Scientometrics (2022) 127:4419–4438
https://doi.org/10.1007/s11192-022-04446-y

Relevance of document types in the scores' calculation of a specific field-normalized indicator: Are the scores strongly dependent on or nearly independent of the document type handling?

Robin Haunschild[1] · Lutz Bornmann[1,2]

Received: 17 May 2021 / Accepted: 20 June 2022 / Published online: 15 July 2022

Abstract
Although it is bibliometric standard to employ field normalization, the detailed procedure of field normalization is not standardized regarding the handling of the document types. All publications without filtering the document type can be used or only selected document types. Furthermore, the field-normalization procedure can be carried out with regard to the document type of publications or without. We studied if the field-normalized scores strongly depend on the choice of different document type handlings. In doing so, we used the publications from the Web of Science between 2000 and 2017 and compared different field-normalized scores. We compared the results on the individual publication level, the country level, and the institutional level. We found rather high correlations between the different scores but the concordance values provide a more differentiated conclusion: Rather different scores are produced on the individual publication level. As our results on the aggregated levels are not supported by our results on the level of individual publications, any comparison of normalized scores that result from different procedures should only be performed with caution.

Keywords Scientometrics · Bibliometrics · Document type · Field normalization

Introduction

According to one of a total of ten principles in the Leiden manifesto for the professional application of bibliometrics in research evaluation, field-normalized scores should be used instead of simple citation counts (Hicks et al., 2015). The citation impact of individual

✉ Robin Haunschild
R.Haunschild@fkf.mpg.de

Lutz Bornmann
bornmann@gv.mpg.de

[1] Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart, Germany

[2] Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

publications from the same year and scientific field is reflected by such scores. Whereas this basic principle of field normalization has been emerged as standard procedure in bibliometrics, specific elements of the procedure are unclear or are applied differently in bibliometrics (e.g., the use of the categorization system to define fields). One of these elements is how the document type should be handled during the normalization procedure. The Leiden Ranking (Waltman et al., 2012) and the SCImago Institutions Ranking (SIR)—two popular institutional rankings—include different types of publications: The Leiden Ranking currently (CWTS, 2022) includes only the document types 'Article' and 'Review' whereas the SIR additionally considers the document types 'Conference Paper' and 'Short Survey'. Both rankings consider the document type when calculating field-normalized scores.

InCites—a citation-based research analytics tool evaluating institutional productivity—includes all document types and normalizes with respect to them separately (Clarivate Analytics, 2021). A similar procedure is applied in SciVal (Elsevier, 2019)—a tool that is very similar to InCites. It seems to be a given for major rankings and tools that documents of different types are treated separately in normalization, although there is to the best of our knowledge no study yet that investigates this effect. Some databases, for example Microsoft Academic Graph (Scheidsteger et al., 2018; Sinha et al., 2015; Wang et al., 2020), its successor OpenAlex (OurResearch, 2021; Priem et al., 2022), or Dimensions (Herzog et al., 2020), do not distinguish between 'Article', 'Review', 'Letter', 'Note', 'Editorial material', etc. (which makes the consideration impossible). Another practical necessity might prevent the consideration of document types during normalization procedures: If fields with rather few publications are separated not only by publication year but also by document type, this might lead to too small reference sets for normalization procedures (leading to unreliable results). Furthermore, the assignment of document types is inconsistent between different databases (i.e., an 'Article' in Web of Science (WoS) might be a 'Review' in Scopus).

Many studies in bibliometrics have shown that publications of different document types not only gather a different average number of citations but also gather their citations at different speeds (see, e.g., Wang, 2013). Based on previous research on the Journal Impact Factor (JIF, provided by Clarivate Analytics), we hypothesized that the handling of the document type in calculating field-normalized scores will lead to different results in research evaluation. In the calculation of the JIF, document types contribute differently to the citations a journal receives (Clarivate Analytics, 2021; Van Leeuwen et al., 1998). Glänzel & Moed (2002) list five factors that may influence the JIF. One of these factors is the document type of a publication, i.e., the distribution of publications across document types in a journal. In previous research on field-normalized indicators, Nederhof & Visser (2004) analyzed in a case study the change in average field-normalized citation scores (significant increases of the indicator between 1989–1993 and 1994–1998) of two Dutch universities. They found a changed document type handling in the two time periods as one reason for the significant increase of indicator values.

In this study, for verifying our hypothesis and possibly generalizing the results of the case study by Nederhof & Visser (2004), we compare field-normalized citation scores, which have been calculated based on three different ways of handling document types in

Table 1 Differences and commonalities of the three datasets used in this study

	dt0	dt1	dt2
Number of papers	34,929,708	26,766,770	26,766,770
Number of document types	35	4	4
Document type included in normalization procedure?	No	Yes	No

the normalization procedure. We are interested whether they lead to the same, similar, or different scores for the same papers—if everything else (i.e., the formula for calculating the scores and the field classification) remains unchanged. This is an important question for the use and interpretation of field-normalized scores in research evaluation: the scores are calculated in different ways in the concrete research evaluation practice (see above). If different document type handling leads to different scores, field-normalized scores from different sources should only be compared with caution—although the field-normalized indicator (and used field-categorization scheme) is the same.

Methods

Data set

We used a custom database developed and maintained by the Competence Center for Bibliometrics (CCB, see https://www.bibliometrie.info/) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), and Arts and Humanities Citation Index (AHCI) of the WoS (Birkle et al., 2020) provided by Clarivate Analytics. All publications ($n = 34{,}929{,}708$) with publication years between 2000 and 2017 were used for the analysis. In total, 35 different document types are included in the data set. We used the affiliation assignment of the CCB for comparing the results of field-normalized scores with respect to countries and universities.

Handling of document types in the normalization procedure

In this study, we compare three approaches of considering the document type in calculating field-normalized scores:

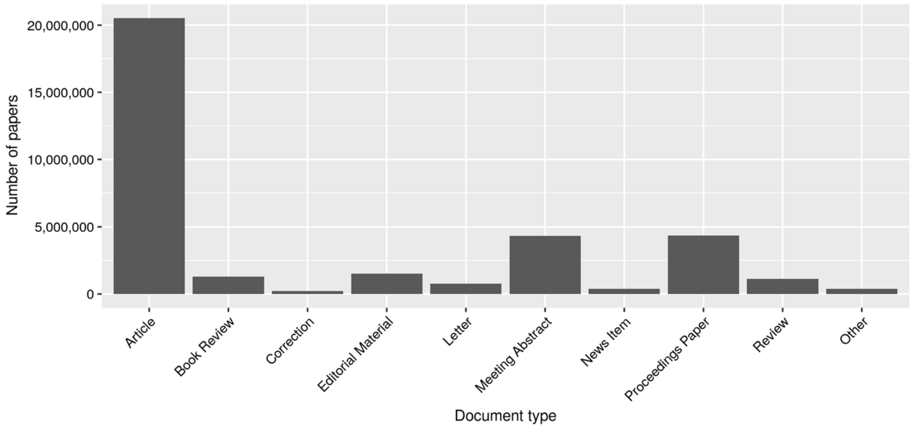

Fig. 1 Most frequently occurring document types in data set dt0

Table 2 Descriptive statistics of the NCS values on the basis of individual publications

	Average	Median	Standard deviation	Minimum	Maximum
dt0	1.25	0.50	3.89	0.00	3,193.09
dt1	0.98	0.44	3.00	0.00	3,179.92
dt2	0.99	0.42	3.12	0.00	2,412.82

Table 3 Descriptive statistics of the MNCS values on the basis of the 84 most productive countries

	Average	Median	Standard deviation	Minimum	Maximum
dt0	1.32	1.28	0.42	0.62	2.40
dt1	1.03	1.00	0.26	0.56	1.71
dt2	1.06	1.04	0.31	0.55	1.83

Table 4 Descriptive statistics of the MNCS values on the basis of the 81 most productive German universities

	Average	Median	Standard deviation	Minimum	Maximum
dt0	1.67	1.68	0.44	0.91	3.56
dt1	1.26	1.27	0.21	0.79	1.87
dt2	1.31	1.30	0.25	0.85	2.33

(1) The data set **dt0** uses all publications in our data set in the normalization procedure. The citation counts are normalized with respect to their WoS subject categories and publication years. Document types are not considered in the normalization of dt0. This data set represents the procedure in a database where no fine-grained distinction is made between journal document types (e.g., 'Article', 'Review', 'Editorial material', and 'Note').

(2) We restrict the publication set of dt0 to the document types 'Article', 'Review', 'Letter', and 'Proceedings Paper'. After this restriction, 26,766,770 publications are remaining. The citation counts of these publications are normalized with respect to their document types, WoS subject categories, and publication years. The resulting data set is referred to as **dt1**. A similar approach is used in SIR.

(3) We normalize the citation counts of the 26,766,770 publications only with respect to their WoS subject categories and publication years without consideration of their document types. The resulting data set is referred to as **dt2**. This means that dt1 and dt2 are subsets of dt0, and handling of dt0 and dt2 is the same whereas handling of dt1 is different. Table 1 shows the commonalities and differences of the three datasets used in this study.

Overall, the data set dt0 contains 35 different document types. Figure 1 shows the most frequently occurring document types in the data set. The document types with less than 100,000 papers each were aggregated in the category ‘Other’.

Statistics

Waltman (2016) provided an overview of several approaches of field-normalization procedures. For this study, we chose the normalized citation score (NCS) for the comparison, because it is still one of the most frequently used approaches. In principle, any other available field-normalized indicator could have been used. For the calculation of the NCS, reference sets are constructed, and the average citation counts within these reference sets is calculated. Each paper’s citation count is divided by the average citation count of its appropriate reference set. For the generation of the reference sets, papers are aggregated when they belong to the same publication year, scientific field (here: WoS subject category), and – depending on the approach (see above)—the document type as the focal paper. If, for example, the paper has 45 citations and the average for the appropriate reference set is 15, the NCS of the paper is 45/15 = 3. This means the focal paper received significantly more citations than an average paper in the same field and publication year (and with the same document type). The NCS is formally defined as

$$NCS = \frac{c_i}{e_i}$$

where c_i is the citation count of a focal paper and e_i is the corresponding expected citation rate in the field (Lundberg, 2007; Rehn et al., 2007; Waltman et al., 2011). The number of citations that a paper receives depends on the time since publication. Thus, the reference sets are also grouped by publication years. Using the different approaches of normalization with regard to the document type handling, we calculated three NCS values for every publication: NCS(dt0) (based on data set dt0), NCS(dt1) (based on data set dt1), and NCS(dt2) (based on data set dt2). In the case of multiple WoS subject categories for a single publication, NCS values are calculated for each category and the arithmetic average is calculated across these scores.

We perform three different analyzes of the data set: (i) on the individual paper level basis, (ii) on the basis of countries, and (iii) on the basis of German universities. We focus on German universities in this study, since we have disambiguated data only for German institutions in our in-house database. 26,058,126 distinct publications are assigned to at least one country, and 1,407,025 distinct publications are assigned to at least one German university. We use the full counting method for assigning publications to countries and German universities. Proper fractional counting might be possible on the country level but not on the institutional level, because we do not have disambiguated institutional data in our database for non-German institutions. Due to the full counting procedure, 33,111,283 publications for countries and 1,778,689 publications for German universities are included in the analysis. We use the arithmetic average for calculating the overall field-normalized impact of publications of a specific

Fig. 2 Scatter plot of NCS values on a log–log scale of the pair dt0 and dt1 in panel (**A**), dt1 and dt2 in panel (**B**), and dt0 and dt2 in panel (**C**). Individual data points are plotted with transparency so that single data points appear light-grey whereas multiple points on top of each other are darker. (Colour figure online) ▶

country or German university. This procedure leads to three impact scores: MNCS(dt0), MNCS(dt1), and MNCS(dt2), for each country and German university.

In order to increase statistical reliability of the results on the aggregated level, we restrict the country analysis to the countries with at least 10,000 publications of the document types 'Article', 'Review', 'Letter', or 'Proceedings Paper' in the analyzed 18 years. The institutional analysis is based on German universities with at least 1,000 publications of the document types 'Article', 'Review', 'Letter', or 'Proceedings Paper' in the analyzed 18 years. This restriction reduces the number of countries in the CCB database from 219 to 84 and the number of German universities from 122 to 81. Despite these significant reductions, 99.3% of the publications are included in the set of the most productive 84 countries, and 99.5% of the publications are included in the set of the most productive 81 German universities. Table 2 shows descriptive statistics of the NCS values on the basis of individual publications. Tables 3 and 4 show the same statistical measures for the aggregated levels on the basis of the most productive countries and German universities. The comparison of the three tables reveals that the standard deviation decreases substantially upon aggregation.

In this study, we compare NCS(dt0), NCS(dt1), and NCS(dt2) on the individual paper level and MNCS(dt0), MNCS(dt1), and MNCS(dt2) of the countries and German universities using (i) Spearman rank correlation and (ii) concordance coefficients following Lin (1989, 2000). Whereas correlations measure the relationship between two variables, concordances measure their agreement: Do we receive the same value despite different handlings of document types? The NCS values for each paper were calculated using SQL in an Oracle database. All other statistical analyses were performed using R (R Core Team, 2019) with the R packages 'tidyverse' (Wickham, 2017) and 'DescTools' (Signorell et al., 2020). Figures were produced using the R package 'ggplot2' (Wickham, 2016).

Results

Analyses on the level of individual papers

Figure 2 shows a scatter plot of NCS values of the pair dt0 and dt1 in panel A, dt1 and dt2 in panel B, and dt0 and dt2 in panel C. Individual data points are plotted with transparency so that single data points appear light-grey whereas multiple points on top of each other are darker. The three scatter plots in Fig. 2 clearly show that the NCS values from dt0 and dt2 are more similar to each other than the NCS values from dt0 and dt1 or dt1 and dt2. This tendency is also visible from the Spearman correlation coefficients in Table 5.

Correlation is a rather weak measure of association. Therefore, we present the concordance coefficients according to Lin (1989, 2000) in the lower triangle and their 95% confidence intervals in the upper triangle of Table 6. We follow the guidelines by McBride

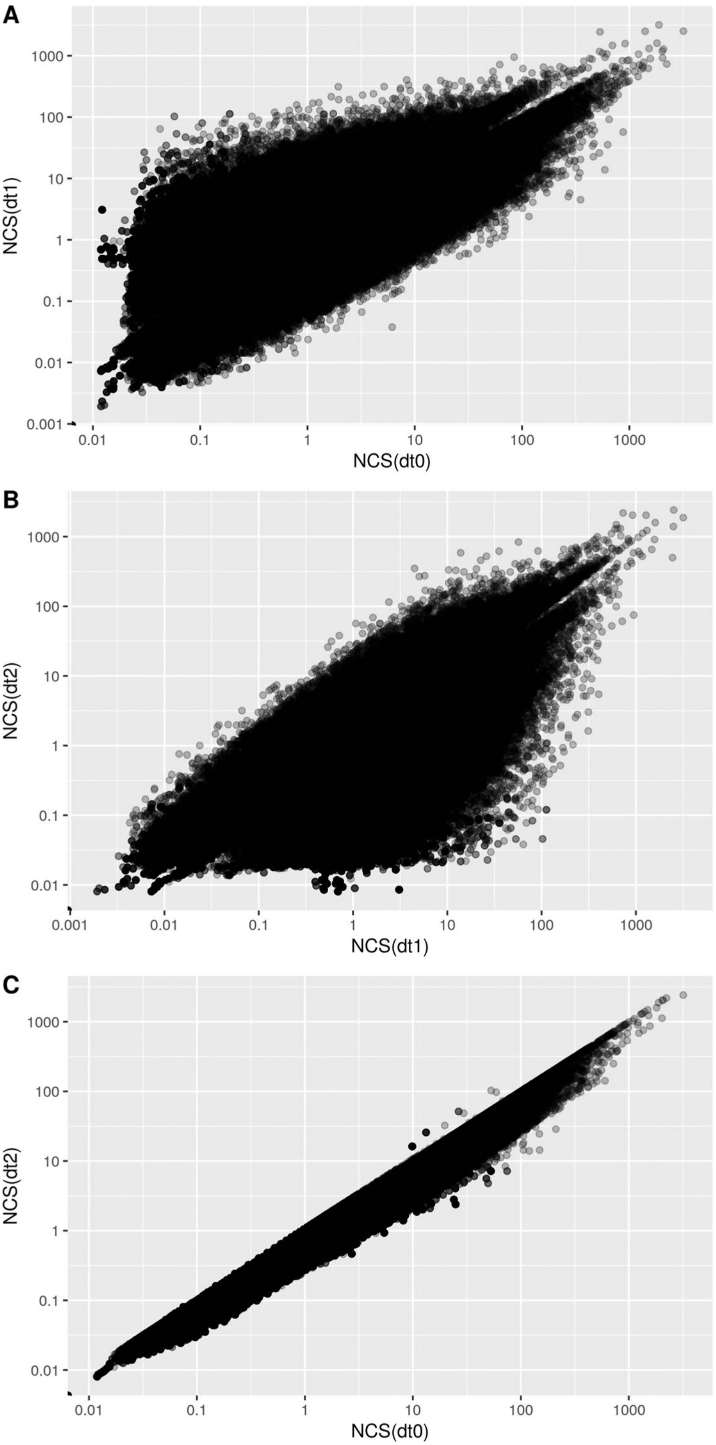
A
NCS(dt1)
1000
100
10
1
0.1
0.01
0.001
0.01
0.1
1
10
100
1000
NCS(dt0)
B
NCS(dt2)
1000
100
10
1
0.1
0.01
0.001
0.01
0.1
1
10
100
1000
NCS(dt1)
C
NCS(dt2)
1000
100
10
1
0.1
0.01
0.01
0.1
1
10
100
1000
NCS(dt0)

Table 5 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2

	NCS(dt0)	NCS(dt1)	NCS(dt2)
NCS(dt0)	1.00	0.92	0.99
NCS(dt1)	0.92	1.00	0.93
NCS(dt2)	0.99	0.93	1.00

Table 6 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2. The upper triangle shows the 95% confidence intervals

	NCS(dt0)	NCS(dt1)	NCS(dt2)
NCS(dt0)	1.00	[0.742, 0.742]	[0.935, 0.935]
NCS(dt1)	0.74	1.00	[0.784, 0.785]
NCS(dt2)	0.94	0.78	1.00

Table 7 Spearman rank correlation coefficients of the MNCS values for the 84 most productive countries

	MNCS(dt0)	MNCS(dt1)	MNCS(dt2)
MNCS(dt0)	1.00	0.95	0.99
MNCS(dt1)	0.95	1.00	0.96
MNCS(dt2)	0.99	0.96	1.00

(2005) for interpreting the strength of the concordance. Table 6 indicates different types of concordance strengths: (i) almost perfect or substantial in black, (ii) moderate in blue, and (iii) poor in red. The concordance coefficients substantiate the impression from the correlation results: The NCS values from dt0 and dt1 as well as dt1 and dt2 show a poor agreement whereas the NCS values from dt0 and dt2 show a moderate agreement.

Appendix A includes tables that show the correlation and concordance for each document type separately. There are only minor changes in the Spearman rank correlation coefficients. The concordance coefficients are rather similar, too, except for the data set combinations dt0/dt1 and dt1/dt2 where the concordance drops significantly for the document types ‘Review’ and ‘Letter’ in comparison with the values in Table 6.

Analyses with respect to countries

Table 7 shows Spearman rank correlation coefficients of the MNCS values for these 84 countries in our data set. The correlation coefficients show a very high rank correlation between all three different approaches for normalizing citation counts. The close relationship is also reflected in Fig. 3 that shows a scatter plot of MNCS values for the most productive 84 countries. The size of the dots in Fig. 3 indicates the number of papers. Appendix B also shows the Spearman rank correlation coefficients for the document types ‘Article’, ‘Proceedings Paper’, and ‘Review’ separately. Some of the 84 countries published too few ‘Letters’ so that a separate analysis for ‘Letters’ would not be useful. The

Spearman rank correlation coefficients change only slightly when focussing on individual document types.

Table 8 shows the concordance coefficients according to Lin (1989, 2000) in the lower triangle and their 95% confidence intervals in the upper triangle. Table 8 indicates different types of concordance strengths: (i) almost perfect or substantial in black, (ii) moderate in blue, and (iii) poor in red. The Table reveals that MNCS(dt1) and MNCS(dt2) show a substantial agreement, whereas MNCS(dt0) and MNCS(dt2) as well as MNCS(dt0) and MNCS(dt1) exhibit a rather poor agreement. This result is in contrast to the results on the level of individual publications: Here, the agreement between MNCS(dt1) and MNCS(dt2) is poor and the agreement between MNCS(dt0) and MNCS(dt2) is moderate.

Appendix B shows separate concordance coefficients for the document types ‘Article’, ‘Proceedings Paper’, and ‘Review’. Most concordance coefficients are rather similar to the ones in Table 8. However, the condorcance coefficients are significantly lower for the document types ‘Review’ and ‘Proceedings Paper’ in the case of the dataset combinations dt0/dt1 and dt1/dt2. Figure 4 shows the rank differences (rank using all papers *minus* rank using a specific document type only) for the 84 most productive countries as a box plot. Outliers are marked separately. The figure shows that most rank differences are rather small for the docuemt type ‘Article’ but larger for the docuemt types ‘Proceedings Paper’ and ‘Review’. The outliers marked in the figure are: Armenia, Belarus, Bulgaria, Cyprus, Indonesia, Kazakhstan, Kenya, Latvia, Malaysia, Peru, Philippines, Sri Lanka, Tanzania, and United Republic of Uganda.

A different perspective on the comparison of the NCS values for the most productive countries in the analyzed time period of 18 years is provided by the dot plot in Fig. 5.

The size of the dots in Fig. 5 corresponds to the number of publications from each country, and the color of the dots corresponds to the handling of the document type in the normalization procedure. In most of the cases, MNCS(dt0) is greater than MNCS(dt1). However, for five out of the 84 most productive countries, MNCS(dt1) is greater than MNCS(dt0). These countries are: Indonesia, Latvia, Malaysia, Romania, and Slovakia. There is no country where MNCS(dt2) > MNCS(dt0). There are 64 out of the 84 most productive countries with MNCS(dt1) < MNCS(dt2).

Analyses with respect to German universities

In this section, we present our results with respect to German universities. Figure 6 shows a scatter plot of MNCS values for these most productive 81 German universities with an MNCS(dt0). The size of the dots in Fig. 6 indicates the number of papers as shown by the legend.

Table 9 shows Spearman rank correlation coefficients of the MNCS values for the most productive 81 German universities in our data set. The correlation coefficients show a lower rank correlation for German universities than for countries between all three different approaches for normalizing citation counts (see Tables 7 and 9).

Table 10 shows concordance coefficients according to Lin (1989, 2000) in the lower triangle and their 95% confidence intervals in the upper triangle. Table 10 indicates different types of concordance strengths: (i) almost perfect or substantial in black, (ii) moderate

Fig. 3 Scatter plot of MNCS values for the 84 most productive countries of the pair dt0 and dt1 in panel **A**, ▶ dt1 and dt2 in panel **B**, and dt0 and dt2 in panel **C**. Individual data points are plotted with transparency so that single data points appear light-grey whereas multiple points on top of each other are darker. (Colour figure online)

in blue, and (iii) poor in red. We observe a substantial agreement between MNCS(dt1) and MNCS(dt2), whereas we find a poor agreement between the other two pairs of MNCS values. This is not in agreement to the results on the level of individual publications: Here, the agreement between MNCS(dt1) and MNCS(dt2) is poor and the agreement between MNCS(dt0) and MNCS(dt2) is moderate.

The concordance coefficients for the German universities in Table 10 are lower than the ones for countries in Table 8. Overall, however, the concordance coefficients on the aggregated levels lead to similar conclusions that are somewhat different from the conclusions one can draw from the analysis on the basis of individual publications. Too few of the German universities have a significant number of publications of other document types than 'Article' indexed in WoS to warrant a separate analysis on the basis of different document types. For example, one of the 81 most productive German universities has only 16 indexed publications of the document type 'Proceedings Paper' and another one has only 13 publications of the document type 'Review'.

Figure 7 provides a different perspective on the comparison of the MNCS values of the 81 most productive German universities. The size of the dots in Fig. 7 corresponds to the number of publications from each university, and the color of the dots corresponds to the handling of the document type in the normalization procedure. In most of the cases, MNCS(dt0) is greater than MNCS(dt1). However, for five out of the most productive 81 German universities, MNCS(dt1) is greater than MNCS(dt0). Those German universities are: 'Universität der Bundeswehr München', 'Technische Universität Ilmenau', 'Technische Universität Bergakademie Freiberg', 'Technische Universität Chemnitz', and 'Universität Passau'. There is no German university where MNCS(dt2) > MNCS(dt0). There are 63 German universities with MNCS(dt1) < MNCS(dt2) and 18 German universities with MNCS(dt2) < MNCS(dt1).

Discussion and conclusions

According to Ioannidis et al. (2016) "the basic premise of normalization is that not all citations are equal. Therefore, normalization can be seen as a process of benchmarking". Usage of field-normalization procedures has become standard in bibliometrics, but various approaches of these procedures are in use. There are different choices available with regard to the specific formula of the field-normalization procedures (percentiles have been proposed as an alternative to scores based on average citations, Bornmann & Marx, 2015), with regard to the choice of field classification (Haunschild et al., 2018, 2022), or with regard to the way how document types are handled. In this study, we addressed the third aspect by comparing the scores that result from different normalization procedures, which

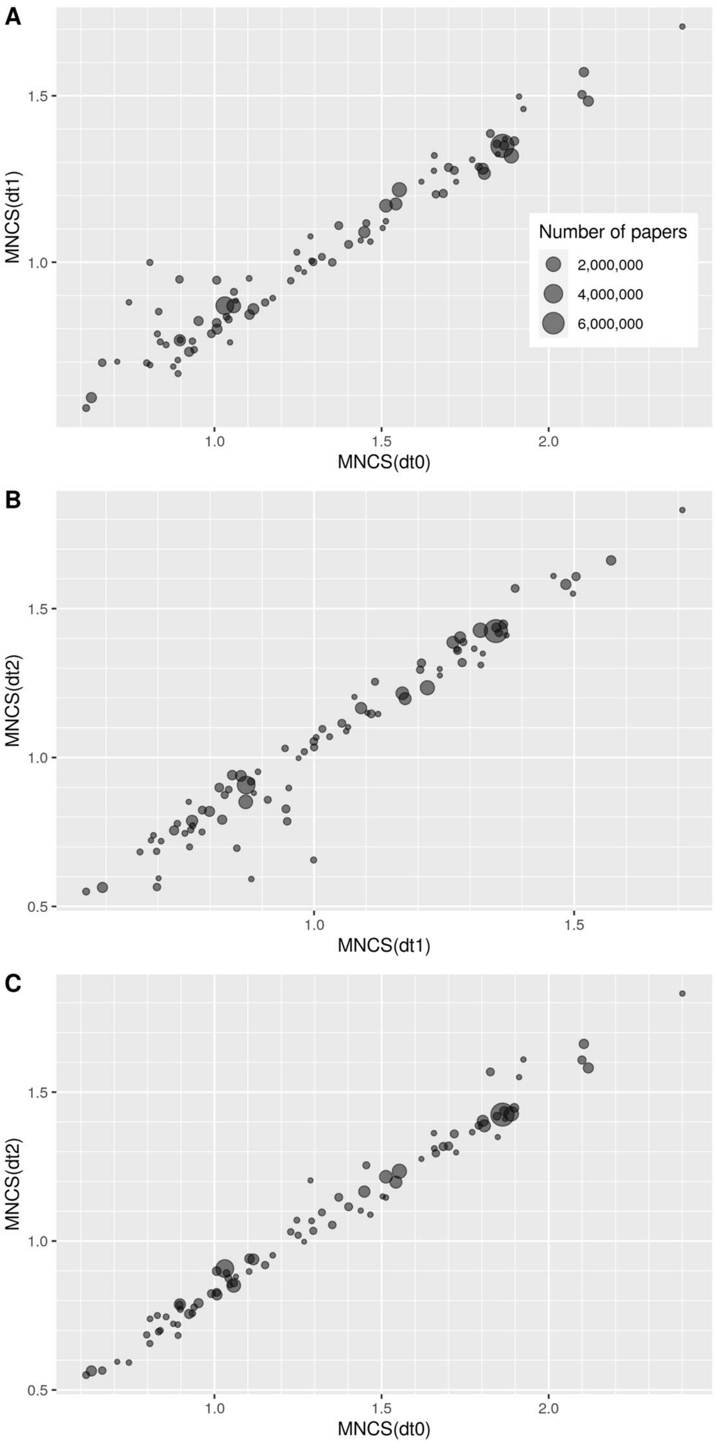
A
MNCS(dt1)
1.5
1.0
1.0
1.5
2.0
MNCS(dt0)
Number of papers
2,000,000
4,000,000
6,000,000
B
MNCS(dt2)
1.5
1.0
0.5
1.0
1.5
MNCS(dt1)
C
MNCS(dt2)
1.5
1.0
0.5
1.0
1.5
2.0
MNCS(dt0)

Table 8 Concordance coefficients of the MNCS values for the 84 most productive countries

	MNCS(dt0)	MNCS(dt1)	MNCS(dt2)
MNCS(dt0)	1.00	[0.57, 0.71]	[0.69, 0.81]
MNCS(dt1)	0.65	1.00	[0.93, 0.97]
MNCS(dt2)	0.76	0.95	1.00

The upper triangle shows the 95% confidence intervals

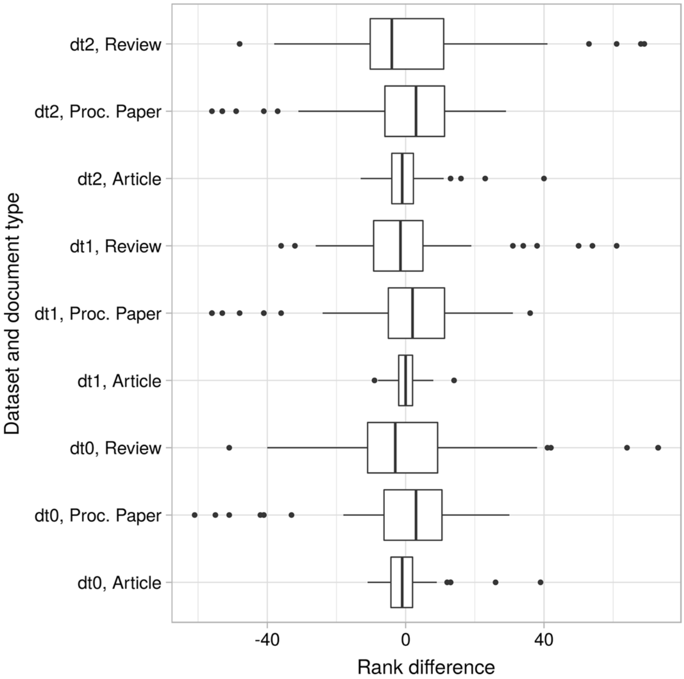

Fig. 4 Rank differences of the 84 most productive countries when different document types are focussed on for the three datasets. (Colour figure online)

have been performed based on three different approaches of handling the document types. At least two of these approaches are in use in popular university rankings.

In this study, we have analyzed field-normalized scores for individual publications, for countries and for German universities with three different ways of handling the document types during the normalization procedure: (i) normalizing all publications irrespective of document type together accounting only for field and age of publications, (ii) normalizing

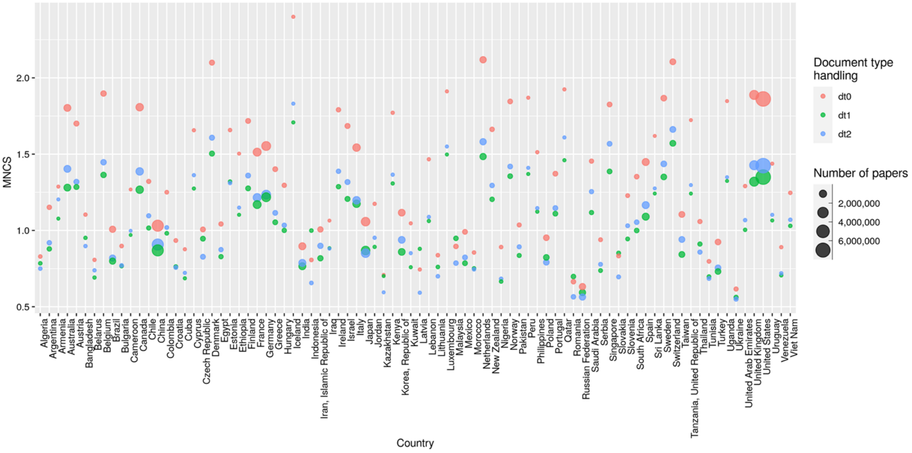

Fig. 5 Dot plot of MNCS values for the most productive countries ($n = 84$) using the three different normalization procedures with respect to the document type. (Colour figure online)

all publications separately with respect to document type, field, and age of publications, and (iii) filtering the publication set for four common document types (‘Article’, ‘Review’, ‘Proceedings Paper’, and ‘Letter’) and normalizing with respect to field and age of the publications irrespective of their document types. Our results showed that the different normalization procedures produce rather different scores on the level of individual publications.

Our results reveal (1) that normalization without any consideration of the document type of the publications (variant 1, i.e., dt0) yields scores that are in moderate agreement on the basis of individual publications with those scores obtained from filtering specific document types before the normalization procedure (variant 3, i.e., dt2). We found (2) that normalization without any consideration of the document type of the publications (variant 1, i.e., dt0) yields very different results on the aggregated levels compared to the variants where the document type is accounted for either by filtering (variant 3, i.e., dt2) or by separate normalization (variant 2, i.e., dt1). Although the results on the paper and aggregated levels do not agree, the results on both levels point to differences between NCS values calculated with different handlings of document types. The reduction in variability of the scores (see Tables 2, 3, and 4) upon aggregation might provide an explanation for the discrepancies between the results on the level of individual publications and on the aggregated levels.

Our results imply that normalized citation scores are not comparable when the normalization procedures are too different with respect to the document type handling, although the same normalization formula and field classification scheme are used. When normalized scores A are calculated with publications of too many too different document types without

Fig. 6 Scatter plot of MNCS values for the 81 most productive German universities of the pair dt0 and dt1 ▶ in panel (**A**), dt1 and dt2 in panel (**B**), and dt0 and dt2 in panel (**C**). Individual data points are plotted with transparency so that single data points appear light-grey whereas multiple points on top of each other are darker. (Colour figure online)

accounting for the document type in the normalization procedure, these scores should only be compared cautiously with normalized scores B that are calculated with some consideration of the document type during the normalization procedure, either be filtering for certain document types or formal normalization with respect to the document types. This is an important result for the use of field-normalized scores in the research evaluation practice. Normalized scores (on the single paper or aggregated level) are comparable only then when the document types have been handled equally or in a similar way.

Although one might expect that our results are transferable to normalization procedures that use a different normalization formula (e.g., percentiles) or a different field classification system, further research might check the transferability of our results in this respect. We analyzed only German universities and countries as aggregated units. Further research could analyze non-German institutional aggregations and check whether our results are transferable to them. Furthermore, our results are specific to the document type definition of the WoS. Different results might be obtained when other databases with different document type definitions are used. It should be checked in future studies, whether this is the case.

For calculating normalized scores, we recommend to perform normalizations with respect to document types besides field and year if practical problems do not prevent such a procedure. For example, some databases do not distinguish between journal document types (e.g., publications of the document types 'Article', 'Review', and 'Letter'). If too small reference sets occur due to separate normalization for each document type, it might be advisable to neglect less frequently occurring document types or group them together.

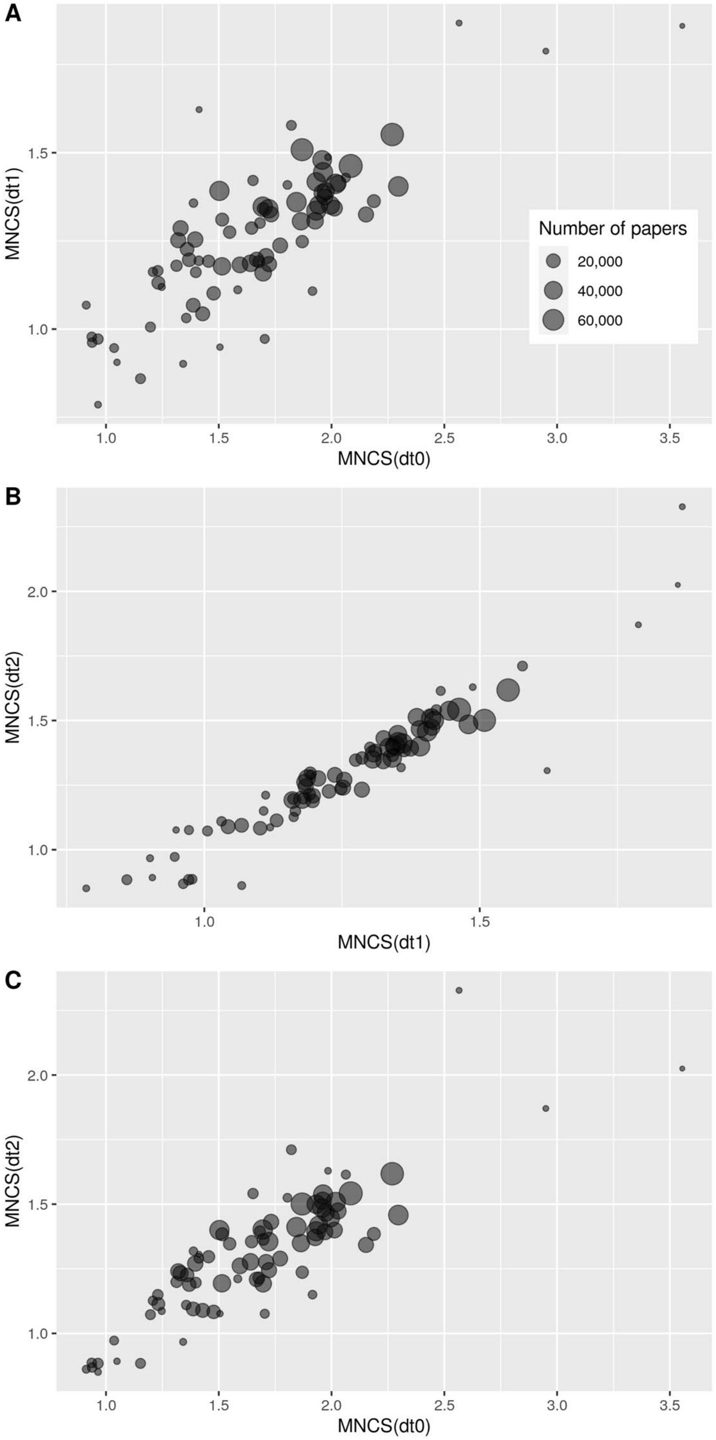
A
MNCS(dt1)
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
MNCS(dt0)
Number of papers
20,000
40,000
60,000
B
MNCS(dt2)
2.0
1.5
1.0
1.0
1.5
MNCS(dt1)
C
MNCS(dt2)
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
MNCS(dt0)

Table 9 Spearman rank correlation coefficients of the MNCS values for the 81 most productive German universities in our database

	MNCS(dt0)	MNCS(dt1)	MNCS(dt2)
MNCS(dt0)	1.00	0.78	0.84
MNCS(dt1)	0.78	1.00	0.96
MNCS(dt2)	0.84	0.96	1.00

Table 10 Concordance coefficients of the MNCS values for the 81 most productive German universities in our database. The upper triangle shows the 95% confidence intervals

	MNCS(dt0)	MNCS(dt1)	MNCS(dt2)
MNCS(dt0)	1.00	[0.279, 0.454]	[0.393, 0.577]
MNCS(dt1)	0.37	1.00	[0.874, 0.940]
MNCS(dt2)	0.49	0.91	1.00

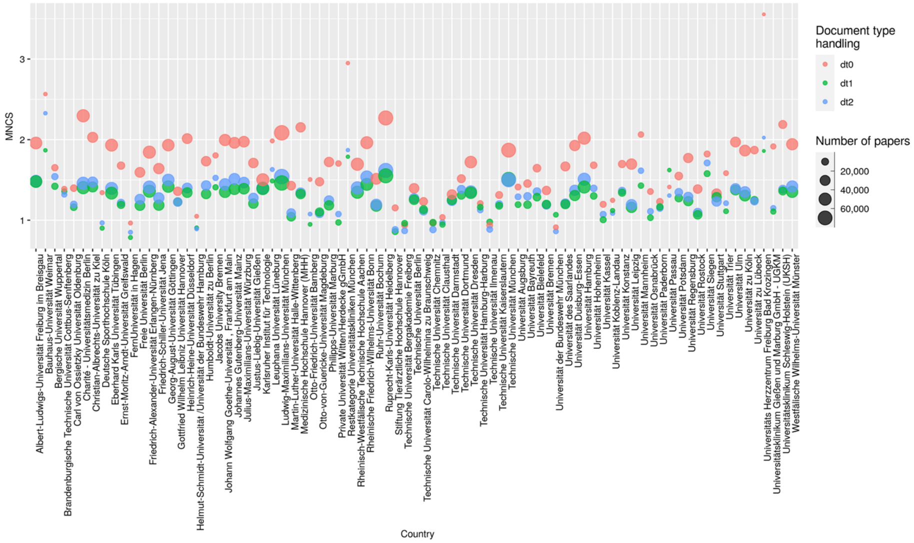

Fig. 7 Dot plot of MNCS values for the most productive German universities ($n=81$) using the three different normalization procedures with respect to the document type

Appendix A: Analyses on the paper level

(See Tables 11, 12, 13, 14, 15, 16, 17, 18)

Table 11 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only articles are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.98	0.99
NCS(dt1)	0.98	1.00	0.99
NCS(dt2)	0.99	0.99	1.00

Table 12 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only proceedings papers are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.98	1.00
NCS(dt1)	0.98	1.00	0.98
NCS(dt2)	1.00	0.98	1.00

Table 13 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only reviews are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.98	0.98
NCS(dt1)	0.98	1.00	0.97
NCS(dt2)	0.98	0.97	1.00

Table 14 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only letters are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.97	1.00
NCS(dt1)	0.97	1.00	0.98
NCS(dt2)	1.00	0.98	1.00

Table 15 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only articles are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.838, 0.839]	[0.932, 0.932]
NCS(dt1)	0.84	1.00	[0.907, 0.907]
NCS(dt2)	0.93	0.90683	1.00

The upper triangle shows the 95% confidence intervals

Table 16 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only proceedings papers are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.628, 0.628]	[0.995, 0.995]
NCS(dt1)	0.63	1.00	[0.616, 0.617]
NCS(dt2)	0.99	0.62	1.00

The upper triangle shows the 95% confidence intervals

Table 17 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only reviews are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.418, 0.419]	[0.920, 0.920]
NCS(dt1)	0.42	1.00	[0.506, 0.507]
NCS(dt2)	0.92	0.51	1.00

The upper triangle shows the 95% confidence intervals

Table 18 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only letters are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.322, 0.324]	[0.924, 0.925]
NCS(dt1)	0.32	1.00	[0.209, 0.211]
NCS(dt2)	0.92	0.21	1.00

The upper triangle shows the 95% confidence intervals

Appendix B: Analyses with respect to the 84 most productive countries

(See Tables 19, 20, 21, 22, 23, 24)

Table 19 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only articles are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.99	0.99
NCS(dt1)	0.99	1.00	0.98
NCS(dt2)	0.99	0.98	1.00

Table 20 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only proceedings papers are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.87	1.00
NCS(dt1)	0.87	1.00	0.86
NCS(dt2)	1.00	0.86	1.00

Table 21 Spearman rank correlation coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only reviews are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	0.97	0.97
NCS(dt1)	0.97	1.00	0.93
NCS(dt2)	0.97	0.93	1.00

Table 22 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only articles are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.442, 0.596]	[0.652, 0.779]
NCS(dt1)	0.52	1.00	[0.827, 0.905]
NCS(dt2)	0.72	0.87	1.00

The upper triangle shows the 95% confidence intervals

Table 23 Concordance coefficients of the NCS values from data sets dt0, dt1, and dt2 when only proceedings papers are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.080, 0.144]	[0.993, 0.996]
NCS(dt1)	0.11	1.00	[0.075, 0.135]
NCS(dt2)	0.99	0.11	1.00

The upper triangle shows the 95% confidence intervals

Table 24 Concordance coefficients of the NCS values obtained from data sets dt0, dt1, and dt2 when only reviews are considered

	NCS (dt0)	NCS (dt1)	NCS (dt2)
NCS(dt0)	1.00	[0.058, 0.105]	[0.612, 0.748]
NCS(dt1)	0.08	1.00	[0.093, 0.167]
NCS(dt2)	0.69	0.13	1.00

The upper triangle shows the 95% confidence intervals

Acknowledgements The bibliometric data used in this paper are from an in-house database developed and maintained by the Competence Center for Bibliometrics (CCB) and derived from the Science Citation Index Expanded (SCI-E), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (AHCI) prepared by Clarivate Analytics (Philadelphia, Pennsylvania, USA).

Funding Open Access funding enabled and organized by Projekt DEAL. This study was supported by Bundesministerium für Bildung und Forschung (Grant No. 01PQ17001).

References

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1*(1), 363–376. https://doi.org/10.1162/qss_a_00018

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics, 9*(2), 408–418.

Clarivate Analytics. (2021). InCites indicators handbook. Retrieved 8 July 2022, from http://incites.help.clarivate.com/Content/Indicators-Handbook/ih-about.htm

Clarivate Analytics. (2021). The Clarivate Analytics Impact Factor. Retrieved 24 March 2021, from https://clarivate.com/webofsciencegroup/essays/impact-factor/

CWTS. (2022). CWTS Leiden Ranking Indicators. Retrieved 2 May 2022, from https://www.leidenranking.com/information/indicators

Elsevier. (2019). *Research metrics guidebook*. Elsevier, Amsterdam, the Netherlands.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics, 53*(2), 171–193. https://doi.org/10.1023/A:1014848323806

Haunschild, R., Daniels, A.D. and Bornmann, L. (2022). Scores of a specific field-normalized indicator calculated with different approaches of field-categorization: Are the scores different or similar? *Journal of Informetrics, 16* (1), 101241, ISSN 1751-1577. https://doi.org/10.1016/j.joi.2021.101241. Accessed 17 May 2021

Haunschild, R., Marx, W., French, B., & Bornmann, L. (2018). *Relationship between field-normalized indicators calculated with different approaches of field-categorization*. Paper presented at the 23rd International Conference on Science and Technology Indicators, Leiden (The Netherlands). https://scholarlypublications.universiteitleiden.nl/handle/1887/65267. Accessed 17 May 2021

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies, 1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431.

Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation metrics: A primer on how (not) to normalize. *PLoS Biology, 14*(9), e1002542. https://doi.org/10.1371/journal.pbio.1002542

Lin, L. I. (1989). A concordance correlation-coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268. https://doi.org/10.2307/2532051

Lin, L. I. (2000). A note on the concordance correlation coefficient. *Biometrics, 56*(1), 324–325. https://doi.org/10.1111/j.0006-341X.2000.00324.x

Lundberg, J. (2007). Lifting the crown: Citation *z*-score. *Journal of Informetrics, 1*(2), 145–154.

McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient*. National Institute of Water & Atmospheric Research: Hamilton, New Zealand

Nederhof, A. J., & Visser, M. S. (2004). Quantitative deconstruction of citation impact indicators. *Journal of Documentation, 60*(6), 658–672. https://doi.org/10.1108/00220410410568142

OurResearch. (2021). We're building a replacement for Microsoft Academic Graph. Retrieved 11 April 2022, from https://blog.ourresearch.org/were-building-a-replacement-for-microsoft-academic-graph/

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions. https://arxiv.org/abs/2205.01833. Accessed 17 May 2021

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/. Accessed 17 May 2021

Rehn, C., Kronman, U., & Wadskog, D. (2007). *Bibliometric indicators: Definitions and usage at Karolinska Institutet*. Karolinska Institutet University Library.

Scheidsteger, T., Haunschild, R., Hug, S., & Bornmann, L. (2018). *The concordance of field-normalized scores based on Web of Science and Microsoft Academic data: A case study in computer sciences* Paper presented at the STI 2018 Conference Proceedings, Leiden, The Netherlands. https://scholarlypublications.universiteitleiden.nl/handle/1887/65358?solr_nav%5Bid%5D=3add44046ea4b81cb330&solr_nav%5Bpage%5D=0&solr_nav%5Boffset%5D=0. Accessed 17 May 2021

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Zeileis, A. (2020). DescTools: Tools for descriptive statistics. Retrieved 24 March 2021, from https://cran.r-project.org/package=DescTools

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). *An overview of Microsoft Academic Service (MAS) and applications.* Paper presented at the 24th International Conference on World Wide Web (WWW '15 Companion), Florence, Italy.

Van Leeuwen, T. N., Moed, H. F., & Reedijk, J. (1998). A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation, 54*(4), 387–419. https://doi.org/10.1108/EUM0000000007175

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365–391.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C., Tijssen, R. J., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F., Visser, M. S., & Wouters, P. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419–2432. https://doi.org/10.1002/asi.22708

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics, 5*(1), 37–47. https://doi.org/10.1016/j.joi.2010.08.001

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies, 1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H. (2017). Tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1. Retrieved 22 June 2020, from https://CRAN.R-project.org/package=tidyverse