STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

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

Exploration of reproducibility issues in scientometric research

Theresa Velden[*], Sybille Hinze[**], Andrea Scharnhorst[***], Jesper Wiborg Schneider[****], Ludo Waltman[*****]

[*] *velden@ztg.tu-berlin.de*
Zentrum für Technik und Gesellschaft (ZTG), Technische Universität Berlin, Berlin (Germany)

[**] *hinze@dzhw.eu*
Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW), Berlin (Germany)

[***] *andrea.scharnhorst@dans.knaw.nl*
Data Archiving and Networked Services (DANS), The Hague (the Netherlands)

[****] *jws@ps.au.dk*
Danish Centre for Studies in Research and Research Policy, Aarhus University, Aarhus (Denmark)

[*****] *waltmanlr@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (the Netherlands)

Introduction

The (lack of) reproducibility of published research results has recently come under close scrutiny in some fields of science (see e.g. Flier 2017 for a discussion of bio-sciences, and e.g. Open Science Collaboration 2015 and Pashler & Harris 2012 for an assessment of the situation in psychology). Aside from genuine error or fraud as sources for the irreproducibility of published research, theoretical investigations (e.g. Ioannidis 2005) and empirical investigations (e.g. John et al. 2012) identify the use of questionable research methods, the overselling of results by overstating claims, and publication bias - the tendency to select positive results over negative results for publication - as further sources for the irreproducibility of results.

In scientometrics, we have not yet had an intensive debate about the reproducibility of research published in our field, although concerns about a lack of reproducibility have occasionally surfaced (see e.g. Glänzel & Schöpflin 1994 and Van den Besselaar et al. 2017), and the need to improve the reproducibility is used as an important argument for open citation data (see www.issi-society.org/open-citations-letter/). We initiated a first discussion about reproducibility in scientometrics with a workshop at ISSI 2017 in Wuhan.[1] One of the outcomes was the sense that scientific fields differ with regard to the type and pervasiveness of threats to the reproducibility of their published research, last but not least due to their differences in modes of knowledge production, such as confirmatory versus exploratory study designs, and differences in methods and empirical objects.

Therefore we suggest that an empirical investigation of the specific challenges to the reproducibility of research in the field of scientometrics would be beneficial to focus the

[1] Workshop report available online at www.issi-society.org/blog/posts/2017/november/reproducible-scientometrics-research-open-data-code-and-education-issi-2017/.

debate and efforts to remedy shortcomings. In the run up of the STI 2018 conference, we decided to use a small sample of different types of studies to explore how an assessment of the specific challenges to the reproducibility of research in the field of scientometrics could be conducted based on a critical review of the content of published papers. To systematize our reviews, we developed a taxonomy of threats to reproducibility to look out for in the review of published papers.

Background

The concept of reproducibility can refer to various approaches to and purposes of reproducing (some aspect of) an original study. What variety of reproducibility is seen as most pertinent, seems to depend on scientific domain. The diversity of perspectives has led to a thorough confusion of terminology around reproducibility, including antithetical definitions of the terms *replicability* and *reproducibility* (Goodman et al. 2016; Barba 2018). To cut through the thicket of terminological confusion, we use in this paper the term *reproducibility* as a generic umbrella term and focus our analysis on two distinct subtypes that we define as follows.

Concepts of reproducibility often differ in terms of the degree of similarity between the original study and a reproduction study, including the study design, methods, and data used (Chen 1994). For the purpose of this explorative study, we distinguish two subtypes that are located at opposite ends of this spectrum and have distinct scientific functions: *direct* and *conceptual* reproducibility (in line with Fidler et al. 2017).

Direct reproducibility is located at the ‘greatest similarity’ end of the spectrum where the same data, tools and methods are used to reproduce and verify a study with the expectation of obtaining identical or very similar empirical results, unless some error is made either in the original study or in the reproduction study.

A precondition for direct reproducibility is that the party attempting the reproduction has all the necessary means, information and resources (access to data, tools, infrastructures, relevant tacit knowledge). In a conflation of terminology, the ability to carry out a direct reproduction attempt is often not distinguished from the factual direct reproducibility of a study. In practice, a reproduction attempt may fail either because the preconditions for direct reproducibility are not met or because the original study is factually irreproducible. Given resource restrictions, in this study we could not carry out attempts at direct reproduction of studies, and therefore restrict our review to the ability to carry out an attempt to directly reproduce the selected studies.

Conceptual reproducibility is located at the other end of the spectrum where a study is reproduced using different data, tools and methods with the aim of testing the robustness of the fundamental knowledge claims made by the original study.

While robustness of scientific knowledge is achieved only in a cumulative and discursive process within the scientific community and not by a single publication, we argue that individual publications provide the foundation for producing robust knowledge. The contribution of an individual publication is twofold: first, through the accuracy of the empirical evidence that it generates, that is by delivering results that are directly reproducible. (Factual) direct reproducibility can be seen as a precondition for conceptual reproducibility, and in a field like scientometrics, that is largely empirical and descriptive, ensuring direct reproducibility could be considered the most fundmental step. Second, an individual publication contributes to the robustness of knowledge by articulating its knowledge claims in accordance with the empirical evidence it produces, that is by not using questionable research methods (see e.g. Simmons et al. 2011; Schneider 2015) and/or overstating claims. In other fields, this aspect of robustness of knowledge is also discussed in terms of the methodological and conceptual rigor of a study, see e.g. Moons et al. 2004 on quality of life assessment studies in medicine, Dube & Pare 2003 on case study research in information systems research, and Eisenhardt 1991 on case studies in management science.

Analytical approach

To explore how one might identify reproducibility issues in scientometric publications, we defined a classification of types of scientometric studies and critically reviewed an exemplar of each type with regard to potential threats to reproducibility. To ensure consistency across our reviews, we developed a taxonomy of potential threats to direct reproducibility, presented further below.

Classification of studies

First, we created a classification of scientometric studies into five categories in order to explore how threats to reproducibility may vary by type of study. We refined the empirical category in order to account for the large number and variety of empirical studies in scientometrics. Our classification is presented in Table 1. As often in classification, many studies do not fit neatly into one of our categories. Nevertheless, an assignment is possible by looking at the primary focus of a study.

Table 1: High-level classification of types of scientometric studies.

Category no.	**Name**	**Description**
1	Theoretical/Conceptual	Studies that are primarily theoretically/conceptually focused.
2	Methods	Studies that are primarily methodologically focused.
3	Empirical (General)	Studies that are primarily empirically focused, aimed at answering substantive research questions in the study of science.
4	Empirical (Case)	Studies that are primarily empirically focused, taking a 'case study' approach, that is, focusing on analyzing particular research domains or particular countries, research institutions, or journals. These studies do not aim to develop more general insights that go beyond the particular case they analyze.
5	Empirical (Data Source)	Studies that are primarily empirically focused, aimed at getting a better understanding of the data sources available for scientometric research.

Figure 1: Taxonomy to identify potential threats to direct reproducibility of a published scientometric study.

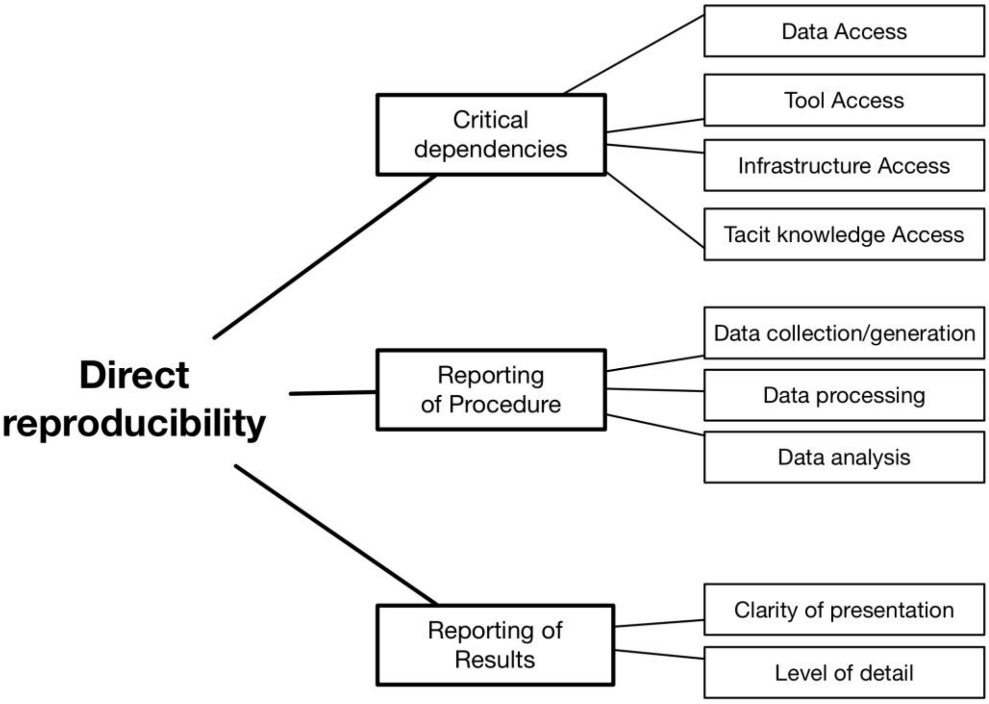
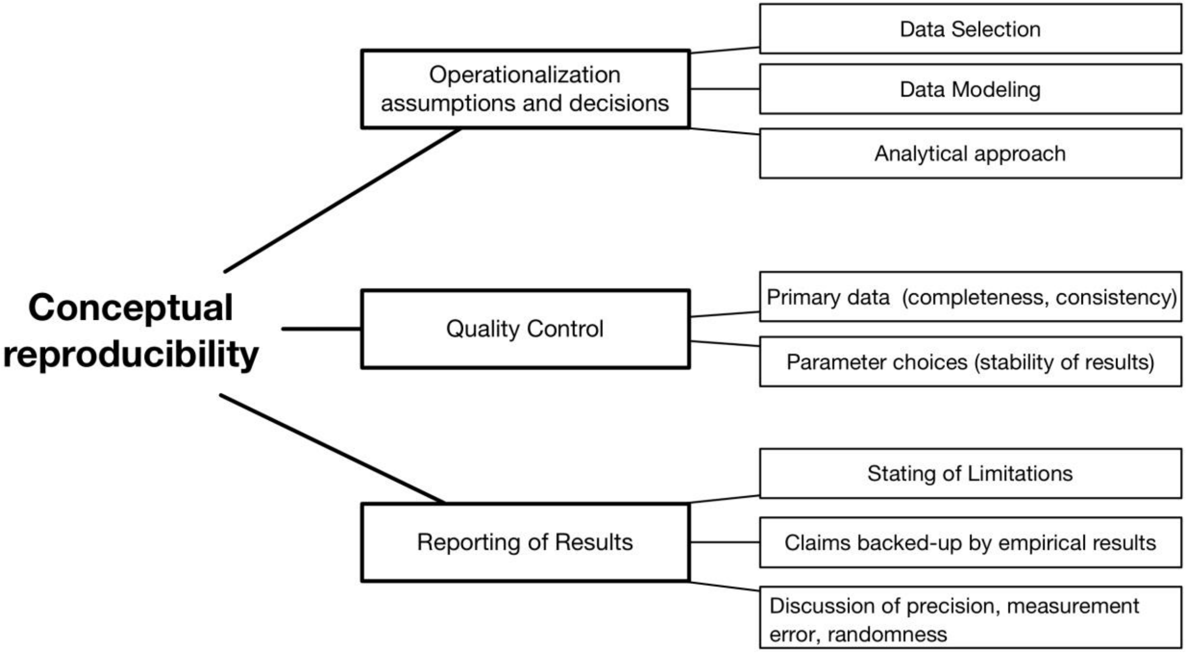


Figure 2: Taxonomy to identify potential threats to conceptual reproducibility of a published scientometric study.


Conceptual reproducibility
Operationalization assumptions and decisions
Data Selection
Data Modeling
Analytical approach
Quality Control
Primary data (completeness, consistency)
Parameter choices (stability of results)
Reporting of Results
Stating of Limitations
Claims backed-up by empirical results
Discussion of precision, measurement error, randomness

Taxonomy of threats to direct reproducibility

The taxonomy that we use captures threats to direct reproducibility by identifying issues that may practically undermine the ability of a third party to conduct a direct reproduction of a study. It distinguishes between critical dependencies (fundamental barriers that cannot be fixed by information provided in publication, such as access to original data or a certain tool used in the study) and issues resulting from incomplete information provided by the publication, either with regard to the procedures used, or with regard to the reporting of the results (see Figure 1.)

Taxonomy of threats to conceptual reproducibility

The taxonomy for threats to conceptual reproducibility identifies substantive issues that undermine our confidence that the central knowledge claims made in a publication are robust and likely to hold up to a test by conceptual reproduction. Based on the debate in other fields that suggests that questionable research methods and overselling play a major role in explaining irreproducibility, we distinguish between Operationalization assumptions and decisions, Quality Control, and Reporting of Results (see Figure 2.)

Within the category of Operationalization assumptions and decisions, we review whether the selection of data, data modeling, and the choice of the analytical methods is appropriate for the chosen research question. Within the category of Quality Control, the firmness of the research design is complemented by looking for evidence for measures for quality control, such as discussions of the completeness and consistency of primary data and of how parameter choices influence the stability of results. Within the category of Reporting of Results, we check whether limitations are explicitly stated, claims backed-up by empirical results, and whether there is an adequate discussion of limits in precision, measurement error, randomness.

Data and method

For each of the five study types in our classification, we selected one paper that was published within the last two years. Two papers were published in *Scientometrics*, two in *Journal of Informetrics* and one paper was made available as a preprint in the arXiv. With the paper selection, we aimed at selecting papers that serve as a good example of one the above five categories. Papers were selected and agreed upon unanimously by all authors of this paper. Each paper was then reviewed by at least two of the authors of this paper, one paper was reviewed by three. Each of the reviewers was asked to assess the papers regarding the elements identified by the taxonomy.

In this paper, we do not reveal the identity of the five papers. Instead we provide an overview of key features of the papers in Table 2. Our focus is on providing general insights into the reproducibility of scientometric research, not about the extent to which specific papers are reproducible. Readers who want to know more about the papers that were reviewed are invited to contact us.

Table 2: Properties of the five papers selected for review in this explorative study.

Paper no.	**Study type**	**Topic area**	**Methods**	**Data**	**Tools**
1	Theoretical/ conceptual	Citation theory	Theoretical reasoning, simulation	Synthetic	Self-developed simulation software
2	Method development	Topic extraction	Network clustering	Bibliometric, proprietary, large-scale (10^7)	Open source software
3	Empirical (General)	Innovation studies	Statistical regression analysis, network analysis	Patent data, proprietary	Standard, proprietary statistical package, network analysis tool (proprietary, free trial)
4	Empirical (Case)	Specialty study at national level	Network analysis and visualization	bibliometric, proprietary, small-scale (10^3)	Freely accessible online tool
5	Empirical (Data source)	Evaluation of sources for citation analysis	Recall and precision measurements, correlation coefficients	Bibliometric, proprietary and freely accessible large-scale (10^5-10^6)	Freely accessible online tool for query generation

Results

Observations regarding threats to direct reproducibility

We organize our report of observations regarding direct reproducibility issues in four parts: Data, software tools, methods, and results.

Data

Empirical data was used in four of the five papers that we reviewed. In all four cases, the data was of a bibliometric nature. Scientometric studies may also use other types of data (e.g., data on peer review outcomes, data on research funding, or survey data), but no such studies were included in our analysis.

Basically, there seem to be two main problems with bibliometric data sources:

1. Some bibliometric data sources (e.g., Web of Science, Scopus, Derwent) are not freely accessible. Especially large-scale data access can be expensive, making it infeasible for many researchers to reproduce studies that rely on large-scale data access. Small-

scale data access (e.g., through the web interfaces of Web of Science or Scopus, based on subscriptions) will often be less problematic and can be sufficient for scientometric case studies (category 4), but it is often insufficient for scientometric studies that aim to draw conclusions that are of general nature and that go beyond one specific case (category 3).

2. All bibliometric data sources seem to lack a systematic approach to version control. Papers sometimes indicate the date at which data was collected from a data source. This may be helpful to approximately reproduce the data collection, but it is not sufficient for exactly reproducing it. To exactly reproduce the data collection, data sources need to adopt a systematic approach to version control or authors need permission to share the primary data on which their study is based.

Software tools

We suggest that from the perspective of direct reproducibility it is useful to distinguish four levels of accessibility of software tools. These levels are listed in Table 3 in increasing order of the degree to which they support direct reproducibility of scientometric research. We found that the software tools used in the five papers reviewed cover all four levels of accessibility.

Table 3: Levels of accessibility to support direct reproducibility

Access level	Description	Example	Implication
0	Custom software developed by the authors of a paper not made available to others		Requires re-implementation
1	Commercial software	SPSS	Accessible only to those that can afford to use these tools
2	Freely available software, not open source	CiteSpace	Accessible to all; one has to rely on documentation for algorithmic details
3	Freely available software, open source	Gephi	Accessible to all; allows scrutiny of code for correctness and algorithmic details

Another relevant issue is the distinction between short-term and long-term availability. We found that various software tools used in scientometric research are made available on personal websites, which does not seem to guarantee their long-term availability.

Finally, we note that some algorithms (e.g., clustering algorithms) implemented in software tools make use of computer generated pseudo random numbers. To achieve full reproducibility of the results, one needs to work with exactly the same random numbers. This means that the same random number generator with the same initial seed needs to be used. Software tools that do not support this will yield results that can be reproduced only in a statistical sense, and not in an exact sense.

Methods

In the case of all five papers that we reviewed, at least some of the reviewers expressed concerns about the lack of sufficient methodological details to enable full direct reproducibility.

Furthermore, although scientometric research relies mainly on quantitative methods, there sometimes also is a qualitative element in the methods, in particular when a quantitative scientometric method is evaluated qualitatively based on expert judgment. Full direct reproducibility of results obtained using qualitative methods may not be possible. For instance, different experts may have different opinions and even the same expert may not have the same opinion at two different points in time. Nevertheless, when qualitative methods are used, one may at least aim to make sure that the methods themselves are reproducible, even though this does not guarantee that the results will be fully reproducible as well.

Results

Two issues were identified related to the way in which the results of a study are reported.

First, results can be made available at different levels of aggregation. Papers tend to focus on reporting results at an aggregate level (e.g., distributions or summary statistics). This means that even if aggregate results have been successfully reproduced, it is not clear whether results at disaggregated levels have been reproduced as well. When it is considered desirable to reproduce the results of a study even at the most detailed level, results need to be available at this level.

Second, when detailed results are made available online in order to facilitate reproducibility, there is the issue of ensuring long-term availability of the results. This is similar to the issue of the long-term availability for software tools that was discussed above.

Observations regarding threats to conceptual reproducibility

Conceptual reproducibility focuses on the question whether knowledge claims published in a field are found to be robust when tested using an alternative approach with different data, methods, and study design. The scope of our assessment of the status of conceptual reproducibility in the field of scientometrics is very limited, as it is restricted to assessing the contribution that individual papers make through using research designs that are appropriate to the research question being asked, and through formulating claims that are not overselling results but are supported by the evidence that the respective study has produced. However, what is seen as appropriate, related epistemic norms and values, are under constant debate and negotiation in a field, and therefore cannot be handled as a simple checklist. Consequently, scrutinizing the papers against those categories leaves space for divergent judgments.

Operationalization (assumptions, decisions)

The question of data selection and modelling is obviously most relevant for *empirical studies*. Data selection and modelling should be consistent with the research problem a paper tackles. Reviewers did not always agree in their critical remarks. For instance, perspectives on how to delineate a field, or if the choice of a database is appropriate for a certain research question, vary within the scientometrics community. *Method papers* need to argue that the choice of data to demonstrate the value of their data analytic method is suitable to prove that claim. *Conceptual papers* can also contain data issues. In our example, the conceptual paper presented a toy data set and a simulation model - choices made for either can be challenged and gauged against empirical phenomena.

For a *method paper*, the subcategory choice of the analytical method is evidently the most central. We found differences in the extensiveness of how authors introduce into concepts and related methods The reviewers welcomed extensive discussion how to operationalise a certain research question; and if methods used were standard in the field. For the *empirical papers* though there was also critique about using standard tools without critical reflecting about limits of a tool.

To summarize, extensive discussions of the choice of data and methods were positively marked by the reviewers. In some cases, standard methods, tools and datasets were found to be taken too much for granted. A critical view on one's own approach and the articulation of pro and cons in the choices made relative to the specific research question pursued would instill greater confidence in the robustness of the results. The conceptual and method paper scored relatively high here, while the empirical papers in the eyes of the reviewers could have been more explicit or more critical.

Quality control
In this category we look for evidence for measures for quality control that could increase confidence in the robustness of results. For the *conceptual papers* this leads to the questions if choices made are thoroughly detailed. For a *method paper*, for instance the influence of noise in primary data on the methodological analysis can be an important point. For the *empirical papers* in our sample questions about the role of missing data, the exclusion of certain data from the analysis, and the representativeness of a certain method of data collection were posed. The *conceptual paper* and the *method paper* scored relatively well on those criteria, but the reviewers were more critical about the *empirical papers*. Either a discussion of completeness and consistency of data and the choice of parameters was entirely missing; or if present the consequences of such omissions for the argument of the article were not discussed.

Reporting of results
Positive is that all papers in our small sample addressed limitations of their studies, so there was clearly a self-critical attitude present. Remarks of authors on the limits to generalisability of results, the risk of obsolescence of the results when the data services used are changed, or the possibility to use another simulation model were usually appreciated by the reviewers. However, there were critical remarks concerning the extent to which specific claims were backed-up by the empirical results. In particular, *empirical papers* of category 4 (case study) seem to be susceptible to such an 'overplaying of your hands', especially when lacking details when discussing limits resulting from sample size. In the case of category 3 (general) critique on the reporting of limits and overstating of claims was voiced, mixed with doubts about the support the research method (in this case regression analysis) lent to the results.

Discussion

This explorative study generated a number of open questions, offered for further consideration below.

Open questions about rationale for enabling direct reproducibility

A key issue relates to the trade-off between efforts invested in and potential benefits expected from improved direct reproducibility. How do we approach this cost-benefit trade-off? Does this trade-off vary by study type - e.g. do publications that produce (potentially) fundamental contributions to theory or method development deserve a higher level of effort to ensure direct reproducibility than publications of case studies with a limited scope and future applicability?

Another important question relates to the exact purpose of enhancing the direct reproducibility of scientometric research: For instance, is the purpose to screen for error or potential fraud, or is it to allow a third party to build confidence in the reported results by independently reproducing the study? Depending on the exact purpose, efforts made to enhance direct reproducibility may need to be focused in different ways.

Open questions about sharing in order to enable direct reproducibility

Our explorative review suggests that certain issues related to direct reproducibility can be addressed by authors merely improving the reporting of their studies. However, complex procedures that require a lot of detail for full documentation, and tacit components that are difficult to convey in writing, constitute an important challenge.

Beyond improvements in reporting, more contentious is the question what to expect in terms of sharing material resources: the (oftentimes proprietary) primary data used, and the software and tools developed to conduct analyses. Here a number of concerns intersect:

1. What is really needed to enable the direct reproducibility of a study?
    a. When is the ability of inspect source code required, and under what conditions can software tools be accepted to reliably function as black boxes (e.g. a standard statistical analysis tool, a visualization tool etc.)?
    b. When is access to detailed results required? If proprietary primary data has been used, the detailed results underlying the analysis often cannot be shared.
2. When are costs that a third party would incur to reproduce a study, e.g. the cost of re-implementing an essential piece of software or infrastructure or of buying a large-scale proprietary data set, seen as prohibitive and what can be done about it?
3. How should the original team's effort involved in enabling direct reproducibility (and its potential sacrifice of 'competitive advantage') be balanced against the investment needed to be made by another team to directly reproduce the original study?
4. What should be our expectations regarding the durability of access to tools and data that enable direct reproducibility? Is ad-hoc archiving and provision of access through personal websites sufficient, or should we develop strong recommendations towards the use of certified archiving services?

Open questions about standards for assessing conceptual reproducibility

In our limited review of scientometrics publications, we found the technical preconditions for direct reproducibility much easier to assess using a checklist approach than the likelihood of their conceptual reproducibility. Reviewing the articles for issues that may present a threat to their conceptual reproducibility largely mimicked the process and effort of conducting a typical peer-review of a journal article submitted for publication - significantly scaling down the number of publications we had hoped to review in each study type category in the time allotted to this explorative study - which in itself is one of the lessons learned.

The taxonomy was helpful in that it directed our attention to specific aspects, such as the adequacy of study designs and methods, and the adequacy of evidence-based claims. However, reviewers diverged on how to interpret conceptual reproducibility. Conceptual reproducibility deeply touches on epistemic norms and values inside a field. Our discussions centered on how to assess the appropriateness of methods and of claims made based on the evidence produced in the light of unsettled methodological debates in our field. We further observed that there exists a diversity of research designs and methods in our field – how would this arguably productive diversity suffer, if journals take a strong stance on enforcing the use of standard methods? Finally, what is the role of a single article in ensuring conceptual reproducibility, and at what point is a debate to be taken to a wider forum in the community, and if so in what form (i.e., methods sections in journals, controversies addressed at conferences, benchmarking tests in training and education)?

Limitations

This explorative study is only a start to empirically assess threats to reproducibility in scientometric research and to identify critical questions to be resolved in order to operationalize reproducibility for our field. The small, hand-selected sample of publications we reviewed is not representative for the entire range of research published in scientometrics, e.g. in terms study designs, methods, and data used, even though we aimed to account for some of the variation we encounter in scientometrics by our high-level categorization of study types. Therefore, our observations can only provide initial pointers towards threats to the reproducibility of research in scientometrics. Due to the smallness of the sample we could not capture the variation in methods and study quality within each category of study type, nor determine whether the distinction made by our categorization of study type is indeed the most productive one to account for major differences in the type of threats encountered. Also we lack empirical data, what types of research best characterize the large majority of studies in our field. These are all topics for future research.

Conclusion

The approach we tested here to identify reproducibility issues in scientometrics has been to conduct a multi-reviewer exercise by a team of researchers with a variety of methodological and epistemic backgrounds, who were guided by taxonomies of threats to reproducibility. Areas of concern that were identified are the dependence on proprietary data, weaknesses in documenting and critically reflecting on choices made with regard to data sets, methods, and operationalizations relative to the specific research questions asked, omissions to demonstrate the robustness of results against parameter variations, and failures to base claims adequately on the empirical results.

Methodologically, the application of the taxonomies has been challenging, revealing remaining confusion about concepts of reproducibility and the need to further consolidate or explicate such taxonomies for use in reviewing exercises. That said, the discussions around the assessment of features of studies relative to our at times diverging ideals of reproducible research, have been productive in eliciting open questions regarding requirements for reproducibility in the specific context of scientometric research.

For the upcoming STI2018 conference, we suggest to discuss some of the open questions identified in this paper. One of the key questions is the trade-off between benefits and costs of improving the direct reproducibility of published research. Can we identify specific areas or instances where lack of direct reproducibility has undermined scientific progress in scientometrics? How could this have been prevented? And what would have been the benefits and costs of preventing this?

Some of the key questions with regard to conceptual reproducibility are how to operationalize expectations for individual articles with regard to the robustness of their knowledge claims, whether the status of methodological debates in our field allows us to be more prescriptive with regard to the appropriateness of methods, and where such debates are most needed and how they could be best supported.

Finally, scientometric journals along with the peer review process as gatekeepers of what gets formally published in our field, are in a key position to set standards for best practices. Hence in terms of a practical outcome of discussions at the STI2018 conference, we might consider taking steps towards developing guidelines for journal editors, reviewers and authors on good practices to ensure and promote reproducibility of published research.

References

Barba, L. A. (2018). *Terminologies for reproducible research*. arXiv:1802.03311.

Chen, X. (1994). The rule of reproducibility and its applications in experiment appraisal. *Synthese*, *99*(1), 87–109.

Dubé, L., & Paré, G. (2003). Rigor in information systems positivist case research: current practices, trends, and recommendations. MIS quarterly, 597-636.

Eisenhardt, K. M. (1991). Better stories and better constructs: The case for rigor and comparative logic. *Academy of Management review*, *16*(3), 620-627.

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience*, *67*(3), 282–289.

Flier, J. S. (2017). Irreproducibility of published bioscience research: Diagnosis, pathogenesis and therapy. *Molecular Metabolism*, *6*(1), 2–9.

Glänzel, W., & Schoepflin, U. (1994). Little scientometrics, big scientometrics ... and beyond? *Scientometrics*, *30*(2–3), 375–384.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 696–701.

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.

Moons, P., Van Deyk, K., Budts, W., & De Geest, S. (2004). Caliber of quality-of-life assessments in congenital heart disease: a plea for more conceptual and methodological rigor. Archives of pediatrics & adolescent medicine, 158(11), 1062-1069.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411–432.

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.

Van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*, *11*(3), 905–918.