



Universiteit
Leiden
The Netherlands

Bibliometric uncertainty: data quality, reliability, and validity

Donner, P.

Citation

Donner, P. (2026, October 1). *Bibliometric uncertainty: data quality, reliability, and validity*. Retrieved from <https://hdl.handle.net/1887/4310427>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4310427>

Note: To cite this publication please use the final published version (if applicable).

6 General conclusion

The preceding series of four empirical studies was designed to generate new insights on the presence and severity of inaccuracies or errors in bibliometric data and their consequences. They contributed to answering the general research question of the dissertation, which is: *How does data quality affect the reliability of bibliometric indicators?* A particular focus was on the data quality dimension of accuracy as a determinant of uncertainty in bibliometric indicators. Beyond this, the studies addressed also the particular gap in the literature of bibliometrics relating to the quantification of inaccuracy and uncertainty for common bibliometric indicators which originates in data quality. In doing so, they have improved our understanding of the relationship between errors in data and uncertainty in indicator values.

6.1 Summary of findings

The preceding studies have made one thing abundantly clear: while bibliometrics researchers may commonly assert the objectivity, numericity, and accuracy of their methods, if we lift the carpet we find that the underlying basic bibliometric data are laced with the dust of errors and inaccuracies. These inaccuracies, far from vanishing and canceling each other out in large scale analyses, affect bibliometric indicator quantities to a degree that cannot be ignored. This is something that many experienced bibliometricians may have vaguely anticipated. It is the “false precision” of the Leiden Manifesto which ought to be avoided (Hicks et al., 2015). The studies have shown how selected data quality issues are responsible for considerable uncertainty in bibliometric indicators from the level of single publications’ normalized citation scores to the level of aggregate scores of output and impact of research groups and research organizations.

The overall common approach that I used was to assemble original data from bibliometric information systems, as they are used in standard bibliometric studies, and contrast this with either manually or automatically corrected versions of the same data. The observed differences were then statistically investigated to derive error rates and characterize error distributions in some detail. Next, using both the error-affected and the corrected data, I looked at how much the found errors affect typical bibliometric studies and indicators to assess the real-

world importance of the findings. This is an essential task because just knowing that the rate of errors in some bibliographic field is 5 or 10 % by itself does not tell us if the inaccuracies cause any serious problems or can be ignored because bibliometric indicator result values are not affected to any degree that matters. Finally, the approach culminated in a proposed framework for quantifying the uncertainty in indicator values by statistical models that are constructed from the empirically collected error data.

The study of the accuracy and reliability of bibliometric databases' assignments of document type information to records in Chapter 2 found that this information is affected by errors with considerable frequency. For the WoS sample of documents, I found that the error rate for document type assignments is about 6 %. I also found that the document type categories assigned by different databases, in this case Web of Science and Scopus, to the same publications, are not very consistent, even though the definitions of categories are comparable. The category assignments for three major document types and a residual 'other' category for a sample of 722 publications were consistent between databases in 83 % of cases. The influence of these errors on citation scores was also studied. Inaccurate document types of a seemingly small fraction of documents do not only affect citation scores of these incorrectly classified records themselves. In any study which uses normalization of citation counts by document type or selection of records by document type, they affect the normalized citation scores of all records. The reasons for this are that (1) they cause the calculation of inaccurate expected citation scores and (2) lead to the inadvertent inclusion or exclusion of the wrong records from analyses. These findings are complemented by another, more unexpected result. I found that a small share of publications cannot be unambiguously classified into exactly one discrete class of document type because they have characteristics of two classes. The findings of Chapter 2 on document type classification error have already been independently substantiated. A similar study was carried out for a more specific sample of highly cited publications by Russian authors (Mokhnacheva, 2023) which confirmed the main findings of my study. The findings were further corroborated for a sample of Italian publications by Maisano et al. (2025), who also exploited the observed between-database discrepancies to semi-automatically identify possible document type errors and verify them at a much larger scale. The study also motivated researchers at the Chinese Academy of Sciences to check their revised journal impact indicator for improved robustness to document type error (Tong

et al., 2023).

The study on publication months (Chapter 3) has shown that publications which appear earlier in a given year have a citation count advantage over those that appear later in the same year when using conventional year-based citation windows. I argued that the existence of this publication month effect means that conventional year-accurate citation window methodology is biased. I showed that month bias persists even for citation windows of over five years. This is a systematic error and hence, with sufficient data on publication month, can theoretically be eliminated. Publication month data are not always present in citation databases. I demonstrated that publication month can reliably be inferred from journals' issue order information. I showed further that when using publication month data to construct equally long citation windows for papers published in different months of a year, this publication month bias is successfully eliminated, as theoretically expected. This means that month-precision citation counts have higher validity than year-precision citation counts for any latent construct for which they are invoked as indicators. One may wonder if this matters in practice. Bibliometrics is primarily used to statistically analyze larger publication sets, such as all publications of some research performing unit in a discipline in a certain year. When such sets of publications are studied as aggregates, it seems intuitive that there will be a mix of early-months and late-months publications and the bias components of citation scores will cancel out. To test this intuition, I carried out a realistic citation analysis of German research organizations at the level of OECD Fields of Science disciplinary categories and compared average normalized citation scores computed either with (month-biased) nominal 5-year citation windows or actual 5-year (60 months) citation windows. I found that this improvement in citation score accuracy leads to clearly discernible score differences which are in many cases of substantial magnitude. It has to be concluded that publication month is also in practice a nuisance variable which introduces bias in citation analysis, even for analysis of larger numbers of publications.³²

The comprehensive investigation of the data quality of the institutional disambiguation systems of mainstream commercial bibliometric databases in Chapter 4 has shown that these systems are very incomplete in their coverage of

³²From the regression results presented in Table 3.7 it can be deduced that publication month accounts for 2 to 4 % of variance in 3-year logarithmized citation counts of individual papers.

organizations and have in many cases poor precision and recall in the identification of papers that should be attributed to organizations. Many organizations are not covered as disambiguated entities in the platform systems of WoS and Scopus at all, although their papers can be found by organization name searches. My co-authors and I found that the WoS and Scopus systems, for the organizations they cover, on average exhibit adequate precision but worse recall. However, precision and recall for organizations within one system varied greatly such that results for some organizations were much worse than for others. We further found that bibliometric indicator values calculated from the vendor databases' disambiguation systems have only very limited accuracy compared to reference values calculated from German organizations' data disambiguated with a known very high quality. The conclusion from this study is that we recommend additional data processing for any use with policy implications. In summary, the platform disambiguation systems were shown to have insufficient accuracy in elementary bibliometric indicators such as publication counts of organizations to be reliable enough for professional bibliometric use. The main results of this study have been independently corroborated by Selivanova et al. (2019) and Dederke et al. (2024) for other countries.

Chapter 5 established, for the first time, that the error distribution of citation counts (specifically missing citations) due to data quality deficiencies (incomplete citation matching) does not follow a normal distribution, as is common for error distributions in the sciences and engineering. Most papers have no or very few missing citations but a small number of papers have high numbers of missing citations. This means that errors, that is to say, missing citations, are closely clustered into relatively few papers which are highly affected and whose actual citation counts are severely underestimated by the observable citations. In other words, the error distribution is skewed. This is a fundamental empirical finding which links the error distribution firmly to the skewed distribution of citation counts and of most other bibliometric quantities. Chapter 5 also confirmed a number of other studies that worked with non-probability samples by showing for a proper random sample the rate of omitted citations in WoS to be on the order of 4 %. Chapter 5 has gone much beyond the mainly empirical-comparative research of the other parts of the dissertation by proposing, introducing and demonstrating a novel framework for the formal propagation of uncertainty from known inaccuracies, such as data quality problems, via measurement models to final bibliometric indicator values. This framework is

general in that it is applicable to any source and any kind of data inaccuracy that can be empirically studied and quantified, and to any bibliometric indicator. Using this new framework and applying it to a realistic bibliometric research assessment scenario, I found that the bibliometric data quality uncertainty from just two sources of error, missing citations and inaccurate document types, is considerable, both on the level of the normalized citation impact of individual publications, but also on the counts and aggregate citation impact scores of the publication sets of research groups.

Incidentally, as an error rate, can we say anything more, qualitatively, about the magnitude of the figure of about 4 % missing citations? Is it a low or a high value in its context? There is, of course, no objective yardstick to compare this to. But I want to offer the following isolated scalar value as a point of comparison. Four hundred years ago, English mathematician Henry Briggs worked out huge logarithm tables by hand. The error rate of this work is surprisingly low:

Briggs calculated each value up to the thirtieth decimal place, and he wrote them down following Stevin's then still novel notation, but already with a period separating the integer from the fractional part (a practice that Napier had initiated). With the help of the rules of multiplication and division of logarithms, Briggs constructed a table with very dense values. What he was able to calculate in his lifetime he published by 1624. This comprised logarithms of all numbers between 1 and 20,000 and then between 90,000 and 100,000, which he wrote as decimal fractions with 14 digits after the period. Throughout the years, about 1100 mistakes were found in the tables. This amounts to no more than 0.04% of the total entries. In most cases, these were deviations of ± 1 in the last digit, plus some simple typos. (Corry, 2015, pp. 169–170)

To put that into perspective, Briggs' logarithms had an error rate two orders of magnitude lower than modern citation counts. This makes the citation matching error rate appear subjectively high.

To return to the finding of a skewed distribution of missing citations, one important implication of this should be pointed out. The seemingly moderate rate of missing citations of 4 % is an average – an average of a skewed distribution. This means that it is not the case that typically every paper will have 4 % of

its citations being missing. Rather, most papers have no or a negligible number of missing citations. But a small number of papers will have a great number of missing citations.

Having reviewed the main results of the studies, I am now in a position to answer the general guiding research question of the doctoral studies project: *How does data quality affect the reliability of bibliometric indicators?*

In the cases of the error rates for document type assignments and omitted reference links, the empirical estimates may appear to be of small magnitude. But the complementing example studies of realistic research assessment analyses have shown that this impression is misleading. These small error rates cause substantial uncertainty in bibliometric indicator values. The case is more straightforward for the organization disambiguation systems of leading proprietary bibliometric databases. These were found to exhibit such poor performance that their use for any serious research or applied studies was cautioned against.

The overall answer to the main research question is therefore that data quality affects the reliability of bibliometric indicators negatively to such a degree that standard bibliometric indicators are usually very imprecise. They are imprecise in the sense that they are typically not very close to the theoretical value that could be calculated if all data errors were corrected. The broad implication is that bibliometric indicator values should routinely be treated as inaccurate estimates with unknown, but substantial, uncertainty. More implications and recommendations follow below.

6.2 Consequences for bibliometric studies and bibliometric data sources

6.2.1 General implications

The studies of this dissertation have demonstrated unambiguously that indicator values calculated from data by commercial data providers using accepted standard methods are affected by substantial levels of uncertainty. Like any measurement, they come with a certain level of unreliability. The cause of unreliability studied in this thesis is insufficient data quality. This core finding has considerable implications that all providers and users of bibliometric

data, indicators, and reports should be aware of and take seriously. Precisely reported numerical values of indicators should be interpreted as estimates with unknown, unspecified uncertainty. Unless uncertainty estimates are actually present, which would be the much preferable option. My studies into the real-life effects of indicator uncertainty have shown that indicator value differences on the order of perhaps 20 to 30 % for research units with hundreds to thousands of publications, such as the multi-year output of research groups, research institutes, or universities, cannot be taken at face value. Such differences could be caused simply by inaccurate base data.

Bibliometric indicators and measurement methods should be continuously reviewed, modified and improved (Hicks et al., 2015). One methodological improvement whose broad implementation I want to advocate for is to substitute month-precision citation windows for year-precision citation windows to remove month bias in citation indicators (Chapter 3). This change is easy to implement, both for database users and for database providers. It can measurably improve the reliability, and hence also the validity, of citation counts, irrespective of the construct for which they are invoked as a proxy.

6.2.2 Research assessment reform

The debate around reform of research assessment practice and the roles of qualitative and quantitative evidence has increased in scope and activity in recent years. Much of this debate has been implicitly about the level of individual researcher assessment, where a common position of advocates of bibliometric methods is that these methods can be a helpful complimentary source of evidence. At this level, their arguments revolve around the low inter-rater reliability of peer review procedures and their susceptibility to particularistic social biases.

My studies have demonstrated clearly that bibliometrics has its own reliability problem. For example, at the publication level, citation counts are frequently inaccurate and document type assignments exhibit significant error rates. The latter issue means that when publications of selected document types are counted, for example the publication of type ‘Article’ of one researcher, these counts are also affected by inaccuracies. I have not studied the researcher aggregation level, but the next higher level, that of formal research groups. The errors from the paper level have been propagated to the research group level indicator

values and due to their non-normal distributions, they do not appreciably statistically cancel out. The indicator value uncertainty levels strongly affect the interpretation of the results. Under the same or similar error levels, uncertainty ranges for indicator values on the individual researcher level can be expected to be larger in relative terms because of the smaller number of observations at this level. Therefore, the use of bibliometrics in researcher-level assessments with unchecked and uncorrected data from major bibliometric data providers at present data quality levels must be cautioned against.

There are two obvious ways of metrics-advocates to respond to these findings. The first is to substantially improve data quality. The the researcher level, this is to some extent possible. One can let researchers check publication lists for completeness and correctness of document type and one can correct citation counts by systematic cited reference searchers. Similar steps would need to be taken for other data quality issues. That would make the base data as accurate and high-quality as it reasonably can be made. The second way would be to explicitly report the proposed uncertainty ranges for indicator values that derive from data quality and error propagation. However, simply pretending that unreliability is a problem only for qualitative assessment of research and not for bibliometrics can no longer be justified.

6.2.3 Improvements in data quality

From studies such as these presented here, which show that poor data quality seriously thwarts reliable and accurate bibliometric studies and from the emerging competition of open databases, one might hope that commercial providers decide to invest in improving data quality to advance the competitiveness of their products. It would be rational to do so because higher data quality means more reliable bibliometric scores which can serve as a basis for more legitimate and evidence-based decision-making. The continued use of data of too low quality may eventually become a liability if these inadequacies are taken up as an argument against specific uses of bibliometric data and methods with relevance for policy or personnel decision-making. It seems likely that commercial database providers have so far not fully appreciated the possible adverse influence of insufficient data quality on the accuracy of results calculated from their data by users and by themselves. Thus, it may be a worthwhile longer term strategy to implement further data cleaning routines or revise existing ones.

In Chapter 4, my co-authors and I found that research organizations and the research publications that they have produced are often difficult to identify comprehensively and accurately in bibliometric databases. The recommendation that follows from these findings is that the on-board institutional affiliation disambiguation systems of Scopus and Web of Science should not be used without further processing and cleaning until the retrieval performance of these systems has improved. The operators of these databases should continue to improve their systems and publish their internal quality assessments to substantiate claims about their performance.

One aspect of the institution identification problem has been partly addressed by the introduction of the Research Organization Registry (ROR). ROR is a “global, community-led registry of open persistent identifiers for research organizations”,³³ which at the time of writing comprised more than 116,000 organizations. ROR’s data is openly available. ROR solves the issue of the usage of different identifiers across different bibliometric and scholarly information systems as its identifiers can be used without cost and licencing restrictions. OpenAlex and Web of Science have adopted ROR IDs.

6.2.4 Open data, open methodology, and transparency about uncertainty

The fundamental possibility to do any data quality studies of bibliometric data sources was until recently very restricted. Access to comprehensive raw data was only available at a few dedicated research centers because of the high costs and restrictive licenses of commercially produced data sources. These were the conditions under which this doctoral research project was started. Meanwhile, conditions have profoundly changed. Large-scale, comprehensive, open bibliometric data sources have been launched; sources with no licencing costs and permissive use licencing. Consequently, studies of the quality and reliability of bibliometric data are now within reach of many more researchers and users. Data quality studies of completely different scale, range, and openness are now possible with open data.

The availability of open bibliometric data has enabled consultants and contractors as providers of bibliometric reporting services to meet the natural and long-standing, but legally difficult to arrange expectation of transparency of

³³<https://ror.org/>

the underlying data to the evaluated units,³⁴ thereby contributing to accountability and mutual trust. A knock-on effect of open data is that bibliometric methodology can also become more transparent. Bibliometric research groups were hesitant to make their technical methodology openly available for several reasons. One reason was that their technical infrastructure, such as processing routines, were so closely bound up with proprietary data that the disclosure of source code may have violated licensing terms.

A prominent example of a transformation from proprietary to open data in bibliometrics is the Leiden Ranking Open Edition³⁵ which uses OpenAlex data and fully open source code³⁶. Another example of improved transparency in bibliometric research methodology is that the KB affiliation disambiguation system, used extensively in Chapter 4, has been open sourced as one early step of a larger movement of the German bibliometrics and science studies community towards open science practices, reproducibility and transparency (Lenke & Taubert, 2024).

This development towards open bibliometric data may seem to somewhat diminish the long-term importance of completed data quality studies of proprietary bibliometric databases. However, studies of proprietary data sources such as these presented here, retain their usefulness in that their results can serve as points of reference for the comparison of the data quality of open bibliometric data. Given the not very high data quality observed in the empirical analyses in this project, the expectation for open bibliometric data sources should be to eventually surpass the data quality of commercial products. The open availability of bibliometric data is a great advantage but this advantage can only be fully realized and utilized if the quality of the data is sufficiently high to meet the quality standards of the research community and of the community of decision-making users of bibliometric reporting products. Consequently, at this moment, there is a lot of research on the data quality of the foremost open bibliometric database product, OpenAlex, with presently a strong emphasis on data completeness (Céspedes et al., 2025; Culbert et al., 2025; Haupka et al., 2024; Hauschke & Nazarovets, 2025; Rodrigues et al., 2025; Scheidsteger et al., 2025; Zhang et al., 2024). This shift of attention of empirical data quality studies from proprietary to open data sources seems to be not only a worthwhile

³⁴San Francisco Declaration on Research Assessment <https://sfdora.org/read/>

³⁵<https://open.leidenranking.com/>

³⁶<https://github.com/CWTSLeiden/CWTS-Leiden-Ranking-Open-Edition>

investment of the research community's effort but also an early indication of an imminent change in the usage patterns of data sources of the scientometrics and science studies community.

A crucial development precipitated by open bibliometric data is the possibility of transparency of bibliometric studies, including both full reproducibility of research work and full disclosure of basic data and calculations for bibliometric research assessments. With this possibility comes of course the expectation that such transparency is the default mode. Heretofore, licensing terms of commercial bibliometric data subscriptions have precluded such transparency. This condition may have contributed to concerns about the legitimacy of bibliometric evaluations since fundamental data could not be independently verified. Clearly, transparency is highly desirable. I believe that studies of data quality, reliability, and validity can contribute further to transparency. Such studies can complement basic data transparency by demonstrating the uncertainty in indicator values which is caused by data inaccuracies and by imperfect validity of bibliometric variables as indicators for theoretical constructs which cannot be directly measured. They can provide transparency about uncertainty. Such a contribution should not be underestimated.

6.2.5 Methodological research in bibliometrics

At present, some part of bibliometric methodology research seems to be caught up in minor, incremental development, improvement, and rediscovery of specific indicators or indicator variants. The bulk of these studies seems to have been written by authors in a rush to propose their own indicator specification without much consideration for the basic reliability and validity of the novel indicator (Bornmann & Marx, 2018). Some topics to which this characterization in my opinion applies are h-index research (Wildgaard, 2019), bibliometric counting methods (Donner, 2020; Gauffriau, 2021), and bibliometric interdisciplinarity measurement (Q. Wang & Schneider, 2020). There seems to be a certain level of dysfunction in the self-organization of this research. Namely, that it is considered a sufficient knowledge contribution to the field when a novel indicator or indicator variant is proposed without so much as to demonstrate improved validity over existing standard indicators (for existing constructs) or basic validity evidence being presented (for newly introduced concepts and indicators). This lenient convention may need to be addressed sooner or later in the interest

of cumulativeness of knowledge. The introduction and adoption of unverified methods may lead to wasteful research – research whose results do not hold up to critical scrutiny because the methods, indicators, tools and so on never actually measured what they alleged to measure (validity) or did so only so poorly as to make the results useless (reliability) (Whitley & Kite, 2012). I recommended that the research community of quantitative science studies adopts the theoretical and methodological apparatus of reliability and validity of psychology and psychometrics (Hunter & Schmidt, 2004) and continues to develop and adapt it for its own needs.

6.2.6 Adoption of explicit uncertainty reporting

A novel framework to calculate uncertainty intervals for arbitrary bibliometric indicators which communicate the uncertainty due to data quality issues was introduced in Chapter 5. For bibliometric studies of critical importance, a further developed model of the introduced novel uncertainty propagation and quantification framework should be applied. This should be built on a solid foundation of empirical error studies for the relevant corpus and data source, which is likely to mean that new data on error distributions has to be collected, covering more metadata fields than investigated in the studies of this project. Moreover, to rule out misguided conclusions from earlier studies, some of them may need to be replicated with the application of a fully worked out bibliometric uncertainty framework that includes data quality and indicator validity as sources of uncertainty.

Table 6.1 gives a concise overview of these and further recommendations organized by the different target groups.

Table 6.1: Summary of recommendations

Indicator	Recommendations
provider/user	
group	

bibliometric data
and analysis
platform providers

- intensify efforts to increase data quality (citation matching, institution affiliation disambiguation, document types, publication months)
- enhance transparency of data quality assurance processes and their results, for example by publishing internal data quality audit reports and benchmark data sets
- offer easily accessible tools for correction and curation of data by user community
- eventually build uncertainty ranges into systems for basic indicators (e.g., ranges of probable citation counts for individual items)
- use month-based citation windows for analysis platforms

bibliometric
research and
evaluation centers

- use month-based citation windows
- continue to conduct independent empirical studies and audits of source data quality, focussing increasingly on open data
- begin to include data quality-based uncertainty ranges for key indicator values
- further develop the framework for uncertainty propagation and expression

research managers,
evaluators, and
other
decision-makers

- build awareness of bibliometric indicator uncertainty and do not base decisions on data and indicator values whose uncertainty from data inaccuracies is unknown
- exercise caution in the interpretation of results if numbers of observations are low and indicator differences are not large
- request uncertainty estimates for key indicators values from data providers and reporting contractors
- report data errors and engage with bibliometric data providers to make corrections

bibliometric
researchers

- exercise caution in the interpretation of results if numbers of observations are low and indicator value differences are not large
- use month-based citation windows
- elevate the standards for the reporting of reliability and validity of bibliometric indicators in research publications
- collect and publish reliability and validity evidence alongside novel indicators and indicator variants
- report data errors and engage with bibliometric data providers to make corrections

6.3 Limitations

The specific limitations of the studies have been discussed separately in the preceding chapters. But there are some broader limitations that apply across

studies that should also be mentioned. The studies presented in the substantive chapters have looked at the practically relevant effects of data quality on indicator values primarily at the intermediate size level of research performing units, namely research groups and organizations. Bibliometric research is also done on the level of individual researchers and on the level of countries or groups of countries, which were not considered in this dissertation but would be relevant extensions for similar research.

Metrics-based assessment of individual researchers, while fiercely contested on methodological grounds, is nevertheless not uncommon. At this level, one of the main advantages of bibliometrics is absent: the statistical robustness of indicator values that large numbers of observations bring about. As individual-level indicator values are inevitably volatile, the influence of inadequate data quality can be expected to be quite large, in other words, their measurement reliability may be quite low. Apart from evaluation purposes, this possibly low reliability may have important ramifications for basic research that are not always recognized. If indicators such as numbers of publications and numbers of citations are affected by measurement errors and are used as proxy variables for unobservable traits of researchers, such as productivity, impact, esteem, or influence, then any associations, like correlations or regression coefficients, will be underestimated because of the random component in the measured values. On the other hand, the much larger number of observations usually available for bibliometric research at the country level would lead one to expect that imperfect data quality may have smaller or even negligible effects here. Yet, this needs to be positively established by empirical research first.

In this dissertation I have argued and demonstrated that poor data significantly impair the reliability and validity of bibliometric indicators. Yet all of this research has been done within the confines of the data that are actually indexed and available in proprietary bibliographic-bibliometric databases. Their coverage of primary research publications is known to be selective and incomplete and the selectivity criteria and their application have been subject to debate (Fenton et al., 2024; Hicks & Wang, 2011; T. N. Van Leeuwen et al., 2001). Nor is the coverage in any way based on random sampling which makes inferential statistics inapplicable as a possible method to generalize from the indexed content to the not indexed content outside of the databases. This confinement to indexed literature masks a major threat to validity that cannot possibly be accounted for in the framework of this research, namely the completeness of the literature in

the available data that is relevant to a given bibliometric study. If the target of a study is as ambitious as all the global literature of basic scientific or scholarly solidity then whatever results one obtains from the major proprietary databases will be highly biased by scientific discipline, by geographical origin of scholarship and by language. Possibly also by the type of publication. For instance, Scopus covers many non-English language journals and many scholarly books but practically all books that it covers are English only³⁷. Thus, it should be noted as a considerable limitation of this research and of other research that similarly relies on these databases, that the estimates of accuracy of data and reliability of indicators are conditioned by this delimited subset of research literature and are necessarily distorted with respect to a representation of the entire global research literature. Such research into database data quality and indicator reliability is always database-specific and cannot simply be generalized beyond the data sources.

The data that are indexed in bibliometric databases are not static, in fact, they are frequently updated. Most updates are newly indexed content but occasionally already indexed metadata are amended or corrected. One would expect that updates to older data are improvements that tend to fix inaccuracies rather than introducing new ones. But this is by no means guaranteed. It is therefore a further limitation that in this dissertation research, the data has been treated from a slightly unrealistic perspective of assuming some longer term stability. It is presently not known if this assumption is justified or not.

Throughout the studies, care was taken to design them in such a way that findings generalize as much as possible, yet some of the results should be considered to generalize better than others. The studies on document types and citation count errors used random samples within specified constraints. Document type inaccuracies and institution affiliation disambiguation were studied across two major commercial bibliometric data sources. Beyond these data sources, error rates for these metadata could be quite different. An important limitation of the study of the affiliation disambiguation systems was that only gold standard data on German institutions was used. The science systems of countries are known to differ in their complexity which likely affects error rates in disambiguation across countries. On the other hand, the results on publication month effects are conceptually independent of data source, scientific discipline, country, and other factors and can be assumed to be reproduce under any conditions.

³⁷See Scopus Book title list available at <https://www.elsevier.com/products/scopus/content>

6.4 Further research directions

From the above-mentioned limitation of my research that it does not fully do justice to the dynamic nature of database content one obvious methodological recommendation for similar studies is to design such studies in a way that makes it easy to repeat them after some time and compare the earlier with the newer results.

There is a certain flexibility in bibliometric practice about which document types to include and exclude in bibliometric studies and in research evaluation and how to handle the interaction of document type and citation count normalization. Some studies and research groups routinely analyze only research articles as assigned in the data source and calculate reference citation count values for normalization on this basis. Others use both ostensible articles and reviews as one undifferentiated kind of substantive research papers and calculate a common normalization score. Yet others may include articles, reviews, and potentially other types such as conference papers as separate categories and treat them distinctly with their own reference scores. For data sources with incomplete or very unreliable document type data, another possibility is the selection of substantive comparable papers based on combinations of criteria such as presence of named authors, a certain page length threshold and number of cited references. It would be an important contribution to systematically study if these different options have a substantial influence on obtained results and if some options are affected more than others by document type inaccuracies.

I have pointed out above the implication of the skewed distribution of missing citations that most publications will have no or few missing citations but a small number of papers will have many missing citations. Table 5.2 showed that of the sample of 372 papers, the majority, 263 papers (71 %), had no missing citations, a further 67 (18 %) had just one missing citation. However, one paper had 15 missing citations and another one had 26. In the first case, this works out to 51 % missing citations and in the second case 65 % missing citations. In my opinion, rather than seeking comfort in the fact that most papers are not affected by substantial numbers of missing citations, the result that missing citations are highly concentrated on a very small fraction of papers should motivate more research into the consequences of this finding. This result is most likely caused by false negatives in citation matching algorithms. One possibility to progress on this is to exploit the results of different citation matching algorithms on

the same data (Franceschini et al., 2013; Olensky et al., 2016; Rodrigues et al., 2025). If these comparisons can identify a sufficiently large number of records with many missing citations, a qualitative analysis of the reasons for matching failure in these cases might produce very valuable insights on the basis of which matching algorithms could be improved.

From the analysis of the original random sample of missing citations in Chapter 5 and various earlier studies using non-representative samples reviewed in that chapter it is clear that citation counts are subject to non-trivial error rates. One problem in developing improved reference matching algorithms, which are the origin of citation counts, is the inevitable trade-off between precision and recall. If an algorithm is made more lenient in order to match more references, such as those with several minor errors, that also means that it will produce more false positives and produce too high citation counts for some papers, as superficially similar reference strings are linked to papers which actually refer to different papers. There seems to me to be one reasonable way to handle this trade-off transparently that citation database providers might want to experiment with. Reference matching algorithms can be designed in a way such that all match candidates that exceed a specific similarity threshold are accepted as matches, where the similarity may be a complex heuristic or a scalar parameter value. Rather than setting one specific threshold (or decision rule) which provides an acceptable balance of precision and recall one could also set two thresholds, one with high recall and still acceptable precision and one with high precision and still acceptable recall (or more than one decision rule). One would then obtain for some papers two different but plausible citation counts, effectively giving a range of plausible citation counts for each publication record which communicates some of the uncertainty inherent in a reference matching algorithm. Since base scholarly data and derived indicator values are uncertain, a reasonable way to move forward is, for every party involved, data producer and users, to always be upfront about which uncertainties they are aware of and how large they are. If a scholarly database uses a reference matching algorithm that cannot determine with complete certainty if one cited reference refers to an indexed document, then this should be visible to users. Citation counts can be displayed as ranges of likely correct values based on fuzzy reference matching results.

The uncertainty propagation framework borrowed from the physical sciences introduced in Chapter 5 can be extended with some further work in an important

direction. While in this dissertation only the type of uncertainty which is attributable to reliability was considered, the uncertainty propagation framework is not restricted to this. Suitably extended, the framework can include data on construct validity (Carmines & Woods, 2005; McDonald, 2005). To this end, the type of outcome quantity has to be changed from directly measurable variables to theoretical latent constructs proxied by one or more measurable variables. The combined assessment of these two cardinal sources of uncertainty, data quality and indicator validity, is proposed as a critically important topic of follow-up research.

The research reported in this dissertation can hopefully contribute to a rational discussion on the opportunities and limitations of bibliometric research methods and the application of bibliometrics to research evaluation. Bibliometric indicators can serve as valid proxies for theoretically and practically relevant concepts if they have been demonstrated to show sufficient construct validity and reliability. I have shown here how much classic bibliometric indicators' reliability depends on data quality and how to propagate and express this reliability component of total indicator uncertainty as uncertainty ranges for indicator values.