



Regular article

Publication counting methods for a national research evaluation exercise

Przemyslaw Korytkowski^a, Emanuel Kulczycki^{b,*}^a West Pomeranian University of Technology, Szczecin, Poland^b Scholarly Communication Research Group, Adam Mickiewicz University in Poznań, Poland

ARTICLE INFO

Article history:

Received 22 December 2018

Received in revised form 31 May 2019

Accepted 4 July 2019

Available online 30 July 2019

Keywords:

Evaluation

Counting method

Whole counting

Fractional counting

Complete counting

ABSTRACT

In this paper, we investigate the effects of using four methods of publication counting (complete, whole, fractional, square root fractional) and limiting the number of publications (at researcher and institution levels) on the results of a national research evaluation exercise across fields using Polish data. We use bibliographic information on 0.58 million publications from the 2013–2016 period. Our analysis reveals that the largest effects are in those fields within which a variety publication and cooperation patterns can be observed (e.g. in Physical sciences or History and archeology). We argue that selecting the publication counting method for national evaluation purposes needs to take into account the current situation in the given country in terms of the excellence of research outcomes, level of internal, external and international collaboration, and publication patterns in the various fields of sciences. Our findings show that the social sciences and humanities are not significantly influenced by the different publication counting methods and limiting the number of publications included in the evaluation, as publication patterns in these fields are quite different from those observed in the so-called hard sciences. When discussing the goals of any national research evaluation system, we should be aware that the ways of achieving these goals are closely related to the publication counting method, which can serve as incentives for certain publication practices.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Multi-authored publications are the key output within various assessments of national research productivity and impact (Huang, Lin, & Chen, 2011; Zacharewicz, Lepori, Reale, & Jonkers, 2018). In ongoing discussions on publication counting, the effects of the different methods are usually considered in relation to country and university rankings based on publications and citations (Gauffriau, Larsen, Maye, Roulin-Perriard, & Von Ins, 2008; Hagen, 2014; Waltman et al., 2012).

The two most often used publication counting methods are whole (full) counting and various variants of fractional counting (Larsen, 2008; Waltman & van Eck, 2015). In the former method, each entity (country/institution/author) gets full credit for co-authored papers. Using the latter method allows to proportionally fractionalize credit across all the contributing entities. Van Hooydonk (1997) argue that fractional counting can be refined into proportional counting to calculate a relative credit depending on the author's rank on a multi-authored publication. However, this method is applicable only in fields in which the order of author list is not alphabetical.

* Corresponding author.

E-mail address: emek@amu.edu.pl (E. Kulczycki).

Gaufrriau et al. (2008) show that whole counting is favorable to certain countries with a high level of internationalization. Waltman and van Eck (2015) present an empirical analysis which illustrates that the best choice is to use fractional counting instead of full counting because only this method allows us to generate the field-normalized results. Aksnes, Schneider, and Gunnarsson (2012) follow this conclusion and show that the difference between whole and fractionalized counts in rankings by citation indicators is greatest for the countries with the highest proportion of internationally co-authored articles. Gaufrriau (2017) provides an overview of arguments for counting methods and shows that there is not often an explicit motivation for choosing a specific method.

Most studies on publication counting methods are based on the international indexes of publications, such as the Web of Science (WoS) or Scopus. However, publication counting is also used in performance-based research funding systems (PRFSs) which calculate a variety of bibliometric indicators to produce rankings across fields or institutions in one country (Aagaard & Schneider, 2015). In PRFS systems, they also use publications collected in national databases that cover not only WoS or Scopus listings but also articles from local journals and all types of scholarly book publications (Sile et al., 2018). This is crucial for the social sciences and humanities, especially where the WoS/Scopus coverage degree is very low (Ossenblok, Engels, & Sivertsen, 2012; Prins, Costas, Van Leeuwen, & Wouters, 2016; Sivertsen & Larsen, 2012) and where scholarly book publications play the major role (Kulczycki et al., 2018). On the basis of the Norwegian database, Piro, Aksnes, and Rørstad (2013) show how publication counting methods can change the picture of a researcher's productivity across different fields. When whole counting is used, researchers from the so-called hard sciences are found to be more productive than from the soft sciences. Changing the method into fractional counting completely reverses this picture.

In this paper we use the term 'publication counting methods for a national research evaluation exercise' (in short: publication counting methods) in a broad sense to cover all aspects of bibliographic data into a score for a field (within an institution) used in a performance-based research funding system.

We identified five dimensions of publication counting methods for a national research evaluation exercise: (1) unit of assessment, (2) counting method, (3) institution limit, (4) researcher limit, and (5) point scale. The publication counting method is just one of many elements of the broader challenge facing a national research evaluation system, which are not focused on just publications. Nonetheless, publications are the most important criterion.

The first dimension reflects the unit of assessment. For instance, in Norway, scientific institutions are assessed, while Poland has changed their unit of assessment from whole institutions, e.g. a faculty comprising researchers from a few fields (the 2013 and 2017 evaluations) to separate fields within institutions (the 2021 evaluation). In this way, the new Polish unit of assessment is similar to the solution used in the Research Excellent Framework (REF) in the United Kingdom.

The second dimension, counting method, is related to a fair approach towards multi-authored publications (Sivertsen, 2018). In Norway, square root fractional counting is used, while Poland use a combination of three methods, i.e. whole counting, fractional counting, square root counting. Sivertsen (2016), in a discussion on redesigning publication counting methods in PRFSs, highlights that there is a need for a method which will allow for balancing across different field-dependent co-authorship practices. According to Sivertsen, such a method is square root fractional counting. Having a similar point of view with Sivertsen, we believe also that square-root counting can be useful not only for balancing results across fields but also for analyses carried out for separate fields. All fields have subfields, which have diverse publication practices, for instance, theoretical physics and high energy physics within the field of physics.

The third dimension concerns an institution limit, i.e. limiting the number of submitted publications from a single unit of assessment (institution or field, depending on the adopted evaluation model). The number of publications per FTE most often expresses this limit. For instance, in Norway, all publications of an evaluated institution are used, whereas in Poland in the 2013 and 2017 evaluations, only a limited number of publications were included in the calculation. This limit is expressed by the $3N-2N_0$ formula, where N is the arithmetic mean of the full-time equivalent (FTE) of academic staff members who work in a given scientific unit during the evaluated four-year period, while N_0 is the number of academic staff members who were not authors of any publication during the period in question (Kulczycki, Korzeń, & Korytkowski, 2017). In the UK Research Excellence Framework 2021, the average number of publications required per FTE in the unit of assessment is 2.5 (Research Excellence Framework, 2018).

The fourth dimension is related to a researcher limit, i.e. limiting the number of submitted publications by one researcher. For instance, there is no researcher limit in Norway. In Poland, according to regulations for the Polish evaluation in 2021, the maximum number of publications required per FTE is 4. In REF 2014, one researcher could submit up to four research outputs whereas in REF 2021 at least one and a maximum of five outputs.

The fifth dimension concerns the points assigned to the publication channels and thus to publications. The point scale can be linear like in the Polish evaluation exercise in 2017 (institutions could obtain from 1 to 50 points for an article), or non-linear as in Finland and Norway (1 or 3 points per article depending on the level of the scientific publication channel). The Norwegian Publication Indicator used within the Norwegian PRFS is a point system which categorizes all publications into two levels and assigns different numbers of points to them (Aagaard, Bloch, & Schneider, 2015). This point system is weighted in terms of both level and publication type (journal articles, articles in anthologies, and monographs). The point systems are also implemented among others in Denmark (Aagaard, 2018), Finland (Pölönen, 2018), Flanders (Engels & Guns, 2018), and Poland (Kulczycki et al., 2017). In each of these counties, various methods of distributing points to institutions which had contributed to the publication might be found.

In this study, we consider publication counting methods rather as science policy tools in research evaluation, than as tools which serve to describe the existing characteristics of researchers' productivity and their publication patterns. We

compare the effects of using four methods of publication counting and of limiting the number of publications, on the results of a research evaluation exercise using Polish data.

We use data from the national evaluation exercise in 2017 in which research outputs from the 2013–2016 period of all universities were assessed. We analyze how the change of publication counting method influences the results of rankings across the fields. Moreover, we investigate the effects of limiting the number of considered publications in the evaluation exercise.

In Poland, the results of evaluation are not directly translated into funding for institutions. Apart from publications, data concerning several other parameters are gathered for the purposes of the evaluation exercise (Kulczycki et al., 2017). These parameters are aggregated into four main criteria, which are later weighted and summed. As a result, the position of a scientific unit is determined among similar units in terms of field. Based on the position of the unit in the ranking, a scientific category (A+, A, B or C) is assigned by the Ministry. Ultimately, the scientific category translates into the size a block grant from the Ministry. The block grants in case of units from universities is about 10% of their annual budget, while for basic and applied research institutes it is up to 30% of their annual budget.

This article adds to the ongoing discussion by showing the effects of the different publication counting methods on the results of a national research evaluation system calculated at the level of fields. We are interested in examining which methods of publication counting favor the so-called hard and soft sciences, and the effects of limiting the number of publications required per researcher.

Our main research question is twofold: how do different publication counting methods influence the field rankings, and how does a researcher limit change the rankings? This paper presents an original study based on bibliographical data from the Polish national research evaluation system. Thanks to this, the study is limited neither to data from the Web of Science nor from Scopus, from which the coverage degree is insufficient to evaluate the social sciences and humanities, especially from a non-English speaking country. Data from a national evaluation is more comprehensive and balanced, *i.e.* data includes scholarly book publications and articles from local scholarly journals.

The usefulness of this article is that we use the point scale (from 1 to 50) – with little modification – that is known to all Polish researchers since 2008. This scale was used in the evaluation exercises in 2013 and in 2017 (details about the point scale are presented in Dataset section). It is also quite often used in promotion procedures and in the periodic assessment of employees. Therefore, researchers are familiar with the framework and the details of the point scale and researchers have adapted – to some extent – their own publishing practices to this scale. Publication channels with a higher number of points are widely recognized as more prestigious and therefore are perceived as channels in which it is worth to publish. Thus, using the one and well-assimilated point scale in all analyses allow us to coherently assess the consequences of the various counting methods for multi-author publications in a national research evaluation system.

The rest of this paper is organized as follows. In Section 2, we present the data, methods focusing on how the data have been prepared for analysis, and what variants of publication counting method were used. In Section 3, we present the results focusing on the effects of variants for fields. Then, in Section 4, we discuss the main findings. In the final Section 5 we present conclusions.

2. Data and methods

2.1. Dataset

In our analysis, we use a data set from the last cycle of research evaluation in Poland conducted in 2017. Scientific units submitted bibliographical records of 581,106 publications, with the FTE of academic staff members at 86,461.84. For each evaluated publication, the given scientific unit obtains a specified number of points, depending on a variety of factors, including the type of publication channel and the number of authors. Articles from journals indexed in the Journal Citation Reports could obtain from 15 to 50 points (on the basis of the five-year impact factor normalized using the Web of Science subject categories). Articles from local journals can obtain from 1 to 15 points. Articles from journals indexed in the European Reference Index for Humanities from 10 to 25 points. Monographs could obtain 25 points and a chapter in monographs obtains 5 points. Detailed information on assessing publications in the Polish system has been presented in previous publications (Kulczycki, 2017; Kulczycki & Rozkosz, 2017).

2.2. Mapping publications to the OECD fields

The Polish evaluation was conducted at the level of scientific institutions, and originally publications are classified to fields according to the organizational classification (Daraio & Glänzel, 2016). For the purpose of this analysis, we have organized all academic staff members into fields according to disciplines (from the Polish classification) declared by them for the purpose of evaluation. We mapped these disciplines to the fields of science and technology in the Organisation for Economic Co-operation and Development (Organisation for Economic Co-Operation and Development (2007) – to make our analyses clearer. Using original classification would be vague for readers because the organizational units were grouped into the Joint Evaluation Groups (Kulczycki et al., 2017) built across the fields. Moreover, Poland has changed field and discipline classification to some variant of OECD FOS classification and reduced the number of disciplines to 44. All these nuances are not relevant for the analyses results but they show rationales for our decision of using the mapping.

Table 1
Characteristics of eight variants.

Variant	Counting method	Researcher limit	Slot	Points
1	Whole counting	No	1	x
2	Complete counting	No	k	$x \cdot k$
3	Fractional counting	No	$k/(k+m)$	$x \left[\frac{k}{k+m} \right]$
4	Square root fractional counting	No	$\sqrt{k/(k+m)}$	$x \sqrt{k/(k+m)}$
5	Whole counting	3	1	x
6	Complete counting	3	k	$x \cdot k$
7	Fractional counting	3	$k/(k+m)$	$x \left[\frac{k}{k+m} \right]$
8	Square root fractional counting	3	$\sqrt{k/(k+m)}$	$x \sqrt{k/(k+m)}$

Finally, in this analysis, all publications published by researchers classified as sociologists are counted as publications from sociology. All scholars representing artistic production were excluded from the analysis, as in the overwhelming majority of cases they submitted artwork to the evaluation.

Performing the mapping required expert decisions due to the fact that some disciplines from the Polish classification can be attributed to several fields from the OECD classification. For example, in the Polish classification, computer science is in mathematical sciences and technical sciences, whereas in the OECD classification it is in natural sciences (the vast majority of researchers) and partly in electrical engineering, electronic engineering, information engineering, and media & communications. For this reason, in the case of several disciplines, the researchers representing these have been entirely attributed to the dominant OECD field. It was our expert decision based on the data of research fields and publications of researchers from a given discipline.

2.3. Variants of publication counting methods

In this analysis, we use eight variants of publication counting methods for the national research evaluation exercise. In terms of the five dimensions presented in Section 1, each variant unit of assessment, institution limit and point scale (Dimensions 1, 3, and 5) of the research evaluation system is invariable, so we investigate the counting method (Dimension 2) and researcher limit (Dimension 4).

We restrict the analysis to cases with an institution limit set at the level of *3FTE* (*3N* in Polish terminology) due to the available data. The data at our disposal does not contain full information about all scientific publications of Polish researchers. During the analyzed period, *i.e.* 2013–2016, according to the national current research information system (Polish Scholarly Bibliography), Polish researchers submitted around 1.03 million scientific outputs, while to the evaluation exercise 0.58 million publications were submitted for the same time period. In the Polish research evaluation only $3N - 2N_0$ publications were taken into account when determining the final result. For this reason, some of the scientific units for evaluation purposes were only a part of the total number of their employees' publications, uploaded to fill the required limit by a certain margin. In the analysis, we could not use data from the Polish Scholarly Bibliography due to a lack of information about assigned points to publications. It should also be highlighted that experts, examining whether publications within the institution limit met the various formal requirements, were in control of the quality of evaluation data.

In the analysis, we took into account four author-level counting methods: whole (full) counting, complete counting, fractional counting, and fractional counting with square root. We separated these four variants into cases either with a researcher publication limit (3 slots per researcher) or without a researcher publication limit. One publication occupies one slot in the case of whole counting. In case of complete counting, one publication occupies k slots, where k is the number of authors affiliated to the institution representing the field. An attribution to k is based on the affiliation in the publication and the field selected by the scholar. In the case of fractional counting, one publication occupies $k/(k+m)$ of a slot, where m is the number of other authors representing other field(s) from the same institution or other institution(s). In the case of square root fractional counting, one publication occupies $\sqrt{k/(k+m)}$ of a slot.

The number of assigned points x is equal to the whole number of points attributed to a single-author publication multiplied by the occupied slot (Table 1).

As was described above, all scientific units submitted at least $3N - 2N_0$ publications. However, only a few institutions provided information about their all publications from the whole period.

In the analysis, therefore, we restricted the data set to only those fields in which we were able to collect *3FTE* publications in Variant 1 after changing the unit of analysis from organization units to fields. Variants with fractional counting and a researcher limit required more publications to fill the available slots, and in the case of some fields within institutions, the publication pool was not large enough.

In order to rank institutions within a field, we only used publication criterion $C \in [0, 50]$. C should be interpreted as the scientific power of a field from a given institution, where a low value means a weak scientific power and a high value means strong scientific power. A value of 50 is achievable if all *3FTE* publications are published in the best channels for which 50

points are awarded according to the Polish list. C is calculated using one of two Eqs. (1 or 2). If there were enough publications to fill all available slots, i.e. $\sum_i s_i = 3FTE$ then

$$C = \frac{\sum_i x_i s_i}{3FTE} \quad (1)$$

If there were not enough publications to fill all available slots, i.e. $\sum_i s_i < 3FTE$ then

$$C = \frac{\sum_i x_i s_i + x_{min} (3FTE - \sum_i s_i)}{3FTE} \quad (2)$$

where x_i is the number of points assigned to a single-author publication according to rules of the Polish research evaluation system 2017 for publication i , s_i is the slot occupied by publication i

$$s_i = \begin{cases} 1 \\ k \\ k/(k+m) \\ \sqrt{k/(k+m)} \end{cases}$$

$x_{min} = \min_i x_i$ is the minimum number of points assigned to a single-author publication from a field from an institution.

The number of available slots for a field from an institution is limited in all analyzed variants (the institution limit) to three times the number of full time equivalents

$$\sum_i s_i \leq 3FTE \quad (3)$$

For Variants 5 through 8, we introduced an additional limit of three slots per researcher (researcher limit)

$$\forall_r \sum_i \frac{s_i}{k_i} \leq 3 \quad (4)$$

where r is the index for the researcher.

The publications were ordered on the basis of the number of points, and then the slots were filled until the publication list was exhausted, e.g. because of the researcher limit. If after exhausting the pool of all publications the slots were not filled with publications, we adopted the principle that unoccupied slots are filled with virtual publications with the lowest scores from those allocated. As it has already been shown, Polish scientists published twice as many works as were reported for research evaluation. Thus, our decision is justified by the observations from the previous Polish research evaluation exercises in 2013 and 2017, where institutions had in their pools a large number of publications not included in the $3N$ limit, with a score similar to that from the cut-off point.

2.4. Field as a unit of analysis

In this study, we analyzed the data at the level of fields. We assigned a code to each field where the first digit represented one of the six major OECD Fields (1: Natural sciences, 2: Engineering and Technology, 3: Medical and Health sciences, 4: Agricultural sciences, 5: Social sciences, 6: humanities).

Table 2 presents each field in terms of the total number of institutions classified in a given field (e.g. universities, research institutes) and the FTE of academic staff members classified in a given field. A field in the institution is subject to research evaluation when the FTE is greater than 12, which is in line with Polish law on science and higher education from 20th July 2018. This analysis is restricted to fields with at least 10 institutions to ensure enough units of analysis for rankings. These two provisions reduced the analyzed data volume down to almost 70,000 FTE.

Finally, the researchers were classified into 29 fields representing all six major OECD fields to produce 875 units of assessment across 245 institutions, and then analyzed. A unit of assessment is one research field in one institution. The largest field in terms of the number of institutions is Economics and business with 77 institutions. The largest field in terms of FTE is Clinical medicine with almost 8000 FTE. The smallest field is Environmental biotechnology with only 11 institutions and about 280 FTE.

Fig. 1 shows boxplots of the FTE of academic staff members across the fields. The mean size of a unit of assessment (a field within an institution) is 78.96 FTE, ranging from 28 FTE in Environmental biotechnology to 248 FTE in Clinical medicine. The median for all fields is much lower than the average at only 46.76 FTE. The single biggest unit of assessment is 971 FTE. There are 170 units of assessment with less than 20 FTE, and 48 units of assessment with FTE higher than 250.

Table 2
Characteristics of the analyzed fields.

Code	Field	Number of institutions	Sum of FTE
11	Mathematics	29	1,167.56
12	Computer and information sciences	38	1,816.28
13	Physical sciences	43	3,052.82
14	Chemical sciences	44	3,024.92
15	Earth and related environmental sciences	23	1,956.02
16	Biological sciences	50	3,281.82
21	Civil engineering	33	2,898.06
22	Electrical engineering, electronic engineering, information engineering	33	3,168.99
23	Mechanical engineering	36	4,143.81
24	Chemical engineering	15	824.6
25	Materials engineering	33	1,744.26
27	Environmental engineering	26	1,816.88
28	Environmental biotechnology	11	279.4
31	Basic medicine	22	3,169.54
32	Clinical medicine	32	7,922.47
33	Health sciences	40	2,684.24
41	Agriculture, forestry, and fisheries	27	3,194.44
42	Animal and dairy science	11	742.31
51	Psychology	19	906.97
52	Economics and business	77	6,397.29
53	Educational sciences	27	1,596.85
54	Sociology	29	1,439.47
55	Law	34	2431.44
56	Political Science	36	1,553.48
58	Media and communications	12	339.5
61	History and archaeology	29	1,904.36
62	Languages and literature	29	4,048.62
63	Philosophy, ethics and religion	25	1,388.99
64	Art (arts, history of arts, performing arts, music)	12	543.81
	<i>Total</i>	875	69,439.20

Note: FTE – full-time equivalent of academic staff members.

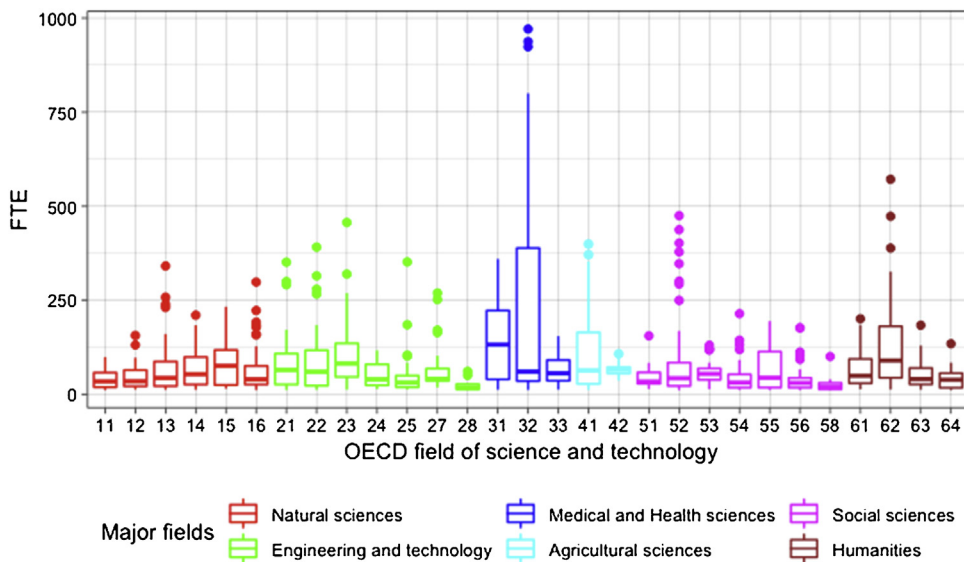


Fig. 1. Boxplots of full-time equivalent of academic staff members across the fields.

2.5. Limitations of the study

For the analysis we included only those scientific units in which at least 12 FTE academic staff members were assigned to a given field. Our decision was inspired by a new regulation for the 2021 Polish evaluation, according to which only scientific units with 12 FTE in a given field will be evaluated. Moreover, we analyzed only those fields to which there are at least 10 scientific units assigned.

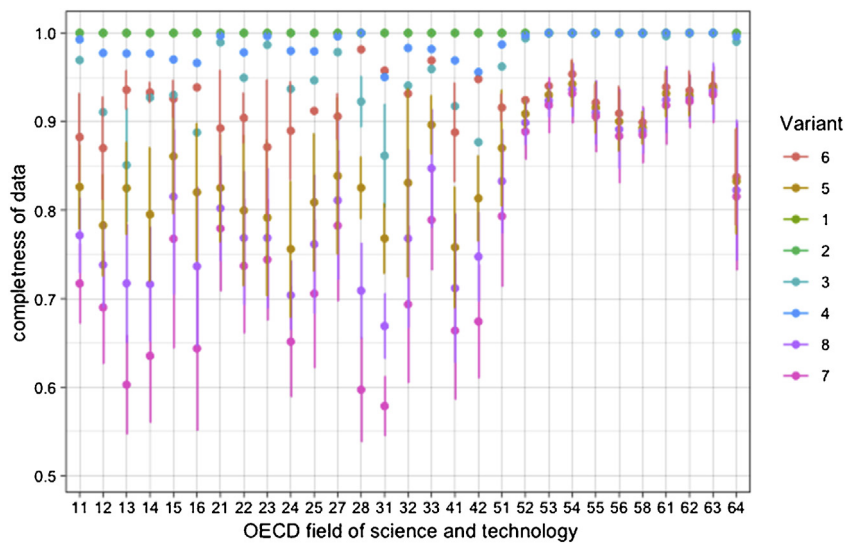


Fig. 2. The completeness of publication data depending on the variant of publication counting methods.

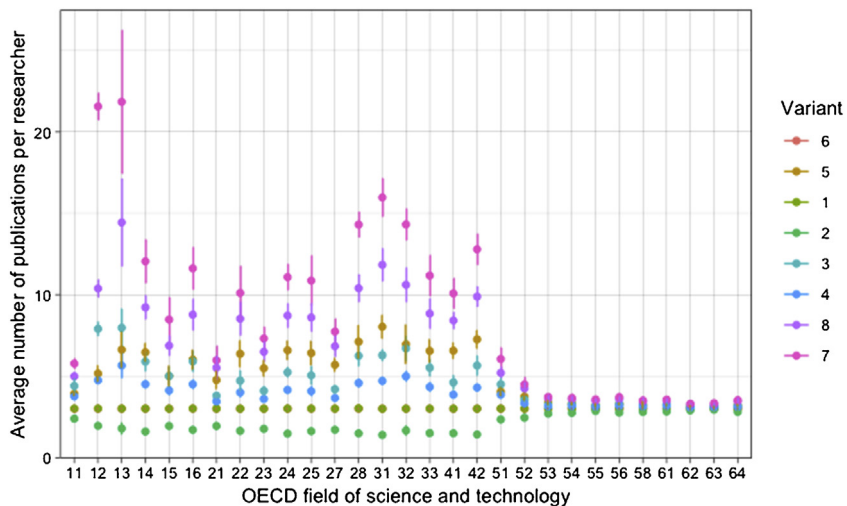


Fig. 3. The average number of publications per researcher across Variants of publication counting methods and OECD fields of science and technology.

In Fig. 2, we present the data completeness for all the analyzed fields and variants. It shows how many virtual publications had to be added in a given variant for a given field. The lowest data completeness is for the so-called hard sciences in Variant 7 (because of fractional counting and the researcher limit of three slots), and a little better completeness for variant 8 due to square root fractional counting.

This incompleteness of the data limits the validity of the results, however, as we have written above, experience from the two last evaluation exercises shows that institutions have a large pool of publications not reported for evaluation. Thus, using virtual publications is a good proxy, which should not significantly affect the results. Due to a lack of necessary information to carry out these analyses (e.g. points assigned to the publications), we could not use data from a national level current research information system, i.e. the Polish Scholarly Bibliography.

Due to low data quality about the number of authors, we have excluded from analysis all edited volumes and monographs with suspiciously large numbers of authors (e.g. 100 editors of an edited volume). In many cases, editors were mixed with chapter authors. This was a result of imprecise provisions of implementing the act and explanations in the data collection software for evaluation purposes.

We analyzed how many publications (shares in publications in case of Variants 5–8) have to be provided for evaluation by a researcher depending on the selected counting method. Fig. 3 presents the results of such an analysis broken down by fields. For Variant 1, it is always three publications, as here whole counting with 3FTE limit is used. In the hard sciences, the number of provided publications in Variant 2 falls below three while in the social sciences and humanities it stays close to three. This shows that in social sciences and humanities, researchers work more often work alone and publish the highest

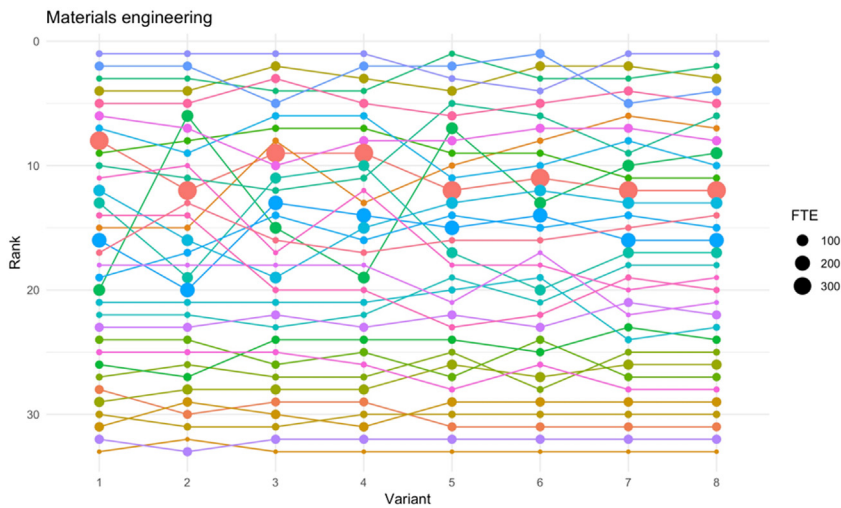


Fig. 4. Field rankings of institutions through the eight analyzed variants for Materials engineering.

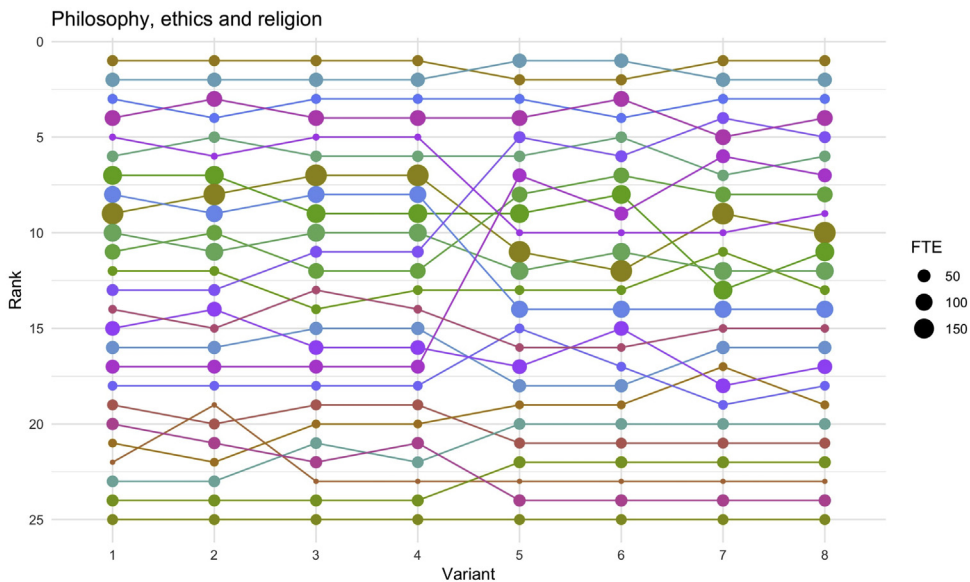


Fig. 5. Field rankings of institutions through the eight analyzed variants for Philosophy, ethics and religion.

share of single-author publications than researchers from the hard sciences. The variant requiring the greatest number of required publications is Variant 7, just before Variant 8. This is especially visible in Physical sciences and in Computer and information sciences.

3. Results

According to the methodology presented in Section 2.3, for each unit of assessment we calculated the value of parameter C (Eqs. (1) and (2)), which determines the scientific power of the field. Next, we built a ranking of institutions for each field based on the value of parameter C . We repeated this procedure for all eight variants of publication counting methods presented in Table 1.

Figs. 4–7 presents charts with field rankings of institutions through the eight analyzed variants. The size of the sign represents the unit size expressed in FTE. Analogous charts for all 29 analyzed fields are in Appendix 1. We use ranks instead of the total number of points because rankings are easier to read as scientific institutions are spread all over the scale. In Appendix 2 we show two plots which allow one to compare these two types of results presentation on the example of Physical sciences.

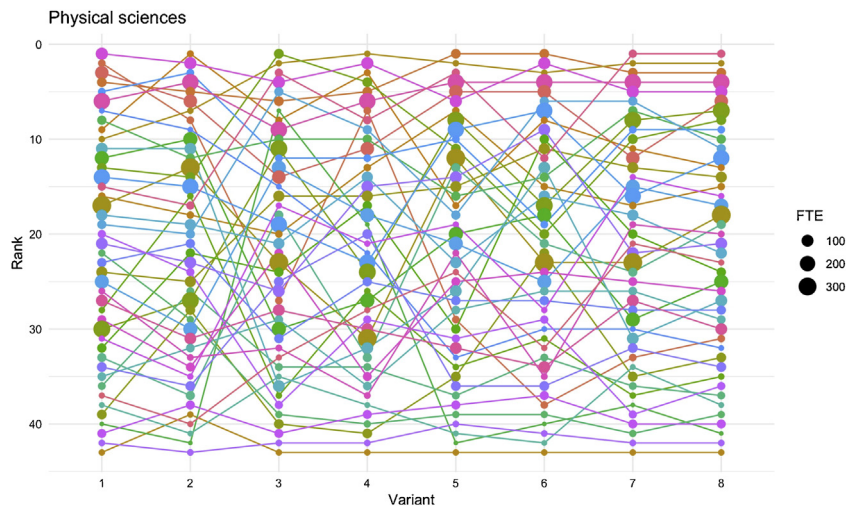


Fig. 6. Field rankings of institutions through the eight analyzed variants for Physical sciences.

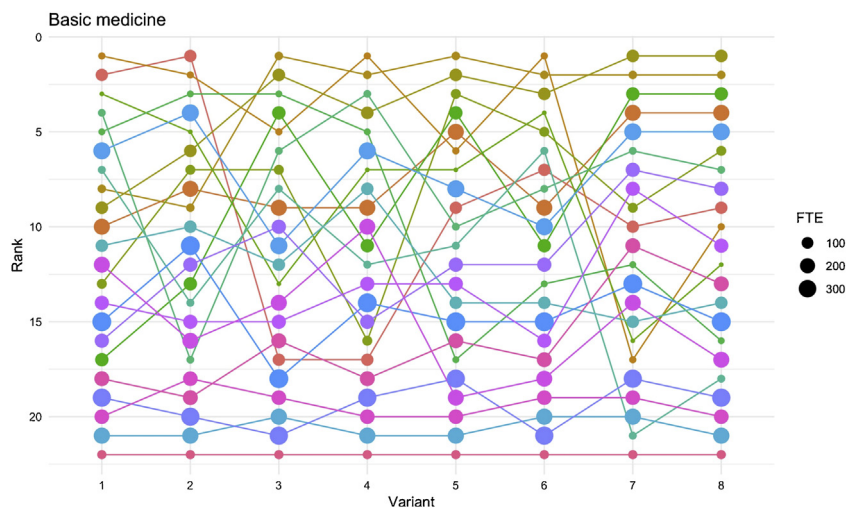


Fig. 7. Field rankings of institutions through the eight analyzed variants for Basic medicine.

In the article, we included charts for only four fields: two of the fields (*i.e.* Materials engineering, and Philosophy, ethics and religion) in which the positions in the ranking of institutions only slightly change because of the counting methods. The next two charts present the results for Physical sciences and Basic medicine. In these fields the positions in the ranking of the institutions strongly depend on the counting method.

In Figs. 4–7, it can be noticed that the largest institutions from the point of view of a given field usually occupy the middle positions in the ranking. At the top of the ranking are most often medium-sized institutions. The best institutions from the field quite often remain one of the best regardless of the analyzed counting variant. Similarly, institutions from the end of the ranking remain there regardless of the counting method. The biggest changes can be observed in the middle of the ranking.

Fig. 8 presents the Spearman's rank correlation coefficient for each field. It can be seen that for some fields the rankings are resistant to changes in the counting method. A low correlation between the rankings indicates that there are various publishing practices within the field, in particular the length of the list of authors. In Physical sciences, different practices are found in high-energy physics where great collaborations dominate (kilo-author publications) rather than in Theoretical physics (in four institution the average number of authors is above 1000 per publication while in six institutions it's less than four per publication). Similarly, in one field, history and archeology are included, but these two subfields have different research methods and cooperation patterns (an average number of authors is 1.28 and 2.22, respectively). Even in mathematics the change of the counting method has a large impact on the ranking (an average number of authors per institution ranges from 1.52 to 7.23).

Fig. 9 presents the Spearman's rank correlation coefficients per a pair of variants aggregated for all fields. One can observe that in general a transition from whole counting to complete counting (1–2 and 5–6) does not cause substantial changes in

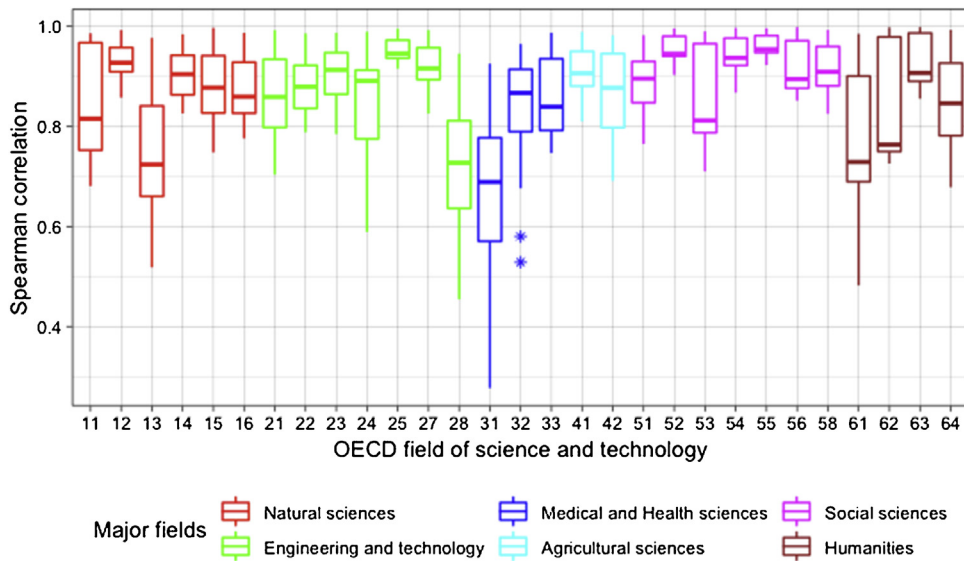


Fig. 8. Spearman's rank correlation coefficient for each OECD fields of science and technology.

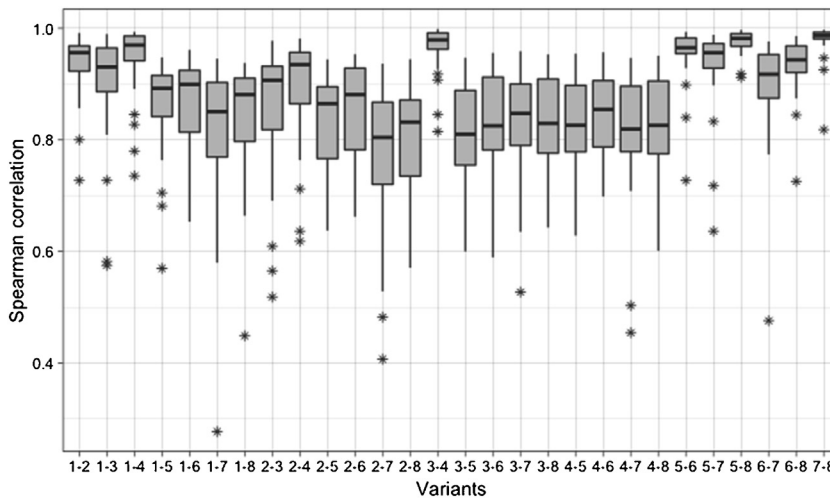


Fig. 9. Spearman's rank correlation coefficients per a pair of variants aggregated for OECD fields of science and technology.

the rankings. The same situation can be observed for a transition from fractional counting to square root fractional counting (3–4 and 7–8). More significant changes occur when a researcher limit is included (variants 5–8). The transition from Variant 1–4 has a higher Spearman correlation than to those where a researcher limit is present. The change is even bigger when the counting method is the same but the researcher limit is present or absent (1–5). Variability, due to the transition among variants with a researcher limit imposed (Variants 5–8), is significantly lower (the Spearman correlation is higher) when comparing variants without a researcher limit (Variants 1–4).

4. Discussion

In this paper, we discussed how publication counting methods for a national research evaluation exercise influence the rankings across the fields of science. We studied over 0.5 million publications submitted by Polish scientific institutions and analyzed eight variants.

Our study reveals that the largest differences are in those fields within which various publication and cooperation patterns (e.g. the number of authors) can be observed. For instance, the substantial effects observed in History and archeology and in Physics show that selecting the publication counting method should be based on the proper granularity of the fields. Determining the proper detailed granularity is to some extent a matter of merit, as the research evaluation system is a

science policy instrument, which serves not only for assessing higher education institutions or fields within them, but also functions as a set of incentives to influence researchers' publication practices.

We position our research also within research evaluation studies because in every evaluation exercise some counting methods (explicitly or implicitly) is used. In this paper, we analyze counting methods together with four other dimensions, that is the unit of assessment, the institution limit, the researcher limit, and the point scale. Therefore, we believe that our results might have implications also outside of evaluation and funding regimes, for example, university rankings.

The substantial effects of the different counting methods can be clearly observed in some fields (e.g. Clinical medicine) while other fields are not so sensitive. Our findings show that social sciences and humanities are not significantly influenced by changes in publication counting method because publication patterns in those fields are quite different from those observed in hard sciences.

Our observations and discussions with the academic community on this topic lead us to a conclusion that social scientists and humanities perceive whole counting and complete counting as an unfair way of assessing publication within the national system. One can say that evaluation is conducted at the level of fields, and researchers from one field are compared only with researchers from the same field. However, researchers from different fields compare themselves with each other because they work in the same higher education and science system. Moreover, public opinion and society treat all researchers as representatives of the same group.

Below we discuss each of the five dimensions of the research evaluation system related to publication counting methods.

Dimension 1: Unit of assessment

An organizational unit within an institution (e.g. faculty, research institute, other higher education institution) or a field (discipline) can be a unit of assessment within the national research evaluation system. In evaluating institutions, it is important to assess homogeneous units, i.e. to assess and compare, for instance, a faculty of history with another faculty of history. When organizational units are heterogeneous (researchers represent various fields), then one field – favored by the publication counting method – can dominate (in terms of obtained points) within this institution. Actually, the same situation exists with evaluating the field. When we constitute a unit of assessment as a field and at the same time aggregate the different fields (in terms of different publication patterns) into a single field, then we have heterogeneous units of assessment. Such a situation can be observed in our results in Physical sciences and in History and archeology.

Dimension 2: Counting method

Four researcher-level counting methods favor various publication patterns and behaviors. Below we discuss each method and argue what is favored and what is underestimated by a given method.

Whole counting: this method favors any type of internal or external cooperation regardless of the contribution from a given unit of assessment. Whole counting underestimates the given unit of assessment when it plays a key role in the publications.

Complete counting: this method favors a unit of assessment from which there are many authors of a given publication, which can reflect the contribution of this unit. Complete counting underestimates single-authored publications. There is a substantial difference to be the only author of a monograph and to be one of four authors of a monograph. Therefore, complete counting is not a balanced method for assessing different fields within one national system from the perspective of fields in which single-authored publications constitute the majority of total volume. Complete counting requires reporting fewer publications for the evaluation than whole counting when an institution limit is used (see Fig. 9).

Fractional counting: this method favors single-authored publications and a high share of contribution (in terms of the number of authors) in publications. However, this bonus is not as significant as in complete counting. Fractional counting underestimates wide cooperation-networks even though participating in many joint studies and projects requires a significant workload and usually is the result of effective networking. Fractional counting requires reporting 1.5–2 times more publications for the evaluation than whole counting when an institution limit is used. Moreover, researchers from social sciences and humanities perceive fractional counting as a more balanced and fair way of publication counting because a single-authored monograph is not equal to one multi-authored article.

Square root fractional counting: this variant of fractional counting is used to mitigate the consequences of fractional counting in order to give more credit to units of assessment with a wide cooperation-network. At the same time, square root fractional counting can be perceived as a balanced and fair way of publication counting by researchers from social sciences and humanities. This method requires a few less publications for the evaluation than fractional counting when an institution limit is used.

Dimension 3: Institution limit

This limit is a pragmatic way of showing that the quality of publications is more important the quantity of research outputs. This common sense intuition might be connected with Bradford's law of scattering or a Pareto distribution, which show that only some part of research output is important. In other words: only articles published in core journals or monographs published by the most important publishers should be reported for evaluation. From the operational point of view, imposing an institution limit is advantageous because it limits the burden related to the acquisition and verification of metadata about publications.

Dimension 4: Researcher limit

When a researcher limit is not used, top-performing researchers are favored in the evaluation. Thanks to this, one unit of assessment (e.g. a faculty or a field within an institution) can be assessed very highly while this result is produced by a

few top-performing researchers and other academic staff-members provide a very small share of evaluated publications. Moreover, such top-performing researchers may not even have co-workers in their institutions.

Using the researcher limit causes all (or almost all) academic-staff member to need to provide some publications for the evaluation. Top-performing researchers might perceive this limit as an instrument, which depreciates their value for institutions. At the same time, for such researchers, this limit can encourage a top-performing researcher to enlarge the research groups in their institutions.

Imposing a researcher limit has an impact on the counting method, making it less important with the researcher limit. In Fig. 8, Spearman correlations across Variants 5–8 are high, which signifies that the ordering of the units of analysis is similar. A researcher limit does not have a strong impact on the number of reported publications.

Dimension 5: Point scale

Points attached to certain publication channels inform researchers what channels are preferred from a science policy point of view. Over time, there should be more publications in channels with a higher rating. Researchers pay more attention to thresholds and their relative difference measured in points than to the width of the scale, from 0 to 1 point or from 0 to 100 points. For instance, in the Polish system, a concomitance of the linear point scale with an institution limit caused a difference between thresholds in terms of points (e.g. a difference between 10-point publications 11-point publications) which could be very substantial. In Norway, from the perspective of a given institution, three 1-point publications could be equivalent to one 3-point publication, whereas in Poland there was a strict cut-point expressed by the institution limit of publication that could be assessed.

5. Conclusion

Our paper shows that selecting the publication counting method for national evaluation purposes needs to take into account the current situation in the given country in terms of the excellence of research outcomes, level of internal, external and international collaboration, and publication patterns in the various fields of science. We have shown how different variants of publication counting methods influence the rankings. We could construct other variants, but it will not make our task, i.e. selecting the proper way of counting, any easier, because there is no external and objective reference point.

In discussing the goals of any national research evaluation system, we should be aware that the ways of achieving these goals are closely related to publication counting methods. For instance, if our goal is to appreciate top-performing researchers in the evaluation, we should not implement a researcher limit. If our goal is to increase the level of international collaboration, we should use rather the square root fractional counting than complete or whole counting. Therefore, one can assess whether a publication counting method was properly selected, not by looking into the field rankings in the evaluation results, but rather by looking into the indicators (showing how incentives actually work), which reflect the goals of the research evaluation system. Publication counting methods have a higher impact on hard sciences than on social sciences and humanities.

Funding

This work was supported by the DIALOG Program (Grant name “Research into Excellence Patterns in Science and Art”) financed by the Ministry of Science and Higher Education in Poland.

Author contributions

Przemysław Korytkowski: Conceived and designed the analysis, Performed the analysis, Wrote the paper.

Emanuel Kulczycki: Conceived and designed the analysis, Wrote the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.joi.2019.07.001>.

References

- Aagaard, K. (2018). Performance-based research funding in Denmark: The adoption and translation of the norwegian model. *Journal of Data and Information Science*, 3(4), 19–29. <http://dx.doi.org/10.2478/jdis-2018-0018>
- Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian Publication Indicator. *Research Evaluation*, 24(2), 106–117. <http://dx.doi.org/10.1093/reseval/rvv003>
- Aagaard, K., & Schneider, J. W. (2015). Research funding and national academic performance: Examination of a Danish success story. *Science & Public Policy*, 1–14. <http://dx.doi.org/10.1093/scipol/scv058>
- Aksnes, D. W., Schneider, J. W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators: A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, 6(1), 36–43. <http://dx.doi.org/10.1016/j.joi.2011.08.002>
- Daraio, C., & Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. *Scientometrics*, 108(1), 391–400. <http://dx.doi.org/10.1007/s11192-016-1914-5>

- Engels, T. C. E., & Guns, R. (2018). The Flemish performance-based research funding system: A unique variant of the Norwegian model. *Journal of Data and Information Science*, 3(4), 45–60. <http://dx.doi.org/10.2478/jdis-2018-0020>
- Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. *Journal of Informetrics*, 11(3), 672–684. <http://dx.doi.org/10.1016/j.joi.2017.05.009>
- Gauffriau, M., Larsen, P. O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, 77(1), 147–176. <http://dx.doi.org/10.1007/s11192-007-1934-2>
- Hagen, N. T. (2014). Counting and comparing publication output with and without equalizing and inflationary bias. *Journal of Informetrics*, 8(2), 310–317. <http://dx.doi.org/10.1016/j.joi.2014.01.003>
- Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science*, 48(10), 944–945. <http://dx.doi.org/10.1163/00000086790082109>
- Huang, M.-H., Lin, C.-S., & Chen, D.-Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. *Journal of the American Society for Information Science and Technology*, 62(12), 2427–2436. <http://dx.doi.org/10.1002/asi.21625>
- Kulczycki, E. (2017). Assessing publications through a bibliometric indicator: The case of comprehensive evaluation of scientific units in Poland. *Research Evaluation*, 26(1), 41–52. <http://dx.doi.org/10.1093/reseval/rvw023>
- Kulczycki, E., Engels, T. C. E., Pölonen, J., Bruun, K., Dušková, M., Guns, R., . . . & Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. *Scientometrics*. <http://dx.doi.org/10.1007/s11192-018-2711-0>
- Kulczycki, E., Korzeń, M., & Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, 11(1), 282–298. <http://dx.doi.org/10.1016/j.joi.2017.01.001>
- Kulczycki, E., & Rozkosz, E. A. (2017). Does an expert-based evaluation allow us to go beyond the Impact Factor? Experiences from building a ranking of national journals in Poland. *Scientometrics*, 111(1), 417–442. <http://dx.doi.org/10.1007/s11192-017-2261-x>
- Larsen, P. O. (2008). The state of the art in publication counting. *Scientometrics*, 77(2), 235–251. <http://dx.doi.org/10.1007/s11192-007-1991-6>
- Organisation for Economic Co-Operation and Development. (2007). *Revised field of science and technology (FOS) classification in the Frascati manual DSTI/EAS/STP/NESTI(2006)19/FINAL* Retrieved from. <http://www.oecd.org/science/inno/38235147.pdf>
- Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science: A comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation*, 21(4), 280–290. <http://dx.doi.org/10.1093/reseval/rvs019>
- Piro, F. N., Aksnes, D. W., & Rørstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology*, 64(2), 307–320. <http://dx.doi.org/10.1002/asi.22746>
- Pölonen, J. (2018). Applications of, and experiences with, the Norwegian model in Finland. *Journal of Data and Information Science*, 3(4), 31–44. <http://dx.doi.org/10.2478/jdis-2018-0019>
- Prins, A. A. M., Costas, R., Van Leeuwen, T. N., & Wouters, P. F. (2016). Using google scholar in research evaluation of humanities and social science programs: A comparison with web of science data. *Research Evaluation*, 25(3), 264–270. <http://dx.doi.org/10.1093/reseval/rvv049>
- Research Excellence Framework. (2018). *REF 2021 Decisions on staff and outputs* (April).
- Sīle, L., Pölonen, J., Sivertsen, G., Guns, R., Engels, T. C. E. E., Arefiev, P., . . . & Teitelbaum, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation*, 1–13. <http://dx.doi.org/10.1093/reseval/rvy016> (June)
- Sivertsen, G. (2016). *A bibliometric indicator with a balanced representation of all fields*. In *International Conference on Science and Technology Indicators*. pp. 910–914. (September).
- Sivertsen, G. (2018). The Norwegian model in Norway. *Journal of Data and Information Science*, 3(4), 3–19. <http://dx.doi.org/10.2478/jdis-2018-0017>
- Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, 91(2), 567–575. <http://dx.doi.org/10.1007/s11192-011-0615-3>
- Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., . . . & 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. <http://dx.doi.org/10.1002/asi.22708>
- Waltman, L., & van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, 9(4), 872–894. <http://dx.doi.org/10.1016/j.joi.2015.08.001>
- Zacharewicz, T., Lepori, B., Reale, E., & Jonkers, K. (2018). Performance-based research funding in EU member states: A comparative assessment. *Science & Public Policy*, 1–11. <http://dx.doi.org/10.1093/scipol/scy041> (June)