STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Data-dependent analytical choices relying on NHST should not be trusted!

Jesper W. Schneider*

*jws@ps.au.dk
Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University, Bartholins Alle 7, Aarhus, 8000-DK (Denmark)

Introduction

Reproducibility and replication seems to be on everybody's lips these days in academia, at least in the "softer" sciences (Fanelli, 2010; Baker & Penny, 2016). Until recently, replication studies in the behavioural and social sciences were not acknowledged (i.e. this is still the case in many fields) and very few researchers conducted them. However, a number recent replication studies have seemingly provided empirical evidence for old warnings and concerns about a flawed knowledge production process that eventually leads to numerous false-positive claims (e.g., Meehl, 1967; J. P. A. Ioannidis, 2005; Button et al., 2013; Collaboration, 2015). Hence, we have seen that previously hailed claims, effects or even theories have failed to replicate (e.g., Hagger & Chatzisarantis, 2016). Currently we discuss the potential causes of this so-called "replication crisis" (e.g., Simmons, Nelson, & Simonsohn, 2011; Munafò et al., 2017). Several of them have been known for decades, while others, such as our understanding of "questionable research practices" (QRP) are relatively new and still under-explored (John, Loewenstein, & Prelec, 2012).

A lot of confusion, however, flurries when it comes to "replication" and "reproducibility". What do the terms mean? Why and when are replication studies important? What are the implications of failure to replicate? How do we define a "failure", and many more issues could certainly be listed. In light of the current attention and debates, several fields including Scientometrics are discussing reproducibility issues and how it relates to their specific knowledge production models, but not everyone agrees that a "replication crisis" actually exists (Peng, 2015; Patil, Peng, & Leek, 2016).

This short opinion paper seeks to cut through some the hype and outline some basic facts of what we know and do not know at the moment about "reproducibility" and "replication". I will discuss these issues and their implications for the field of Scientometrics, and I will briefly list some of the arguments brought forward for why data and code sharing should be universally promoted. However, my focus will be on the kernel of the current debate, the reliance of null hypothesis significance tests (NHST) for knowledge claims and the implications this have on reproducibility issues. The latter is as important for Scientometric and research evaluation studies, as it is for all other studies using such tests.

The paper is organized as follows: The next section outlines the characteristics of the so-called replication crises and how it relates to scientometric studies. Then I discuss some implications of "confirmatory" studies focusing on data-dependent analytical choices known as "garden of forking paths". Finally, I briefly discuss the implications of "garden of forking paths" for scientometric studies.

The characteristics of the "replication crisis"

A "replication crisis" or a "reproducibility crisis" has been declared and discussed after a string of failed replication studies (e.g., Prinz, Schlange, & Asadullah, 2011; C. Glenn Begley & Ellis, 2012; Pashler, Coburn, & Harris, 2012; Button et al., 2013; Collaboration, 2015; J. P. A. Ioannidis, 2016). This is crucial as replication is a cornerstone in the idealized view of the scientific method and the supposed ability of science to be self-correcting and cumulative (John P. A. Ioannidis, 2012). The low replication rates are claimed to reflect that many published findings are essentially based on flawed research designs (C. G. Begley & Ioannidis, 2015), excessive reliance on NHST (J. P. A. Ioannidis, 2005; Button et al., 2013; Andrew Gelman & Loken, 2014), and more deliberate questionable research practices such as HARking or p-hacking (Kerr, 1998; Simmons, Nelson, & Simonsohn, 2011). Indeed, some authors see the flawed use of statistics and over-reliance on statistical evidence as *the* focal issue that has led to a proclamation of a "statistical crisis in science" (Andrew Gelman & Loken, 2014). However, it is rarely acknowledged that the discussions of replication and reproducibility and a potential crisis are almost exclusively based on challenges linked to experimental knowledge production models and the potential pitfalls linked to their designs and reliance upon NHST. It is also often ignored that the vast majority of the empirical evidence comes from a relative restricted number of areas mainly behavioural and biomedical, both clinical and pre-clinical, but also single cases from for chemistry, biology and bioinformatics (see references above). Indeed, the majority of failed replication studies come from social psychology. It is therefore highly relevant to examine to what extent the problems in experimental fields linked to reproducibility and replication issues are also relevant for non-experimental knowledge production models and obviously whether such models themselves bring in further challenges.

The characteristics of experimental knowledge production models in the soft sciences

At the heart of experimental knowledge production models is the aim to test hypotheses and treatments effects "predicted" from "weak theories" (Meehl, 1967). The research approach is mainly confirmatory and often framed as an all-or-nothing game where outcomes below some arbitrary significance level, typically 5%, means success and the ability to publish. "Confirmation" comes in a flawed roundabout manner, as strawman null hypotheses are rejected without having tested the experimental hypothesis directly. Weak theories are simply not capable of making strong predictions; however, being weak means that they can often explain any observed pattern, positive or negative, contradictory to Popper's notion of strong theory testing. Ideally, experiments can have strong internal validity due to randomization, but "control" is more challenging when study objects are humans. So is measurement and most devices used are susceptible to both random and systematic noise and imprecision. Experiments are also characterized by relatively small sample sizes, which brings higher variability and lower statistical power.

Finally, it is very important to stipulate that many experimental fields in the soft sciences are characterised by having a substantial publication bias towards "statistically significant" findings, much more than we should expect from theory. In principle, low power means that only relative strong effects should be detected, it is therefore a paradox that so many studies claim "significant" effects when it is generally acknowledge that most effect sizes in the soft sciences are of a weaker kind and difficult to entangle from the noise surrounding them. Hence, it has been claimed that the "magic" 5% significance level has led to researchers "chasing" significance by using researcher "degrees-of-freedom" in data processing, analyses, and presentations (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016).

The characteristics of scientometric studies

In general, scientometric studies are very diverse. Studies using NHST are mainly non-experimental. Data most often come from one of the major bibliographic databases and/or surveys. Sample sizes are often relatively large, sometimes very large, whereas measurements usually are unobtrusive counts or indices from these databases. Having large sample sizes mean that effect sizes can be estimated with less variability and more precision. It also means that statistical power is stronger and that almost anything can turn out to be statistical significant. Whereas experiments typically use simple significance tests or ANOVA designs, non-experimental knowledge production models typically rely on regression-designs. The latter bring with them a whole string of intricate issues and assumptions which experimental settings are exempted from, most notably the model specification, which are the proposed data generation mechanism (Berk, 2010). Most importantly, such a specification has no randomization scheme and therefore no experimental control. Scientometrics is a data-driven field and most studies do not test hypotheses derived from theories. Nevertheless, using NHST still means that "theories" are tested and thus all the implications for confirmatory studies should be abided to if the statistical evidence is to be utilized as I will discuss below.

From a brief analytical point of view, Scientometrics is only partially susceptible to the challenges that may have been the main causes for the replicability challenges in say social psychology. Our studies have larger sample sizes and thus more power and less variability. We should therefore not expect to see so many false-positive results *if and only if* we have followed the research plan outlined before data was examined! It is also questionable whether our journals decline the publication of null results; this is not my impression and hence, we should not expect to have a serious publication bias in our field. Does this mean that we should expect that most published findings in Scientrometrics do replicate either directly or conceptually? Not necessarily as I will discuss below. Two important aspects should be kept in mind. Non-experimental study designs are much weaker compared to experimental designs and combined with the numerous often undisclosed researcher-degrees-of-freedoms seemingly open for the researcher to explore, means that we should in fact expect that published studies and claims relying on NHST can come out very differently, when different analytical choices and paths are made in replication studies.

The implications of "confirmatory" studies

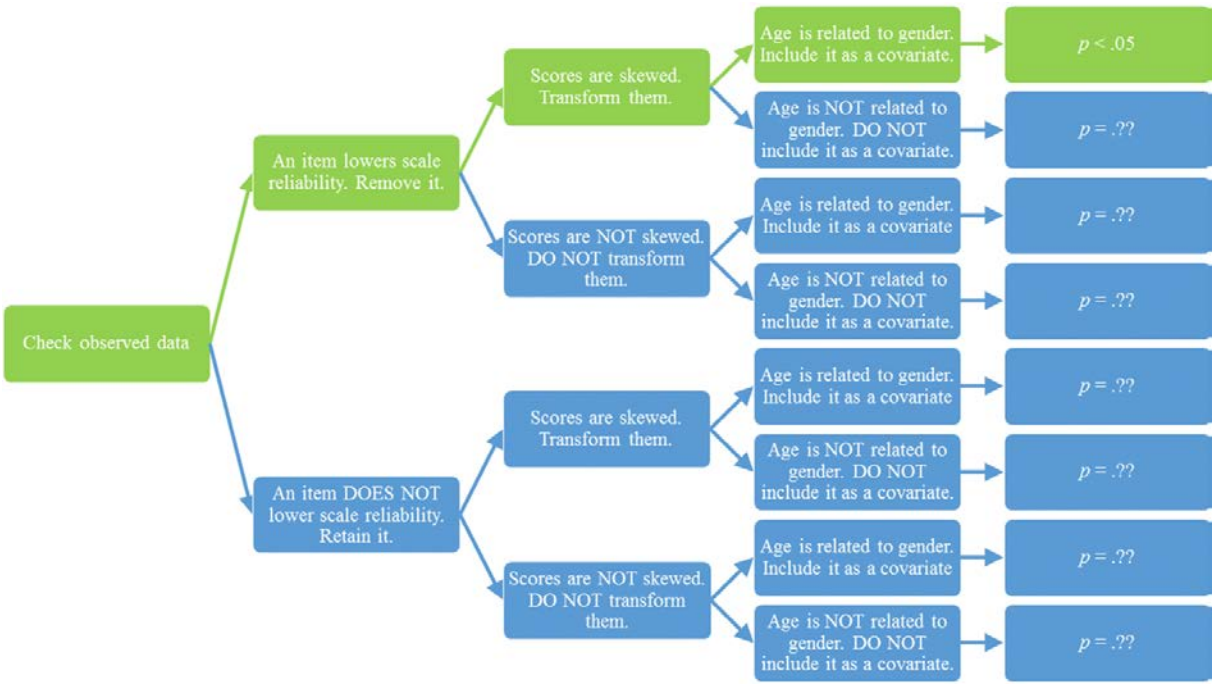
The focus of this paper is upon statistical reproducibility. There are of course several other important issues linked to replication and reproducibility that goes beyond statistics. Openness, data and code sharing are promoted with arguments of various kinds of benefits for both the researchers' sharing but also their research communities, and eventually the norms and progress of science (Munafò et al., 2017). Benefits include the ability to do reanalysis that can detect honest as well as fraudulent errors, provide more in-depth information about findings, detect researchers' susceptibility to biases when examine the results of their own studies, more exploration of the data and so forth. It is not my intention here to discuss these issues. However, based on Schmidt (2009), I shall point to the functional approach to replication and interpret this into the context on non-experimental studies. There are undoubtedly many reasons for studies not to replicate, but in many cases, I suspect that investigators fool themselves due to a poor understanding of statistical concepts and the implications they come with and this may very well be most pertinent for non-experimental studies based on NHST (Motulsky, 2015), also in Scientometrics.

The "Garden of forking paths" or simply "p-hacking"

Above I argued that most Scientometric studies are data-driven and seems to be more exploratory than confirmatory in character. However, when NHST is used implications follow. If the study is perceived as exploratory then the p-values will have different meanings compared to a confirmatory study (Rubin, 2017a), if indeed any meaning at all. In reality most claims about findings are strongly depended on the outcomes of the NHST procedure and then the study is by definition confirmatory. A dichotomous "truth" situation is set up where "truth" is parametrized in a theoretical null hypothesis of no difference or correlation. When this is the premise, an important implication follows: *Statistical results can only be interpreted at face value when every choice in the data analysis was performed exactly as planned, and documented as part of the research design!* Hence, the analytical path should be chosen before the data is seen and one must stick to this plan.

A data set can have many different analytical paths as shown in Figure 1 below and thus an endless number of researcher-degrees-of-freedom; this is what Gelman and Loken (2013) have called a "garden of forking paths". What this essentially means is that the analytical path chosen is data-dependent. We examine the data and then make our analytical choices. The problem comes when you realize that there are a multitude of ways to approach a hypothesis, and there are numerous informal decisions we make when looking at our data about which tests or data we use. For example, which covariates to focus on, how data is transformed, how many categorical bins to divide data into, and so on. These endless choices lead to "garden of forking paths", where researchers may explore their data extensively, but only report a subset of the statistical methods they ended up utilizing (e.g. the models that fit the data better). The mistake is in thinking that, if the particular path that was chosen yields statistical significance, this is strong evidence in favour of the hypothesis. If the data had come out slightly different another path may have been chosen and the results probably come out different. Data-dependent analysis explains why many statistically significant comparisons do not hold up and this is where we fool ourselves!

Figure 1: Illustration of a data-dependent analytical path known as a "garden of forking paths" (from Rubin, 2017b)



Indeed, the metaphor of the "garden of forking paths" is actually an attempt to describe an unintentional research practice that nevertheless is flawed. The "garden of forking paths" changes name when these data-dependent choices become intentional in order to chase statistical significance; then we speak of p-hacking (Simmons, Nelson, & Simonsohn, 2011) which is a serious questionable research practice. In reality, it can be very difficult to separate the two, but their implications are the same: results become uninterpretable!

Discussion: The implications for Scientometric studies

I am not claiming that most published research findings in Scientometrics based on NHST are false, but I do claim that they are at best very susceptible to "garden of forking paths", there is basically no evidence to the contrary, and at worst are deliberately p-hacked, but there is also no evidence for that either. What it does entail is that many of these findings as they stand are very difficult to interpret. They cannot be extrapolated and other analytical paths can lead to different outcomes on the same data sets, as for example demonstrated in Silberzahn and Uhlmann (2015).

What should we do then? In order to be able to examine the stability of a finding, it necessary in the context of NHST-based non-experimental studies to know the intent of the researcher and the research plan set out before data is seen. Sharing data has been promoted, and while certainly laudable and needed, it seems highly sub-optimal in this situation without a time stamped pre-registration of the research plan. Notice, exploring your data can be a very useful way to generate hypotheses and make preliminary conclusions, but all such analyses need to be clearly labelled, and then retested with new data. There is no free lunch when it comes to statistical inference and my conjecture is that the premise of outlining a research plan before data is seen is never fulfilled in our field, which essentially means that most Scientometric studies using NHST are rife with "gardens-of-forking-paths" or worse p-hacked data. For this reason, replication studies are certainly needed!

References

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, *533*(7604), 452-454. doi:10.1038/533452A

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531-533.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in Science Improving the Standard for Basic and Preclinical Research. *Circulation Research*, *116*(1), 116-126. doi:10.1161/circresaha.114.303819

Berk, R. (2010). What You Can and Can't Properly Do with Regression. *Journal of Quantitative Criminology*, *26*(4), 481-487. doi:10.1007/s10940-010-9116-4

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

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

Fanelli, D. (2010). "Positive" Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(3). doi:10.1371/journal.pone.0010068

Gelman, A., & Loken, E. (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*.

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist, 102*(6), 460-465.

Hagger, M. S., & Chatzisarantis, N. L. D. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science, 11*(4), 546-573. doi:10.1177/1745691616652873

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696-701. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science, 7*(6), 645-654. doi:10.1177/1745691612464056

Ioannidis, J. P. A. (2016). Why Most Clinical Research Is Not Useful. *PLoS Medicine, 13*(6), e1002049. doi:10.1371/journal.pmed.1002049

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196-217.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*(2), 103-115. doi:10.2307/186099

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology, 172*(8), 2126-2132.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. doi:10.1038/s41562-016-0021

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of Social Distance? Failure to Replicate Effects on Social and Food Judgments. *PLoS ONE, 7*(8), e42510. doi:10.1371/journal.pone.0042510

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should we expect when we replicate? A statistical view of replicability in psychological science. *Perspectives on psychological science : a journal of the Association for Psychological Science, 11*(4), 539-544. doi:10.1177/1745691616646366

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30-32.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov, 10*(9), 712-712.

Rubin, M. (2017a). Do p values lose their meaning in exploratory analyses? It depends how you define the familywise error rate. *Review of General Psychology, 21*(3), 269-275. doi:http://dx.doi.org/10.1037/gpr0000123

Rubin, M. (2017b). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology, 21*(4), 321-329. doi:http://dx.doi.org/10.1037/gpr0000135

Schmidt, S. (2009). Shall we really do it again? the powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13(2), 90-100*(2), 90-100.

Silberzahn, R., & Uhlmann, E. L. (2015). Many hands make tight work. *Nature, 526*, 189-191.

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. doi:10.1177/0956797611417632

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing,

and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*(1832). doi:10.3389/fpsyg.2016.01832