



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).

2 Document type assignment accuracy in the journal citation index data of Web of Science

This chapter was first published under the same title in *Scientometrics*, vol. 113, pp. 219–236, at <https://doi.org/10.1007/s11192-017-2483-y>, on 04 August 2017.

Abstract

This article reports the results of a study of the correctness of document type assignments in the commercial citation index database Web of Science (SCIE, SSCI, AHCI collections). The document type assignments for publication records are compared to those given on the official journal websites or in the publication full-texts for a random sample of 791 Web of Science records across the four document type categories articles, letters, reviews and others, according to the definitions of WoS. The proportion of incorrect assignments across document types and its influence on document specific normalized citations scores are analysed. It is found that document type data is correct in 94 % of records. Further analyses show that within records of one document type as assigned in the data source, the records assigned to the type correctly and incorrectly have different average page counts and reference counts.

Keywords

citation normalization; document type; data accuracy; bibliometric data; citation impact; Web of Science; Scopus; data quality

2.1 Introduction

Scientific journal publications can be classified by their intended information transfer function into distinct document types (also referred to as DTs in the following) such as research articles, reviews or letters. The DT data in citation index databases contain useful information about scientific publications that is routinely used in bibliometric studies.

The specific importance of DT in bibliometrics lies in their usage for 1) defining included and excluded publications in analyses and 2) in creating appropriate reference sets for calculating expected values for normalized citation scores and, less commonly, 3) in weighting publications according to their DT.

Any studies making use of the document type data have so far implicitly assumed that the document type data of commercial citation index databases is sufficiently accurate for these purposes. However, to the knowledge of the author, there is no published study in which this assumption was tested as of yet.

The reason for making distinctions across publications according to their DTs in scientometrics and research assessments is that because of their specific purposes and contents, they are utilized, that is, read and cited, differently, which leads to different citation distributions (e.g. Braun, Glänzel, & Schubert, 1989b, p. 392, Table 6; Lundberg, 2007, p. 156, Fig. 2; Vinkler, 2010, p. 176 ff.). In a thematically homogeneous set of publications, reviews are cited on average more often than articles, which are cited more often than letters (Glänzel, 2008, p. 14). Publications of different document types also exhibit different onset and decay in the distributions of citations over time, as well as in average “citation speed” (J. Wang, 2013). This leads to very different results in calculations of citation-based indicators when constructing reference sets by DT and when disregarding this kind of normalization (Moed & van Leeuwen, 1995; Sirtes, 2012). Therefore, when computing relative indicators of citation impact, the correctness of the assignment of document types to publications is crucial for fair comparisons and valid results. Furthermore, as DT is also commonly used to restrict publication sets from bibliometric datasets to those containing primary research results, DT data quality is highly important for obtaining correct publication counts for research units. As an example, in a comparison of pharmacology and pharmacy journals indexed in Web of Science and Scopus, Gorraiz & Schloegl (2008) found that in one fifth of the titles, the sum of articles and reviews differed in the two databases by more than 10 %.

There is no common standard of DTs in scientific publishing. Publishing houses, journals, and databases all use their own nomenclatures and definitions. The producers of Web of Science and Scopus do not directly copy the DT information of journals. Both producers maintain fixed systems of DTs for use in their products which can be accessed in the respective database documentation¹². How the internal DTs are exactly assigned based on these systems is not appar-

¹²For Web of Science: Web of Science® Help. Searching the Document Type Field. [accessed 2016/10/07] http://images.webofknowledge.com/WOKRS59B4/help/WOS/hs_document_type.html. For Scopus: Scopus Content Coverage Guide Jan. 2016; page 10. [accessed 2016/10/07] https://www.elsevier.com/___data/assets/pdf_file/0007/69451/scopus_content_coverage_guide.pdf

ent from the documentations. The correctness of DT assignment in commercial citation indexes has been called into question. Van Leeuwen et al. (2007) draw attention to the treatment of letters and ‘research letters’ from medical journals that were considered to be the same type in Web of Science (WoS). Spodick & Goldberg (1983) categorized samples of letters to the editor in four general and four specialist medical journals, published in 1979, according to their function. In the specialist journals these were comprised mainly of letters concerning articles published earlier (79 % of letter items). Very few items were replies to such letters. Letters in response to editorials and “letters presenting initiatives”, which often reported primary research in brief form, were also found. In the general journals as much as one quarter of letters was of the type “presenting initiatives”. These results indicate that a sizable proportion of items in letter sections do not address previously published research or continue such discussion through author replies. Across the eight journals it was found that within letters concerning published articles, 11-45 % of the items were criticism, 4-45 % were replies to such criticism and 5-22 % were supportive or supplemented the findings.

A similar study of Tierney et al. (2015) scrutinized the roles of letters in four otorhinolaryngology specialty journals published in 2012 (92 letter items). Responses to original articles accounted for 33 % of letters, author replies to them accounted for 20 %. Responses to letters made up 9 % and letters not related to original journal material 38 % of the total. For example they find:

Clinical Otolaryngology [a journal] was the most varied when it came to letter theme with 55% being unrelated to original journal material, of which technical innovations were to the fore, representing 13.2%. 9.4% (n = 6) were pilot studies, with literature reviews representing 7.5% (n = 4). Other categories’ letters fell into were spontaneous discussion/educational topics, new tools and case reports. (Tierney et al., 2015)

These two studies illustrate that the letters sections of medical journals contain a variety of material, a part of which is discussion of articles, a significant part is wholly different in content.

Another point of contention is the correction of assignment of review articles. Sigogneau (2000) and Harzing (2013) illustrate how WoS was using some highly

questionable assignment criteria for classifying publications as reviews in the past. Another case in point are the discrepancies in the number of items of DT article and review found by Gorraiz & Schloegl (2008) for the same journals across the two investigated data sources. As DT data is important in practice and its accuracy has been questioned in the literature, there is a need for an assessment of the correctness of DT data in citation index databases.

Besides document type errors, there are numerous other classes of error in bibliometric databases which threaten the validity of results obtained from these databases. To name but a few striking observations discussed recently in the literature, Valderrama-Zurián et al. (2015) found a surprisingly high incidence of duplicate publication entries in Scopus. Extensive studies on omitted citations have been performed by Franceschini and colleagues. Using a method that uses WoS and Scopus data in parallel (Franceschini et al., 2013) they obtain rates of missing citations on the order of approximately 5 % in these sources. This allows them among other things to indicate the accuracy of journal impact indicators (Franceschini et al., 2015b). In the same paper it is shown that journal omitted citation rates not infrequently reach figures above 10 %, in the extreme even up to 40 %, in the studied field of manufacturing. They furthermore report large numbers of duplicate DOIs, which are commonly used to match external records to bibliometric database items (Franceschini et al., 2015a). Further evidence is provided by Gorraiz & Schloegl (2008), who noted large discrepancies of the number of articles and reviews of the same journals indexed in both WoS and Scopus.

The overall impression of considerable negligence of data quality increases concerns about the basic validity of any bibliometric study using these data sources, independent of disputes regarding conceptual foundations.

In this article, DT assignments for a sample of WoS records are compared to those of the publishers on official journal homepages and in the original publications' full-texts, using WoS' own definitions of DTs. We also compare the DT assigned in WoS to the DT assigned in Scopus to the same publications by linking the two datasets for the WoS-sampled records. As we found a considerable proportion of disagreements between our independently coded DT and those of either data source, we also investigate the effect that this data inaccuracy will have on typical bibliometric studies. To this end we calculate two field normalized citation scores for each publication with an incorrect DT, one based on the DT as available in the data source and one with the corrected DT. We use the

distribution of score differences to shed light on this possible distorting source of error. By reporting on the error distribution across the four DT classes, we hope to enable researchers and practitioners in bibliometrics to better judge the influence of data inaccuracy on their results.

2.2 Methods

2.2.1 Data and Sampling

For this study, data licenced from Thomson Reuters¹³ Web of Science (WoS) and Elsevier's Scopus and loaded into custom SQL databases was used. The WoS data is comprised of the collections Science Citation Index Expanded, Social Sciences Citation Index and Arts and Humanities Citation Index and is the primary object of study and the data source actually sampled. The Scopus data linked to the sampled records is used only in a supportive role for some comparisons but is not sampled and hence not studied in depth. The WoS sample was restricted to publications published in journals (no proceedings papers, books or book chapters were considered), and restricted to publication dates between 2002 and 2014. The distribution of WoS-assigned DT in journals for that period is: articles: 67.6 %, letters: 2.6 % reviews: 3.5 %, others: 26.3 %. The third restriction is that only those records that had an assigned DOI in WoS were retained in the sample. The reason for this is that the DOI is used as one of several data fields for linking WoS records to Scopus records to make comparisons across data sources feasible. As a consequence, only relatively recent publications are sampled, see Appendix B for the distribution of publication years of the sampled records. There is one complication to this. It was reported that DOI in citation indexes are not as unique as they should be by definition (Franceschini et al., 2015a). For this reason, before the sampling was carried out, both databases were checked for DOIs that occur more than once and the combined list of those DOI was used as a further exclusion criterion for the sample so that no duplicate DOI would lead to ambiguities in the matching step¹⁴. The population size resulting from these restrictions is approximately 11.07 million records.

¹³The department of Thomson Reuters producing the citation indexes recently changed ownership and now runs under the company name Clarivate Analytics.

¹⁴For our WoS data (publications from 1980 to 2014) this list of multiply assigned DOI contains 19,096 entries and the list for Scopus data contains 81,715 entries (publications from 1996 to 2014).

Random samples of document identifiers were drawn from the WoS database, stratified by document type as assigned in WoS after four document type groups were formed. The groups are the types *article*, *review* and *letter*, which were used directly as available in the data, as well as records not assigned to one of those three types, in this study referred to as *other*. Subsamples of the document types were drawn in approximately equal numbers, using the WoS-assigned DT. It was attempted to evaluate approximately 200 publications of each of those four categories to get a reasonably sized sample with subsets of similar size. This allows for an overall estimate of accuracy with a 5 % margin of error. It is necessary to oversample the more rare DTs to be able to make accurate estimates of the error rate within these subsamples. Stratification by WoS DT was used because articles account for the vast majority of all publications and we were interested in estimates of the correctness of DT assignments of the population and the four relevant DT class subsets. The grouping into articles, letters, reviews and others follows the convention of distinguishing between primary research results (articles), secondary, synthesizing overviews (reviews), discussion about published results (letters) and others, such as editorial content or book reviews. The latter are usually deliberately excluded from bibliometric analyses, because they do not relate closely to research. This is necessarily a simplification. Editorials and book reviews can in some contexts be considered items relevant for bibliometric studies (Van Leeuwen et al., 2013; Zuccala & van Leeuwen, 2011).

The sampled records were linked to the Scopus records detailing the same documents using a matching table constructed for this purpose. The matching procedure is described in Appendix A.

To summarize the sampling process:

1. restrict WoS data to journal publication records
2. restrict to records from 2002 onwards
3. restrict to records with unique DOI
4. independently sample the four WoS/synthetic DT categories article, review, letter and “others” (not of the former three types) on the order of 200 records each.

2.2.2 Data collection

Bibliographic description data of each sample record comprised of article title, first author family name and initials, publication year, journal name, volume, issue and first page were queried from the WoS data and saved along with internal record IDs into a separate table. The table rows were randomized and the data exported as a spreadsheet file. The author and one student assistant independently searched for the publications' abstract web pages and full-texts online, using the bibliographic information to query academic and web search engines, blind to the DT in the original data. On the article web page of the journal or in the full-text, we attempted to find the officially assigned document type, if specified, and initially coded it as *article*, *letter*, *review*, *other*, *document not found*, or *ambiguous*. The individual DTs collected into the synthetic "other" category in this study are not distinguished as they are of only minor interest to bibliometrics, which is primarily concerned with publications providing scientific results. If no document type was stated, all useful information such as title, abstract and full-text was taken into account and used to decide on the DT to be assigned. We did not rely on journal section headings exclusively, only in combination with other evidence. The source URL and the term used to describe the DT in the source were recorded. When we found an article from the field of medicine, we also recorded the fact whether the publication is a case report or not.

The definition of document types used for the independent assignment exactly followed those of the two data source providers, which are sufficiently similar on the level of the four classes used. To summarize, an *article* is a report on original research of any length, including meta-analyses. Proceedings papers published in journals are also included in this class. A *review* condenses many previously published original research findings on a specific topic. A *letter* is a correspondence with the editors and readers of a journal about an item previously published in the same journal. The special rest category *other* comprises everything else, for example editorials, book reviews, meetings abstracts and corrections. This includes comments on papers published in different journals and invited discussion contributions that follow focus articles.

After the results were obtained the assignments of the two raters were compared side by side and any disagreements and remaining uncertainties were resolved by reviewing any available information and further searching for clarifying in-

Table 2.1: Sample description - DT frequencies

DT	independent assignment	WoS	Scopus
Article	294	204	279
Letter	111	193	177
Review	165	185	166
Other	208	209	100
Hybrid	13		
Total	791	791	722

Note: document type category “Hybrid” only applies to the results obtained by independent assignment

formation when necessary, including all cases when at least one rater indicated a record as *not found* or *ambiguous*.

While initially a total of 793 publication records were assessed, two records had to be excluded from the analysis because no certain DT assignment was possible as no reliable information for the publications could be found. This resulted in a final sample of 791 cases. The basic properties of the sample are described in Table 2.1 and Table 2.10 (Appendix B). In 13 cases we found it not to possible to assign the publication to strictly one of the DT categories exclusively. Instead, they appeared to be hybrids between two types. Consequently, it was decided that they will be recoded to have two permissible DTs, such that they are counted as correct in the comparison if the DT in the citation indexes is one of these two DTs. These publications comprise one “article/letter”, 4 “article/reviews” and 8 “article/others”. Of the final sample, 722 records (91 %) could be matched successfully with Scopus records.

2.2.3 Data analysis

The collected data was compared with the DT assigned by WoS and Scopus, respectively, and agreement was coded as 1, disagreement as 0. The transformation of WoS DT classes into our four relevant DT classes was straightforward. In the case of Scopus data, records coded as “short survey” were recoded as *reviews* (15 cases), as their description in the Content Coverage Guide explains that they are a shorter type of review. Records coded as “proceedings paper” that were published in regular journals are recoded as *articles* (19 cases) and

one record coded as “in press” was recoded as an *article*.

The data quality of the DT field was assessed by the share of correct DT assignments in the databases. Weighting by population proportions on the DT strata (as coded in WoS) was used for statistical inference from the sample to the population which is possible only for WoS, being the source of the sample. The population frequencies for the values of the variable are known from the original data. Overall and DT category inferential statistics were calculated with the *survey* package for the R statistical programming environment, version 3.3 (Lumley, 2004).

The measure of Precision, as used in information retrieval evaluation, serves as the main indicator of DT assignment correctness in this study. In information retrieval, Precision is defined as the proportion of retrieved documents that are relevant to a query among all returned documents. Analogously, Precision is understood in this study as the proportion of correctly assigned records among all records assigned a specific DT and can be calculated by the following general formula (Baeza-Yates & Ribeiro-Neto, 1999, p. 75)

$$P = \frac{\text{number of relevant records retrieved}}{\text{total number of retrieved records}},$$

or in the terms of this study

$$P = \frac{\text{number of correctly assigned records of a specific DT}}{\text{total number of records assigned to a specific DT}}.$$

In the case of Precision calculation only, hybrids are also counted as correct. The applied method of calculation of the strata weighted confidence intervals follows Korn & Graubard (1998). To estimate the proportion of correctly coded DTs in the population across all DT categories it is necessary to take into account the stratified sampling design. This is accomplished by defining the sampling design using the four strata groups and their population proportions in the *survey* R package and then computing the overall mean of agreement between database-assigned and independently assigned DT.

It is similarly possible to estimate population values for the other main quality criterion from information retrieval, Recall. In information retrieval, Recall is the proportion of retrieved documents that are relevant. In this study this is equivalent to the proportion of correctly assigned records among all records that

actually have that DT. If there are assignment errors, those records cannot be ‘retrieved’ under their correct DT. The basic formula for Recall is

$$R = \frac{\text{number of relevant records retrieved}}{\text{total number of relevant records}},$$

which in terms of this study becomes

$$R = \frac{\text{number of correctly assigned records of a specific DT}}{\text{total number of records of a specific DT}}.$$

For Recall calculation, 14 records assigned as hybrid DT were excluded, as there is no sound way to use them. It is not appropriate to directly calculate *sample* Recall values in this study, because they have little relevance in light of the uneven population proportions of DT categories as opposed to the approximately equal strata sample sizes based on WoS DTs. Since records were sampled with probabilities different to their population frequencies, the calculation of the population Recall estimate needs to be weighted accordingly. The modified formula, following the method for estimation of population proportions from stratified sampled data (Lohr, 2010, p. 80-81), is

$$R'_d = \frac{\pi_{d,d} \times w_d}{\sum_{i \in D} \pi_{d,i} \times w_i},$$

where subscript d denotes a specific document type from the set $D = \{A, L, R, O\}$ of all document types (strata), $\pi_{d1,d2}$ refers to the proportion of records of DT $d1$ in the stratum of DT $d2$ and w_d is a weight derived from the population proportion of a DT d . For instance, the Recall for article DT records in WoS is calculated this way:

$$\begin{aligned} R'_A &= \frac{\pi_{A,A} \times w_A}{(\pi_{A,A} \times w_A) + (\pi_{A,L} \times w_L) + (\pi_{A,O} \times w_O) + (\pi_{A,R} \times w_R)} \\ &= \frac{\frac{197}{204} \times 0.675}{(\frac{197}{204} \times 0.675) + (\frac{69}{193} \times 0.026) + (\frac{10}{203} \times 0.263) + (\frac{19}{181} \times 0.035)} = 0.96 \end{aligned}$$

For a simplified illustration of the calculation please refer to Appendix C.

2.3 Results and discussion

Qualitative findings

There is no standard for the designation of document types in scientific publications. Many different designations for the same DT were observed on different journal web pages and publication full-texts in collecting the data. This creates ambiguity, as the following examples of alternative designations which were found for articles, reviews and letters illustrate.

- Article: original research article, short report, short communication, full paper, original investigation, rapid communication, communication, progress report, case study, brief report, research letter, method article, clinical report, letter (“reports an important novel research result, but is less substantial than an Article” in the journal *Nature Biotechnology*)
- Review: mini review, critical review, review article, survey
- Letter: comment, correspondence, letter in reply, communication to the editor, response, discussion paper (“critical comments on papers already published in the Journal” in the journal *Electrochimica Acta*)

This list is not exhaustive and many variations of the terms occurred. The examples also show that merely using the term as given by the journal is not sufficient as these can be ambiguous and are not standardized, which is in line with the conclusions of Montesi & Mackenzie Owen (2008).

The publication of short original research findings and case reports as “research letters” addressed to the journal editors is common in medicine and closely related disciplines, confirming the findings of Van Leeuwen et al. (2007). Because they report original research, they are considered *articles* in this study. This follows directly from the official definitions of the two data sources in their documentations, which state that case reports are considered as articles in terms of their DT. Furthermore, many journals make distinctions between articles based on publications of completed research and manuscripts presenting preliminary results or specifying different DTs simply based on the length of the paper (“full length paper” – “brief communication”). A similar distinction was also observed for reviews in some cases (“full review” – “mini review”). Special types of articles are often treated and flagged separately, such as case studies, meta-analyses, methodology articles and tutorial papers. This information is useful for bibliometrics (Barrios et al., 2013; Patsopoulos et al., 2005; Romero et al., 2009). For example, case reports make up a significant fraction of what are presently coded as *articles* in medical journals. If they were identified distinctly it would

Table 2.2: Independent DT assignment vs. that of WoS

independent assignment	WoS				Total
	Article	Letter	Other	Review	
Article	197	69	10	19	295
Letter	1	105	5	0	111
Other	1	17	185	5	208
Review	5	0	3	157	165
Article/Letter Hybrid	0	1	0	0	1
Article/Other Hybrid	0	2	6	0	8
Article/Review Hybrid	0	0	0	4	4
Total	204	194	209	185	791

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

be possible to attempt to reduce noise in indicators by excluding them from analyses or treating them separately, as they are rarely cited (Patsopoulos et al., 2005). In fact, treating them separately might be appropriate as there are a number of journals which only or mostly publish case reports. At present, this is possible by using PubMed data, but not by using data from WoS or Scopus only.

Descriptive results

The complete result for the independent assignment compared to those of WoS for the same publications is given in Table 2.2 and the result for Scopus in Table 2.3. In the case of the WoS data, 655 of 791 cases were assessed as correct (83 %). For Scopus the figures are 549 correct cases out of 722 (76 %). It is important to note that no estimate of the Scopus population values can be derived from this because only WoS was sampled and the two data sources differ in the inclusion of content types. For the WoS sample we note that according to the data source's DT assignments, there were 194 letters, while we found only 111 plus one article/letter hybrid. WoS data indicates 204 articles in the sample while we found 295 plus 13 hybrids.

Case reports

As noted before, many medical case reports appeared in the letters section of issues or were addressed directly to the editors. This might lead to them being incorrectly assigned the DT letter. In the documentations of both Web

Table 2.3: Independent DT assignment vs. that of Scopus

independent assignment	Scopus				Total
	Article	Letter	Other	Review	
Article	218	59	3	10	290
Letter	4	99	7	0	110
Other	27	16	88	18	149
Review	23	0	2	135	160
Article/Letter Hybrid	0	1	0	0	1
Article/Other Hybrid	4	2	0	2	8
Article/Review Hybrid	3	0	0	1	4
Total	279	177	100	166	722

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

Table 2.4: DT assigned to case reports in WoS and Scopus

independently assigned DT	DT in data source	frequency WoS	frequency Scopus
Article	Article	3	11
Article	Letter	33	33
Article	Other	6	1
Article	Review	4	1
Article/Other Hybrid	Article	0	1
Article/Other Hybrid	Other	1	0

Note: Agreement between independently assigned document type and database assigned document type is indicated in bold

of Science and Scopus case reports are explicitly stated to be assigned the DT article. In the WoS sample, there were 47 case reports. The majority of them were assigned incorrectly: 33 letters, 4 reviews, 7 other; 3 were correctly coded as articles. Hence this is a major source of error but does not account for all the inaccuracies in records coded articles by WoS. All 47 case reports are also contained in Scopus, where the error is not as large: 12 records are correctly identified as articles while 33 were coded as letters. One case report fell under the DT review and one under the DT other. The findings are summarized in Table 2.4.

Comparing WoS and Scopus results directly

A cross-tabulation of the DT assignments of WoS and Scopus for the same pub-

Table 2.5: Comparison of DT assigned by WoS and Scopus to the same articles (N=722)

according to WoS	according to Scopus			
	Article	Letter	Other	Review
Article	191	1	1	9
Letter	11	174	5	0
Other	38	2	93	19
Review	39	0	1	138

Table 2.6: Comparison of DT assigned independently and by WoS for publications not included in Scopus

independent assignment	according to WoS			
	Article	Letter	Other	Review
Article	2	1	1	0
Letter	0	1	0	0
Other	0	1	56	2
Review	0	0	0	5

lications is given in Table 2.5. The two databases give the same DT, according to our four categories, in 83 % of the cases in this sample (N=722). Notable is the discrepancy in what Scopus considers *articles*.

It could be possible that the differences in coverage of the data sources influence the observed disagreements. To show that this is unlikely, Table 2.6 shows the independently assigned DT and the DT in WoS for publications found only in WoS but not in Scopus. No systematic pattern is apparent. The high number of records of type *other* not found in Scopus is consistent with Scopus’ policy of not including meeting abstracts, which WoS does cover.

Statistical analysis

The statistical inference is restricted to WoS as a data source, as only WoS was sampled, while Scopus data was subsequently linked to these sampled records. The coverage of WoS and Scopus differ, as does their DT inclusion policy. Hence, no reasonable statements of DT assignment accuracy for Scopus can be made from this data.

Table 2.7 contains the results for the estimation of Precision across DT classes

Table 2.7: Population estimates for Precision in DT assignment in WoS (n=791)

DT category	Web of Science
	Precision (95 % confidence interval)
Article	0.97 (0.93, 0.98)
Letter	0.55 (0.48, 0.62)
Other	0.91 (0.87, 0.95)
Review	0.87 (0.81, 0.91)
total	0.94 (0.92, 0.96)

Table 2.8: Population estimates of Recall in DT assignment in WoS (n=778)

DT category	Web of Science
	Recall (95 % confidence interval)
Article	0.96 (0.95, 0.98)
Letter	0.58 (0.36, 0.79)
Other	0.97 (0.95, 1.00)
Review	0.57 (0.39, 0.75)
total	0.94 (0.92, 0.96)

for WoS. The population estimate of the WoS overall proportion of correctly assigned DT is 0.94; 95 % confidence interval: (0.92, 0.96). The results for the estimation of DT class and overall Recall values are given in Table 2.8.

To summarize the results, Precision is particularly low in the DT *letter*. In terms of Recall, the assignment quality for the DT *article* and the synthetic DT of *other* are high and for *letter* and *review* the values are low. A considerable part of the inaccuracies in the DT *letter* results from incorrectly assigned case reports. Even with manual coding effort there remains a small set of publications which cannot easily be assigned to exactly one category.

Differences in quantitative publication characteristics between correctly and incorrectly DT coded publications

Because of their different communicative purpose and requirements for authoring, publications of different document types have specific distributions of various characteristics. For example, articles are usually longer than letters and reviews usually have more references than articles. For the WoS data collected in this study, the mean of the distributions of the number of pages, number

Table 2.9: Quantitative characteristics of different DT correctly and incorrectly assigned (WoS data)

DT in WoS	correctly assigned	N	mean pages	mean authors	mean references
Article	no	7	10.9	2.4	49.7
	yes	197	8.9	4.5	31.9
Letter	no	87	2.7	4.3	7.6
	yes	106	1.7	2.2	5.8
Other	no	18	3.8	2.9	15.7
	yes	191	2.2	5.9	5.1
Review	no	24	8.1	3.6	47.1
	yes	161	13.1	3.1	99.7

of authors and number of references are given in Table 2.9 for correctly and incorrectly coded DT assignments, respectively. The data confirm the above assumptions: the averages differ notable for correctly and incorrectly assigned records. This suggests the possibility of using these and possibly additional data to find publications which are likely coded incorrectly, namely those that deviate in such characteristics from the expectations of the typical publications of that DT. While we did not attempt to exploit this possibility in this study, this observation suggests the usefulness of a semi-automatic method to reduce the needed manual checking in situations where the DT data correctness is particularly important.

Influence on citation indicators

It was found that a certain share of DT assignments is inaccurate. Insofar as the uncorrected DT data are used to define reference sets to compute normalized citations scores (NCS), the data inaccuracies will influence citation impact studies. Impact distortions occur because the observed citation count is divided by the wrong DT-based reference group’s average citation score (expected value). Large-scale assessments that use NCS with normalization by DT include the Times Higher Education World University Ranking¹⁵ and the National Science Foundation’s Science and Engineering Indicators 2016¹⁶.

¹⁵Documented for the 2015/2016 edition <https://www.timeshighereducation.com/news/ranking-methodology-2015> but not the 2016/2017 edition <<https://www.timeshighereducation.com/world-university-rankings/methodology-worlduniversity-rankings-2016-2017>>.

¹⁶<https://www.nsf.gov/statistics/2016/nsb20161/#/>

To study the extent of this effect, the normalized article citation scores for 3-year citation windows for publications that were assigned an incorrect DT in WoS were computed. We only included records with both a corrected DT and a WoS DT of article, review or letter that were published prior to 2012 (73 cases). The calculation follows the method described in Waltman et al. (2011). WoS Subject Categories were used for disciplinary aggregation of publications necessary for defining the reference set. The incorrect values are the NCS computed with the reference set citation rate averages¹⁷ for the WoS-assigned DT while the corrected NCS are computed using the reference set average citation rate of the actual DT of the publication, as determined by this study, see Fig. 2.1.

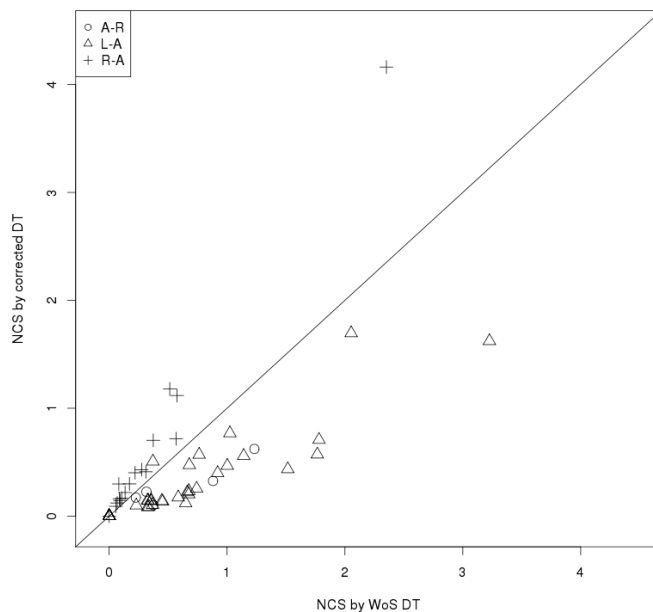


Figure 2.1: Scatterplot of WoS and corrected normalized citation scores for incorrectly coded publications ($N=73$). *Circle*: corrected from article to review (5 obs.); *triangle*: corrected from letter to article (51 obs.); *cross*: corrected from review to article (17 obs.)

The figure shows that these publications' NCS are over- or underestimated, de-

¹⁷It should be kept in mind that all the expected values are themselves affected by DT inaccuracies. No corrected values for these reference scores are available, so this issue must be put aside for this study although it is relevant in general.

pending on the specific pair of incorrect and correct DT. The diagonal indicates the line of no error. Publications with correctly computed values would lie exactly on the diagonal. Consider for example the publications whose DT was corrected from letter to article (triangle symbol). Most of them fall below the diagonal, indicating that their citation impact was overestimated because they were divided by expected values of letter publications even though they should have been divided by the expected value of articles. One relatively isolated data point is the cross symbol for an article mistakenly assigned the DT of review in WoS at the coordinates (2.4, 4.2). This publication, the paper Chaiworapongsa et al. (2008) is clearly an article, as the full text reports a full medical study and is labeled on the abstract page as “Research article”¹⁸ and on the issue contents web page it is listed in the section “Original Article”. It was cited 29 times in three years, while articles in its subject class were cited 7.0 times and reviews 12.3 times on average.

The error distribution of the NCS in terms of corrected scores minus WoS scores is shown in Fig. 2.2. 24 of the 73 publications have zero citations and are consequently not affected by distortion. It should be kept in mind that this histogram is not of the NCS error distribution of the entire sample but just the subset of incorrectly assigned DT.

These results confirm that within this specially selected subset of inaccurate DT data normalized citation scores of individual publications are strongly affected by these inaccuracies. Unless the citation count is zero the change in NCS is usually substantial and increases as the absolute citation count gets higher. The systematic influence of the kind of incorrectness, that is, which specific DT assignment error occurred, is also evident.

Limitations

This study has some methodological limitations that need to be considered when interpreting its results. The sample was restricted to publications that had a (unique) DOI assigned. In the earlier years of the sampling frame, this is not the greater part of publications. If the accuracy of DT assignment in citation indexes differs for publications without DOI, the sample would be biased. The reference DT was assigned clerically by two raters. Such ratings are subjective to some degree. However, in the case of ambiguous DTs in the original source,

¹⁸<http://www.tandfonline.com/doi/abs/10.1080/14767050701832833> accessed Sept. 21, 2016.

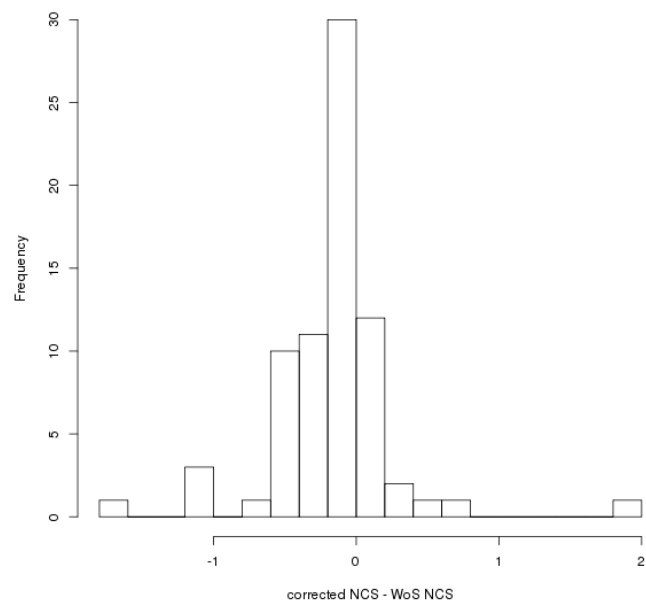


Figure 2.2: Error distribution NCS

some leniency was allowed for in the comparison. Because the sample is drawn from WoS data and afterwards matched with Scopus data where possible, the sample is not representative for Scopus and no estimates for the DT assignment accuracy of Scopus are possible.

2.4 Conclusion

In the present study, a stratified random sample of Web of Science-indexed publications was checked for accuracy of the document type as recorded in the data source in comparison to independently assigned document type data based on the data source's DT definitions. Document type assignments were found to be correct in Web of Science in 94 % of cases, when using the four DT categories of this study (article, letter, review, other). Significant proportions of publications may be missed when making selections of records from these data sources constrained to particular document types. In particular, records assigned the types letter and review are affected by inaccuracies. For identical publications indexed in both data sources, the DT assigned by WoS and Scopus are often conflicting. Because of different citation distributions of publications of different document types, the normalized citations scores of publications with incorrect DTs become systematically distorted.

In analyses of modest numbers of publications, publications should not be excluded from citation analysis a priori because of their apparent document type. DT assignments should be verified if feasible, especially for indicator supported research evaluation. When this is not viable, the results of studies can be accompanied by sensitivity analyses that take the information about inaccuracies of DT into account, using the error rate found empirically in this paper or determined independently for the publication set at hand. In interpreting the results of bibliometric studies, it should be kept in mind that when publications of some document types are excluded, this might lead to the unintentional exclusion of publications of interest. When citation normalization is based on reference classes of specific document types this can lead to inaccurate normalized citation scores. This problem is present in general in any bibliometric study that is based on calculations taking document type data into account implicitly or explicitly.

Acknowledgements

This study was supported by the German Federal Ministry of Education and Research (BMBF) grant 01PQ13001, project “Kompetenzzentrum Bibliometrie”. I want to thank Anastasiia Tcypina for help with data collection and Nees Jan van Eck for discussion of the manuscript.

Appendix

A: Matching of WoS and Scopus records

An initial basic matching table was created by FIZ Karlsruhe for general use in projects by the database users based on combinations of exactly corresponding values in 10 metadata fields. The fields were assigned weights for their discriminative power. For example, the publication year field has a low discriminative power while that of article title and DOI is very high. Field values were normalized for differences in character sets, capitalization, special characters and some structural aspects (e.g. removing separating dashes in ISSN) to make them more uniform between the two data sources. All journal records are then mutually compared across all fields and a score is calculated from the number and weights of exactly matching fields. Only pairs with a predetermined threshold score are kept. Through this procedure, a pair of records may be included multiple times in the resulting table because all reasonable field combinations are used. Furthermore, a record from one data source can occur as a plausible match pair with multiple records from the other data source. The matching quality of this method was assessed before the study was conducted and found to be satisfactory. For this, a random sample of 2450 matching pairs was selected from the table and manually checked for equivalence using full bibliographical data from both data sources. In 9 cases the match was incorrect, in 15 cases the decision was unclear and in 2426 cases the match was found to be correct, the percentage of correct matches being greater than 99 %.

This basic data was slightly modified for matching the sampled WoS records to those in Scopus uniquely. For WoS records being assigned only one possible Scopus record, the entry rows were copied directly to the final matching table. For those WoS records with more than one plausible match, the single row with the highest matching score was copied. This concerns 2.6 % of the entries in the base table. Using this look-up table, of the 793 records that were initially assessed, it was possible to identify matches in Scopus in 711 cases. This was

extended by matching only on the DOI for remaining records and manually assessing the matches. Finally, for all unmatched records that were left, the title was searched in the Scopus online platform and the results checked for a correct match. These two steps produced another eleven verified matches. The remaining unmatched records are mostly meeting abstracts records, which are deliberately not included in Scopus. According to the DT assigned for this study, these unmatched records comprise, after exclusion of publications which could not be found, four articles, one letter, five reviews and 59 others. Within these, there were five misclassifications by WoS.

B: Distribution of publication years in the WoS sample

Table 2.10: Distribution of publication years in the WoS sample

publication year	cases
2002	18
2003	31
2004	35
2005	42
2006	55
2007	60
2008	63
2009	79
2010	99
2011	90
2012	108
2013	93
2014	18

C: Equality of population estimates of Precision and Recall

In this study, the overall population estimates of Precision and Recall, as opposed to the estimates for particular DT, for each data source are equal, unlike in typical information retrieval evaluation. This is so because each case is “relevant” for one of the four DT categories. To illustrate, a simplified example is worked through.

Consider a dataset with only two document types, A and B, for which we have the data source’s assignments and the true DTs, as assigned manually. There

is no sampling and stratification. Data are cross-tabulated data source and independently assigned DT counts like this:

Table 2.11:

independently assigned DT	data source DT	
	A	B
A	20	6
B	5	40

There are 26 true A records, they have a population proportion of 0.37 while the 45 B records have a proportion of 0.63. Calculating first the DT Recall (R) and Precision (P) values for A and B gives:

$$R_A = \frac{20}{20+6} = 0.77 \quad R_B = \frac{40}{40+5} = 0.89$$

$$P_A = \frac{20}{20+5} = 0.80 \quad P_B = \frac{40}{40+6} = 0.87$$

Then the overall Recall and Precision are:

$$R = (R_A \times w_A) + (R_B \times w_B) = (0.77 \times 0.37) + (0.89 \times 0.63) = 0.85$$

$$P = (P_A \times w_A) + (P_B \times w_B) = (0.80 \times 0.37) + (0.87 \times 0.63) = 0.85$$