described above, together with publication bias, conspire to unduly increase confidence in our research outcomes. A final exculpatory note is due, though, in that these practices are not necessarily intentional or fraudulent, and in some cases lie within a 'grey area' of accepted standard procedures.

4 Where we stand and where we are heading

Given the similarities in methods between the psychological sciences and phonetics/phonology, it is reasonable to assume that the situation does not fare better in the latter. As mentioned above, sample size, coupled with the effects of increased variance due to between-subject designs, can have a big impact on statistical power. Kirby & Sonderegger (2018) suggest that the number of participants in phonetic studies is generally low, and that, even with nominally high-powered sample sizes, estimation of small effect sizes is subject to the power-related issues discussed above (especially Type S/M errors). Nicenboim et al. (2018) further show how low statistical power has adverse effects on the investigation of phonetic phenomena characterised by small effect sizes, like incomplete neutralisation. Winter (2015) further argues that the common practice of using few items (e.g. word types) and a high number of repetitions increases statistical certainty of the estimates of idiosyncratic differences between items rather than those of the sought effects. Roettger (2019) discusses how the inherently multidimensional nature of speech favours exploration of the researcher’s degrees of freedom, by allowing the researcher to navigate through a variety of choices of phonetic correlates and their operationalisation.

In a review of 113 studies of acoustic correlates of word stress in a variety of languages, published between 1955 and 2017, Roettger & Gordon (2017) show that the majority of studies include 1 to 10 speakers (mode = 1), 1 to 40 lexical items, and 1 to 6 repetitions. A follow-up analysis conducted on the same data indicates that the median number of participants per study is 5 (see Appendix A). A few recent studies (2010 onwards) constitute a clear exception by having more than 30 participants. However, no apparent trend of increasing average number of speakers can be observed and the situation has been fairly stable over the years. Finally, the language endangerment status has a small but negligible negative effect on participants’ number in vulnerable and definitely endangered languages, but not so much in severely and critically endangered ones. It is reasonable to assume that, based on this cursory analysis, sample size in phonetic studies is generally very low, independent from publication year and endangerment status.

As a partial remedy to the issues discussed so far, researchers have proposed two solutions: pre-registrations and Registered Reports. Pre-registration of a study consists in the researchers’ commitment to an experimental and analytical protocol before collecting and seeing the data (Wagenmakers et al. 2012; van ’t Veer & Giner-Sorolla 2016). Pre-registering a study establishes a clear separation between confirmatory (hypothesis testing) analyses and exploratory (hypothesis generating) research. While both types of research are essential to scientific progress (Tukey 1980), presenting exploratory analyses as confirmatory is detrimental to it. Pre-registrations ensure researchers comply to such demarcation, while leaving space to generate new hypotheses via exploratory research. A more recent initiative proposes Registered Reports as a publication format that can counteract questionable research practices and the exploitation of the researcher’s degrees of freedom (Chambers et al. 2015). At the time of writing, no journal specialised in phonetics/phonology offers this article format, although it is currently under implementation at the Journal of the Association for Laboratory Phonology and a few other journals focussed on other linguistic fields.[2]

Another incentive to developing a transparent research attitude comes from aspects of reproducibility. As discussed above, a research analysis is reproducible when different researchers obtain the same results as in the published study by running the same analysis on the same data. Ensuring full reproducibility also means ensuring computational reproducibility, or in other words enabling researchers to

[2] See the spreadsheet at this link for a curated list: `https://docs.google.com/spreadsheets/d/17dLaqKXcjyWk1thG8y5C3_fHXXNEqQMcGWDY62BOc0Q/edit?usp=sharing`.

perform the original analysis in an identical computational environment (Schwab et al. 2000; Fomel & Claerbout 2009). Peng (2009) mentions exposed cases of fraudulent data manipulation and unintentional analysis errors that call for policies of reproducibility to ensure accountability of published results. Our field is not immune from these issues (see for example the 'Yokuts vowels' case, Weigel 2002, 2005; Blevins 2004), and the idea of reproducibility is not new to linguistics in general (Bird & Simons 2003; Thieberger 2004; Maxwell & Amith 2005; Maxwell 2013; Cysouw 2015; Gawne et al. 2017) nor to phonetics/phonology specifically (Abari 2012; Roettger 2019).

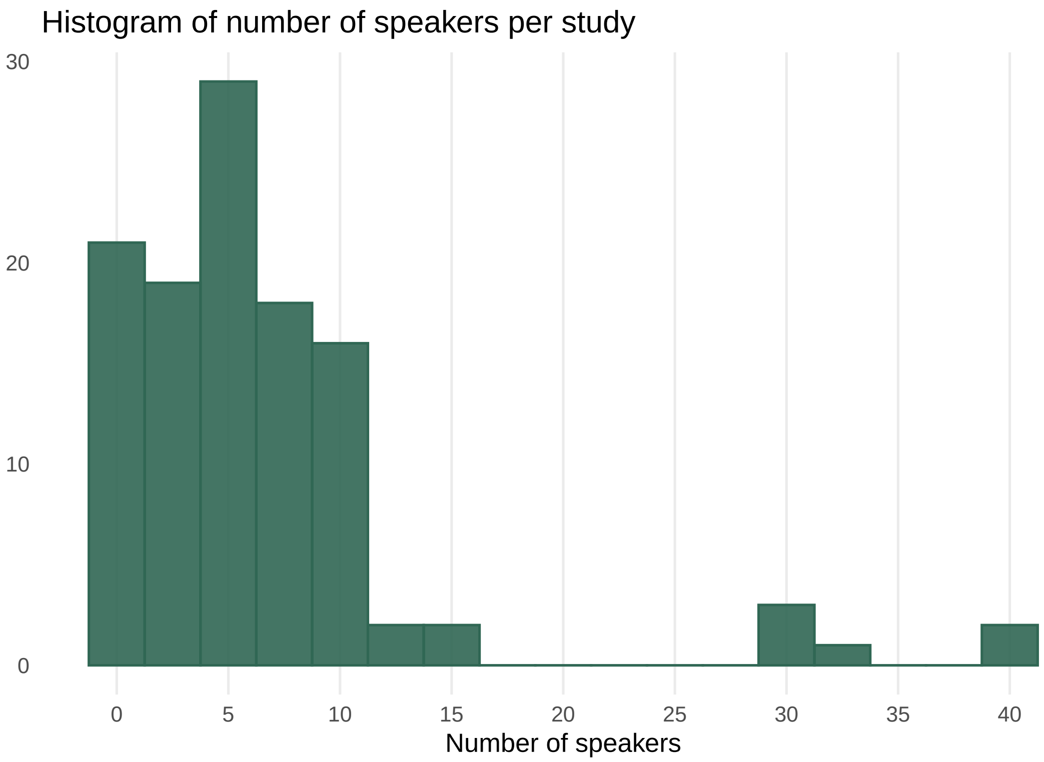
The objective of making research accountable can be achieved by publicly sharing data (subject to ethical restrictions), analysis code, and detailed information on the software that produced the results (Sandve et al. 2013). Sharing data is also fundamental for the accumulation of knowledge, for example in the context of meta-analytical studies. Several services are now available which offer free online data storage and versioning, like the Open Science Framework, GitHub, and DataHub. Extensive documentation of code takes on an important role, and the paradigm of literate programming offers a practical solution (Knuth 1984). Within the literate programming framework, code and documentation coexist within a single source file, and code snippets are interwoven with their documentation. Reproducible reporting further implements this concept (Peng 2015) by automating the generation and inclusion of summary tables, statistics, and figures in a paper using statistical software like R (R Core Team 2019). In a reproducible report, data and results are computationally linked via the statistical software, and changes in data or analyses are reflected in changes in the results appearing in the text. This workflow reduces chances of reporting errors and facilitates validation of the data analyses by other researchers.

A An informal analysis of number of speakers per phonetic study by year and endangerment status

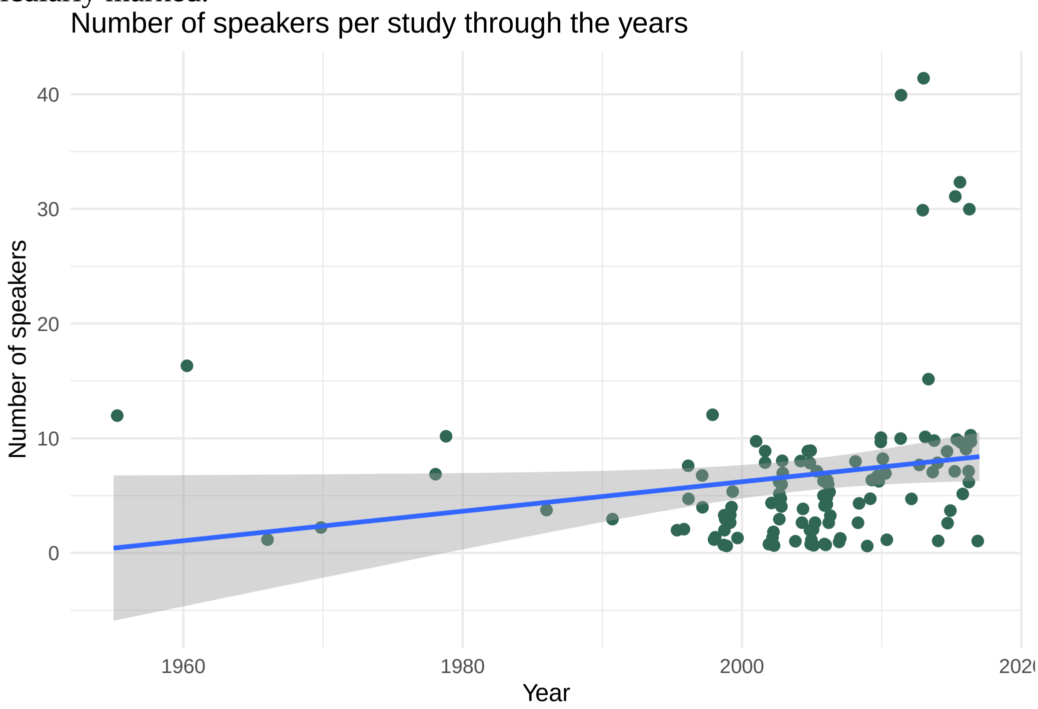
This analysis is based on the dataset used in Roettger & Gordon (2017) and Gordon & Roettger (2017) (Gordon & Roettger 2018).[3] The dataset contains information on number of participants from 113 studies, published between 1955 and 2017 (the majority of the studies are within the range 1990–2017).

The median number of speakers per study across the entire dataset is 5. The histogram below shows that most studies have 10 speakers or less, and that there are a few outliers with 30-40 speakers.

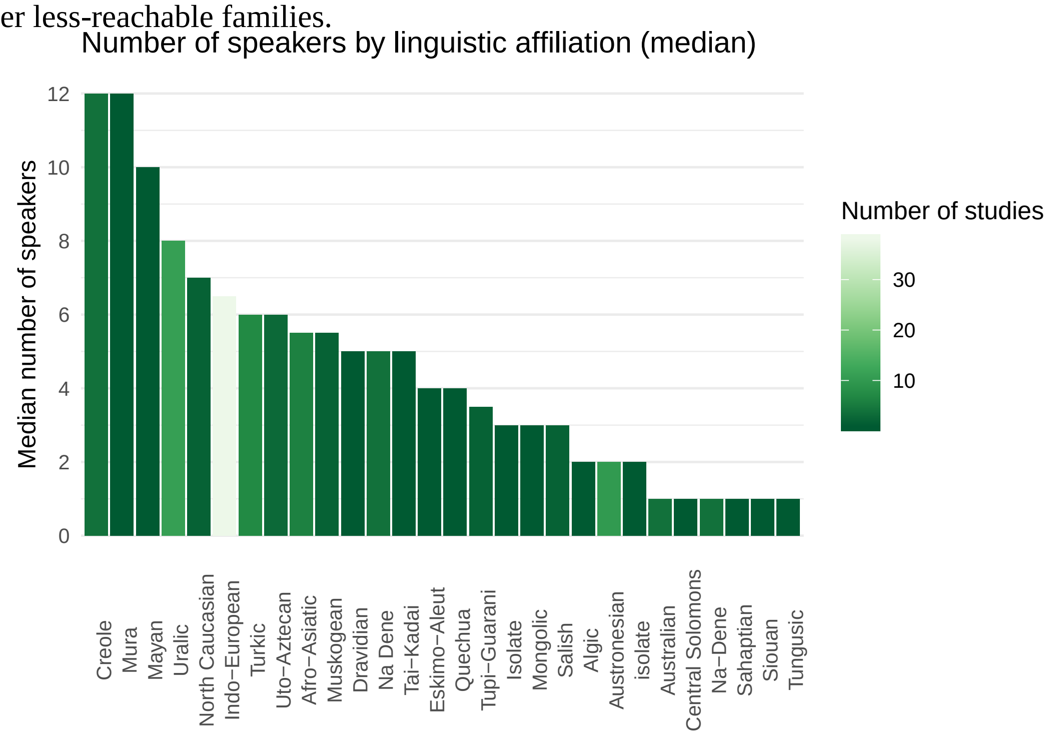
[3] A previous version of this appendix appeared as a blog post at `https://stefanocoretta.github.io/post/an-estimate-of-number-of-speakers-per-study-in-phonetics/`.


Histogram of number of speakers per study

The following plot shows the number of speakers across publication year. There is a tendecy for an increase in number of speakers, although the trend is not particularly marked.

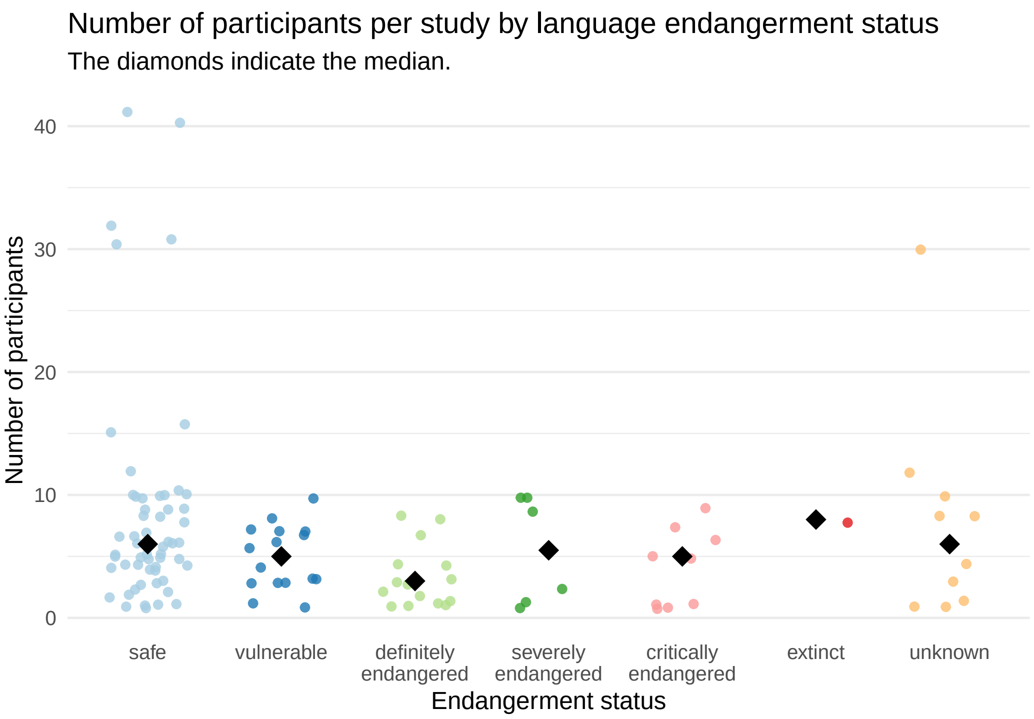

Number of speakers per study through the years

The following bar chart shows the median number of speakers in studies grouped by linguistic affiliation. The colour of the bars indicates the number of studies. Indo-European languages stand out in terms of number of studies (> 30), but the median number of speakers in this family does not fare much better than other less-reachable families.



Information on the endangerment status of the languages in the dataset was obtained from GlottoLog.[4] The following strip chart shows the number of speakers for each of the studies (each point) categorised by the endangerment of the target language. With the caveat that there are more studies on safe languages, there is a trend of decreasing number of speakers from safe, to vulnerable, to definitely endangered languages. The very low number of studies on languages of greater endangerment status makes it harder to establish patterns. Note also that the decreasing trend is in fact small (1/2 speakers).

[4] `https://glottolog.org/meta/downloads`.



While generalisations based on this cursory analysis would not be wise, there seems to be a tendency for studies to have a very low number of speakers (median 5 speakers per study). The majority of studies analysed data from 10 speakers or less. This estimate is independent of publication year and endangerment status of the language enquired.

References

Abari, Kálmán. 2012. Reproducible research in speech sciences. *International Journal of Computer Science Issues* 9(6). 43–52.

Bakker, Marjan, Annette van Dijk & Jelte M. Wicherts. 2012. The rules of the game called psychological science. *Perspectives on Psychological Science* 7(6). 543–554. doi:10.1177/1745691612459060.

Bird, Steven & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 557–582. doi:10.1353/lan.2003.0149.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14(5). 365. doi:10.1038/nrn3475.

Chambers, Christopher D., Zoltan Dienes, Robert D. McIntosh, Pia Rotshtein & Klaus Willmes. 2015. Registered reports: realigning incentives in scientific publishing. *Cortex* 66. A1–A2. doi:10.1016/j.cortex.2015.03.022.

Crüwell, Sophia, Johnny van Doorn, Alexander Etz, Matthew Makel, Hannah Moshontz, Jesse Niebaum, Amy Orben, Sam Parsons & Michael Schulte-Mecklenbeck. 2018. 8 easy steps to open science: An annotated reading list. PsyArXiv. doi:10.31234/osf.io/cfzyx.

Cysouw, Michael. 2015. Accountable and reproducible research. `https://github.com/cysouw/Reproducible-Research/blob/master/README.pdf`.

Dumas-Mallet, Estelle, Katherine S. Button, Thomas Boraud, Francois Gonon & Marcus R. Munafò. 2017. Low statistical power in biomedical science: a review of three human research domains. *Royal Society open science* 4(2). 160254. doi:10.1098/rsos.160254.

Easterbrook, Phillipa J., Ramana Gopalan, J. A. Berlin & David R. Matthews. 1991. Publication bias in clinical research. *The Lancet* 337(8746). 867–872. doi:10.1016/0140-6736(91)90201-Y.

Flake, Jessica Kay & Eiko I. Fried. 2019. Measurement schmeasurement: Questionable measurement practices and how to avoid them. Pre-print available at PsyArXiv. doi:10.31234/osf.io/hs7wm.

Fomel, Sergey & Jon Claerbout. 2009. Guest editors' introduction: Reproducible research. *Computing in Science and Engineering* 11(1). 5–7. doi:10.1109/MCSE.2009.14.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. doi:hdl.handle.net/10125/24731.

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. doi:10.1177/1745691614551642.

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, `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

Gelman, Andrew & Francis Tuerlinckx. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics* 15(3). 373–390. doi:10.1007/s001800000040.

Gordon, Matt & Timo B. Roettger. 2018. Studies on acoustic correlates of word stress - an online corpus. (Last updated: 2018 April 17). doi:10.17605/OSF.IO/9R2CD.

Gordon, Matthew & Timo B. Roettger. 2017. Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard* 3(1). doi:10.1515/lingvan-2017-0007.

Greenland, Sander. 2017. Invited commentary: the need for cognitive science in methodology. *American Journal of Epidemiology* 186(6). 639–645. doi:10.1093/aje/kwx259.

Ioannidis, John P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2(8). e124. doi:10.1371/journal.pmed.0020124.

John, Leslie K., George Loewenstein & Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23(5). 524–532. doi:10.1177/0956797611430953.

Kerr, Norbert L. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2(3). 196–217. doi:10.1207/s15327957pspr0203_4.

Kicinski, Michal. 2013. Publication bias in recent meta-analyses. *PLoS ONE* 8(11). e81823. doi:10.1371/journal.pone.0081823.

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

Knuth, Donald E. 1984. Literate programming. *The Computer Journal* 27(2). 97–111. doi:10.1093/comjnl/27.2.97.

Lieber, Rochelle. 2009. *Introducing morphology*. Cambridge University Press.

Marszalek, Jacob M., Carolyn Barber, Julie Kohlhart & B. Holmes Cooper. 2011. Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills* 112(2). 331–348. doi:10.2466/03.11.PMS.112.2.331-348.

Maxwell, Michael. 2013. A system for archivable grammar documentation. In Georg Rehm, Cerstin Mahlow & Michael Piotrowski (eds.), *Systems and Frameworks for Computational Morphology. Third International Workshop,* 72–91. Springer.

Maxwell, Michael & Jonathan D. Amith. 2005. Language documentation: the Nahuatl grammar. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, 474–485. Berlin Heidelberg: Springer-Verlag.

Morin, Olivier. 2015. A plea for "shmeasurement" in the social sciences. *Biological Theory* 10(3). 237–245. doi:10.1007/s13752-015-0217-z.

Motulsky, Harvey J. 2014. Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Arch Pharmacol* 387. 1017–1023. doi:10.1007/s00210-014-1037-6.

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. doi:10.1016/j.wocn.2018.06.001.

Nickerson, Raymond S. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2(2). 175–220. doi:10.1037/1089-2680.2.2.175.

Nissen, Silas Boye, Tali Magidson, Kevin Gross & Carl T. Bergstrom. 2016. Publication bias and the canonization of false facts. *Elife* 5. e21451. doi:10.7554/eLife.21451.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349(6251). aac4716. doi:10.1126/science.aac4716.

Pashler, Harold & Eric-Jan Wagenmakers. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science* 7(6). 528–530. doi:10.1177/1745691612465253.

Peng, Roger D. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3). 405–408. doi:10.1093/biostatistics/kxp014.

Peng, Roger D. 2015. *Report writing for data science in R.* Lulu.

R Core Team. 2019. R: A language and environment for statistical computing. `https://www.R-project.org/`.

Roese, Neal J. & Kathleen D. Vohs. 2012. Hindsight bias. *Perspectives on psychological science* 7(5). 411–426. doi:10.1177/1745691612454303.

Roettger, Timo B. 2019. Researcher degrees of freedom in phonetic sciences. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1). 1–27. doi:10.5334/labphon.147.

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

Rosenthal, Robert. 1979. The file drawer problem and tolerance for null results. *Psychological bulletin* 86(3). 638. doi:10.1037/0033-2909.86.3.638.

Rossi, Joseph S. 1990. Statistical power of psychological research: What have we gained in 20 years? *Journal of consulting and clinical psychology* 58(5). 646. doi:10.1037/0022-006X.58.5.646.

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor & Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology* 9(10). 1–4. doi:10.1371/journal.pcbi.1003285.

Schooler, Jonathan W. 2014. Metascience could rescue the 'replication crisis'. *Nature News* 515(7525). 9. doi:10.1038/515009a.

Schwab, Matthias, N. Karrenbach & Jon Claerbout. 2000. Making scientific computations reproducible. *Computing in Science & Engineering* 2(6). 61–67. doi:10.1109/5992.881708.

Silberzahn, Raphael, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard & Evelina Bonnier. 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. doi:10.1177/2515245917747646.

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. doi:10.1177/0956797611417632.

Song, Fujian, Sheetal Parekh, Lee Hooper, Yoon K. Loke, J. Ryder, Alex J. Sutton, C. Hing, Chun Shing Kwok, Chun Pang & Ian Harvey. 2010. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 14(8). 1–193. doi:10.3310/hta14080.

Thieberger, Nicholas. 2004. Documentation in practice: Developing a linked media corpus of South Efate. In *Language documentation and description*, vol. 2, London: SOAS.

Tressoldi, Patrizio E. & David Giofré. 2015. The pervasive avoidance of prospective statistical power: major consequences and practical solutions. *Frontiers in psychology* 6(726). doi:10.3389/fpsyg.2015.00726.

Tukey, John W. 1980. We need both exploratory and confirmatory. *The American Statistician* 34(1). 23–25. doi:10.1080/00031305.1980.10482706.

van 't Veer, Anna Elisabeth & Roger Giner-Sorolla. 2016. Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology* 67. 2–12. doi:10.1016/j.jesp.2016.03.004.

Wagenmakers, Eric-Jan. 2007. A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review* 14(5). 779–804. doi:10.3758/BF03194105.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas & Rogier A. Kievit. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7(6). 632–638. doi:10.1177/1745691612463078.

Weigel, William Frederick. 2002. The Yokuts canon: A case study in the interaction of theory and description. Paper presented at the annual meeting of the Linguistics Society of America, January 2002, San Francisco.

Weigel, William Frederick. 2005. *Yowlumne in the Twentieth century*: University of California, Berkley dissertation.

Winter, Bodo. 2015. The other N: The role of repetitions and items in the design of phonetic experiments. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow: The University of Glasgow. `https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0181.pdf`.