Errors of measurement in scientometrics: Classification schemes and document types in citation and publication rankings

Nicolas Robinson-Garcia[1], Benjamín Vargas-Quesada[1], Daniel Torres-Salinas[1], Zaida Chinchilla-Rodríguez[2] and Juan Gorraiz[3]

[1] *elrobin@ugr.es; benjamin@ugr.es; torressalinas@ugr.es*
Unit for Computational Humanities and Social Sciences (U-CHASS), department of Information and Communication, University of Granada, Granada (Spain)

[1] *zaida.chinchilla@csic.es*
Consejo Superior de Investigaciones Científicas (CSIC), Instituto de Políticas y Bienes Públicos (IPP), Madrid, Spain

[3] *juan.gorraiz@univie.ac.at*
University of Vienna, Vienna University Library, Dept of Bibliometrics, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract

This research article delves into methodological challenges in scientometrics, focusing on errors stemming from the selection of classification schemes and document types. Employing two case studies, we examine the impact of these methodological choices on publication and citation rankings of institutions. We compute seven bibliometric indicators for over 8,434 institutions using 23 different classification schemes derived from Clarivate's InCites suite, as well as including all document types versus only citable items. Given the critical role university rankings play in research management and their methodological controversies, our goal is to propose a methodology that incorporates uncertainty levels when reporting bibliometric performance in professional practice. We then delve into differences in error estimates within research fields as well as between institutions from different geographic regions. The findings underscore the importance of responsible metric use in research evaluation, providing valuable insights for both bibliometricians and consumers of such data.

Keywords Responsible metrics; institutions rankings; citation indicators; publication counts; classifications of science; professional bibliometrics

Introduction

General Context

Errors constitute an inherent and inevitable aspect of the scientific process. Achieving perfect accuracy is unattainable, as tools for measurement will always include some level of uncertainty (Scuro 2004). According to the Joint Committee for Guides in Metrology (BIPM et al. 2008), uncertainty is defined as a 'parameter associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand' (p. 2), where the measurand refers to the object being measured.

In the field of scientometrics, uncertainty and errors in measurement are usually overlooked. This is problematic for various reasons. First, neglecting uncertainty leads to the misuse of bibliometric indicators, which are then employed to legitimize decisions in conditions of limited trust or political controversy (Ràfols et al. 2016). The professionalization of bibliometrics in academic libraries (Gorraiz et al. 2020; Gumpenberger et al. 2012) and Higher Education planning and research administration (Cox et al. 2019) has elevated bibliometric reporting to a valuable resource for decision-making. The widespread of metrics either explicitly or implicitly in research assessment exercises to support and judge individuals, departments or institutions (e.g., Hammarfelt and Rushforth 2017; Moed 2008) has created a landscape of tools and commercial solutions designed to respond to such demand. Bibliometric suites such as InCites (from Clarivate) or Scival (Elsevier) offer a battery of research indicators based respectively on bibliographic data from Web of Science and Scopus aimed at responding at this demand. These indicators are not only built based on different sets of publications, but their definition also differs leading sometimes to contradictory results for monitoring a common object (Robinson-Garcia et al. 2020).

Second, bibliometric indicators are usually reported with high levels of precision, being the more evident example the inclusion of up to three decimals of the Journal Impact Factor. This is particularly worrying given the fact that even Garfield himself considered the impact factor accurate only up to one decimal place (Bensman 2007). In later years he admitted that the only reason ISI calculated the impact factors reported in the JCRs out to three decimal places was to avoid the large number of ties that would have resulted in listing many journals alphabetically in the impact factor rankings. Any analysis carried out under such a premise would de facto lose their validity, and we should not forget that this error has a profound effect particularly at the lower frequencies on ordinal rankings by the impact factor, on which most journal evaluations are based (Schloegl and Gorraiz 2010).

This sense of false precision is also present in league tables and rankings in which variations in positions may be the result of noise rather than improvement or decay, affecting the prestige of those being portrayed in such tables (Bastedo and Bowman 2010; Gadd et al. 2021). While there have been some efforts to recognize this uncertainty, such as the inclusion of stability intervals in the Leiden Ranking and the introduction of position intervals in the Shanghai Ranking below a certain threshold, these measures are, at best, modest. Among the many causes for error or uncertainty in league tables we identify four types: 1) those derived from the misassignment of research outputs (Waltman et al. 2012), 2) those derived from unequal coverage of fields and locations in the database (Hicks 1999; Rafols et al. 2019; van Leeuwen et al. 2001), 3) those inherent to the metadata such as incompleteness or low quality metadata (Franceschini et al. 2015, 2016; Guerrero-Bote et al. 2021; Selivanova et al. 2019), and 4) those derived from methodological choices. By the latter we refer to choices which can affect the results without having *a priori* criteria set to justify such choices.

Objectives of the study

Our ultimate goal is to propose a methodological framework for evaluating measurements of error and variability when reporting bibliometric indicators in professional practice. This is particularly important if we aim to advocate and promote a responsible use of metrics in research evaluation in times when their use are more questioned than ever (Torres-Salinas et al. 2023). Both, bibliometricians and producers of bibliometric data and indicators, have the responsibility of bridging towards professionals and scientists (Leydesdorff et al. 2016) consuming this data and promoting good practices on the use of such metrics.

In this paper we showcase an specific case study in which errors can be quantified, derived from methodological choices. Specifically, we will focus on two case studies:

- **Diverging classification schemes.** Here errors arise from the selection of a given classification scheme over another one when producing field normalized indicators (Ruiz-Castillo and Waltman 2015). In this study we will focus on this latter aspect of indicator variability by computing the same set of indicators on the same set of publications using up to 23 different classification schemes. Differences are due to how publications are categorized differently according to each classification. Also, they can be due to the inclusion or exclusion of some of the records due to their unfitness regarding the classification scheme (e.g., the Essential Science Indicators classification scheme does not consider the fields of Arts and Humanities).

- **Selection of document types.** The second case study is derived from the definition and selection of document types included in each analysis (Moed and Van Leeuwen 1995). Here we will focus on estimating errors derived from including either all document types in an analysis or only citable items.

In both cases we will be using publication and citation rankings of institutions as a means to showcase how different indicators are affected by these errors and how it affects the positioning of institutions in these rankings. It is important to emphasize that the indicators used in our study, including those available through InCites, are part of a larger methodological framework focused on evaluating measurement errors and variability. This framework aims to improve the accuracy and reliability of bibliometric analyses, particularly when applied to institutional rankings, which use in in research management is specially controversial (Gadd 2020).

Structure of the paper

The paper is structured as follows. Next, we review literature related to methodological challenges imposed by the selection of classification schemes and definition of document types. Then we define our methodological design for measuring errors in scientometrics. Here we build on a previous study (Robinson-Garcia et al. 2023) to compute error estimates of 7 bibliometric indicators by using up to 23 different classification schemes. We use the InCites bibliometric suite to calculate these indicators at the institutional level, analyzing a total of 8,433 institutions. Third, we report average relative error estimates when considering different classification schemes at an aggregated level. We analyze differences in errors when accounting for all document types versus when considering only citable documents (articles, reviews and letters). Furthermore, we look into differences in errors when focusing on specific research fields as well as on geographic regions. We conclude by reporting the implications our findings have for professional practice and the use of scientometric indicators for decision-making.

Literature review

Selection of field classifications of science

The selection of an appropriate classification scheme is a matter of concern in the field of scientometrics that has been discussed extensively in the literature using both quantitative and qualitative data (Gómez et al. 1996; Janssens et al. 2009; Minguillo 2010; Perianes-Rodriguez and Ruiz-Castillo 2018; Shu et al. 2019; Sugimoto and Weingart 2015). It is problematic for various reasons. First, in relation to the level of analysis at which the field delineation is made. Here, Gómez et al. (1996) point at four potential levels: document level, journal level, affiliation level and author level. Determining the level at which field delineation is done is crucial in terms of interpretability (Robinson-Garcia and Calero-Medina 2014) and accuracy (Shu et al. 2019) as there is a problem of attribution especially when embedded in evaluation practices (Hansson et al. 2017). This issue is elegantly illustrated by Shu et al. (2019), who applied the Chinese Library Classification system to a set of publications using two different levels of analysis: at the journal level and the article level. They reported differences when rankings both: institutions and authors in terms of productivity, although this influence was mitigated at the institutional level.

Second, there are many classification systems available databases (Gusenbauer 2022), each data sources has its own delineation of fields with *ad hoc* proposals (Gómez-Núñez et al. 2014;

Muñoz-Écija et al. 2019). But these classifications are not standardized, dealing to contradicting and non-comparable results. This is an important point in bibliometric studies in order to allow comparisons. The evaluation of the same content illustrates the structural differences in the databases, which is informative for interpreting bibliometric analyses (Stahlschmidt and Stephen 2022). A comparative study between the subject categories classification system in Web of Science and the All Science Journal Classification (ASJC) in Scopus, indicated that both classification were "too lenient in assigning journals to categories" (Wang and Waltman 2016, p. 359). That is, they tended to include journals in multiple categories regardless how well connected they were in terms of citations with such categories. Another example is the classification proposed by Thijs et al. (2015), which combined hierarchical clustering and bibliographic coupling to create a 24 field classification system. This classification seemed to deviate greatly from the 22 fields from Essential Science Indicators included in Web of Science. These irregularities and discrepancies can lead to inconsistent outcomes (Reuven and Rosenfeld 2022) hindering the interpretation of bibliometric indicators. As a means to improve these classification systems, some authors have suggested the use of hybrid methods, that is, refine journal-level classifications with paper level citation clustering and text mining to improve these classifications (e.g., Janssens et al. 2009).

Third, there is the issue of balancing between accuracy, granularity and interpretability (Börner et al. 2012). In this sense, there are many multi-level classification systems which try to offer combinations by which users can zoom in or panning out. For instance, the publication level classification implemented by the CWTS (Waltman and van Eck 2012) is a three-tier classification system which goes spans from broad areas to micro topics based on direct citation networks between papers. Another example is the Australian and New Zealand Standard Research Classification (Australian Bureau of Statistics 2008) used by Dimensions, which also includes a three-level hierarchical classification system. The level of granularity of a classification will depend on the purpose for which the classification system is used. While broad fields may be desirable when reporting findings, more accuracy can lead to more robust indicators when normalizing by field (Ruiz-Castillo and Waltman 2015). In this sense, it is important to note that different indicators will show different levels of variability when moving from one classification system to another (Perianes-Rodriguez and Ruiz-Castillo 2018).

Definition of document types

The definition and inclusion of document types will have a vital influence on the outcome of a bibliometric report. Furthermore, their definition and typology will vary depending on the database used as a data source. The underlying principle behind this distinction of documents is that different types of documents serve different functions in the scientific system and hence are read and cited differently, leading to differences in citation distributions (Lundberg 2007; Moed and Van Leeuwen 1995). Here we observe inconsistencies between databases, finding that a paper categorized as 'Article' in Web of Science (WoS) might be defined as a 'Review' in Scopus, while Dimensions makes no distinction whatsoever (Visser et al. 2021).

Scientometric studies have historically distinguished between citable and non-citable items (Heneberg 2014). But databases often misclassify records, being letters and reviews the most affected by these inaccuracies (Donner 2017). Gorraiz and Schloegl (2008) reported that there was a difference of over 10% between the sum of articles and reviews reported in Web of Science and Scopus. On a different study, Haunschild and Bornmann (2022) compared the scores that result from different normalization procedures, which 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. They showed that field-normalized scores strongly depend on the choice of different document types and that the results on the aggregated level (country, institution) are not supported by results on the level of individual publications.

Data and methods

Data collection

Data was retrieved using the bibliometric suite InCites, which includes the Web of Science Core Collection including the Emerging Sources Citation Index (ESCI), for the period 1980-2022. In this study we worked at the institutional level, using their "Organizations" option, that is, their disambiguated list of institutions (more on issues for disambiguating institutional names in Waltman et al. 2012). We include a total of 8,434 institutions which have published at least 1,000 documents (all document types) in the study period. InCites provides for each unit under analysis a battery of bibliometric indicators which can be computed according to some customizable parameters. One of them being the election of a given classification system. We select 23 of the classification schemes currently available in InCites, including in here multi-level classification systems. Table 1 provides a description of each of them, indicating the number of categories per level and the assignment method Web of Science uses to create such tables. Except of the Citation Topics classification, which follow the methodology developed by Waltman and van Eck (2012), all schemes follow either a journal based method or aggregate Web of Science subject categories (which also are journal based).

Table 1. Description of the 23 classification schemes employed for the analysis*

Acronym	Denomination	Description	Levels and categories	Method
ANVUR	**ANVUR Category Schema**	Official academic fields and disciplines list for Italian Universities Research and Teaching.	**(1)** 17 broad categories	Category-to-category mapping (WoS)
FOR	**Australia ERA FOR**	Revised Australian and New Zealand Standard Research Classification (ANZSRC 2020)	**(1)** 24 FoR2 **(2)** 212 FoR4	Journal mapping
CAPES	**CAPES Brazil**	Classification created by the Foundation CAPES, linked to the Ministry of Education (Brazil)	**(1)** Capes 9 **(2)** Capes 49 **(3)** Capes 121	Category-to-category mapping (WoS)
CHINA	**China SCADC Subject Categories**	State Council Academic Degree Committee (SCADC) and Ministry of Education of China	**(1)** Broader 13 **(2)** Granular 96	Journal and other sources mapping
SHANGHAI	**Shanghai Ranking Global Ranking of Subjects**	Rankings of universities in 54 subjects across, Natural, , Life, Medical, and Social Sciences…	**(1)** 54 academic subjects	Category-to-category mapping (WoS)
TOPICS	**Citation Topics**	Algorithmically derived citation clusters (using an algorithm developed by CWTS, Leiden)	**(1)** Macro 10 **(2)** Meso 326	algorithmically on citation relationships
ESI	**Essential Science Indicators Research Areas**	All documents from Science Citation Index Expanded and Social Science Citation Index	**(1)** 22 broad categories	Journal mapping
FAPESP	**FAPESP Brazil**	Created by the São Paulo Research Foundation	**(1)** 9 High Level **(2)** 72 Detailed categories	Category-to-category mapping (WoS)
GIPP	**Institutional Profiles Research Areas**	Clarivate Analytics has been profiling the world's leading universities and research institutions	**(1)** 6 broad academic fields	Category-to-category mapping (WoS)
KAKEN	**KAKEN Category Schema (10 and 66)**	From Japan called the Kakenhi Program (Grants-in-Aid for Scientific Research).	**(1)** 10 L2 **(2)** 66 L3	Category-to-category mapping (WoS)
OECD	**OECD Category Schema**	Revised Field of Science and Technology (FOS) Classification of the Frascati Manual.	**(1)** 42 fields	Category-to-category mapping (WoS)
PL19	**PL19 Category Schema**	The Polish PL19 category schema is used for annual evaluation exercise	**(1)** 44 underlying categories	Journal mapping
RIS	**Research and Innovation Strategies for Specialization**	The Research and Innovation Strategies for Smart Specialization (RIS3) for Latvia	**(1)** 7 specialization fields	Category-to-category mapping (WoS)
UKREF14	**UK RAE Units of Assessment 2018**	UK 2014 Research Assessment Exercise (RAE) Units of Assessment (UoA)	**(1)** 36 categories	Category-to-category mapping (WoS)

UKREF21	**UK REF Units of Assessment 2021**	UK 2021 Research Assessment Exercise (RAE) Units of Assessment (UoA)	**(1)** 36 units of assessment	Category-to-category mapping (WoS)
WOS	**Web of Science Research Areas**	The Web of Science schema comprises approximately 250 subject areas	**(1)** 250	---
* Information extracted from https://incites.help.clarivate.com/Content/Research-Areas/research-areas.htm The data presented reproduce higher levels of categorization. Specifically, certain categories (FOR, CAPES, CHINA, TOPICS, FAPESP, KAKEN) encompass multiple levels, which are not detailed in this table. The total of 16 listed categories, when including the sublevels of the mentioned categories, sums up to the 23 classification schemes analysed in this paper.				

For each institution we focused on seven different indicators: total number of publications, times cited, the Category Normalized Citation Indicator, top 1% most cited papers, top 10% most cited papers, average percentile, and H-Index. Each indicator is computed 46 times; according to the 23 different classification schemes and including all document types versus only citable items.

Calculation of errors

The calculation of the errors followed the standard definition used in the experimental sciences, where the absolute error of a measurement is the difference between the measured value and the true value. In cases where the true value is unknown, it is replaced with the mean value obtained after multiple iterations of the measurement. In our case, each indicator is calculated a total of 23 times for each classification scheme and twice when looking into document types.

To accurately determine the absolute error for each institution, we first calculate the mean value of the repeated measurements of each indicator. The absolute error for a given institution, Δx_i is then computed as the difference between its measured value and the reference mean value. Mathematically, the absolute error of an institution can be defined as:

$$\Delta x_{mean} = (ABS(\Delta x_1) + ABS(\Delta x_2) + \ldots.. + ABS(\Delta x_n))/n$$

Where x_i corresponds to an institution and $ABS(\Delta x_i)$ is its absolute error.

Let's consider we want to compute the absolute error after retrieving the number of publications produced by three departments from three different bibliometric databases. The measurements obtained for from each database are shown in Table 2.

Table 2. Mock up example of number of publications obtained for three institutions from three different databases

	Database I	Database II	Database III
Department A	250	260	240
Department B	300	310	290
Department C	280	270	290

For each research department, we first calculate the mean number of publications across the three databases:

- Department A: 250
- Department B: 300
- Department C: 280

Next, we calculate the absolute error for each database by comparing the measured values with the mean value for each department and dividing by the number of measurements. Hence, the absolute error for Department A would be:

$$\Delta x_{\text{mean}} = \frac{|250 - 250| + |260 - 250| + |240 - 250|}{3} = 6.67$$

Following the calculation of absolute errors, we also compute relative and percentage errors, facilitating comparisons between indicators across different classification schemes. The relative error is obtained by dividing the absolute error by the mean value, while the percentage error is calculated by multiplying the relative error by 100. Following the example of Department A, we now show its relative and percentage errors:

$$Relative\ Error = \frac{6.67}{250} \approx 0.0267$$

$$Percentage\ Error = 0.0267 \times 100 \approx 2.67\%$$

To account for the variability observed in repeated measurements, we calculated the confidence intervals associated with each error measurement. These intervals were derived from the standard deviation of the repeated measurements and provide a range within which the true value is expected to lie with a specified level of confidence.

Results

General overview of error estimates by classification scheme and document types

In table 3 we include an overview of our final dataset. For each classification scheme we include the total number of institutions covered when considering all document types and when considering only citable documents. As observed, only the WOS classification includes the 8,434 institutions originally included in our dataset when considering all document types. Simply by filtering to citable documents, we lose up to 559 institutions in the best of cases (WOS). The classification scheme with the lowest coverage is FOR1 which includes 42% of the institutions and 18% of documents, a share that increases up to 20% when considering only citable documents.

Table 3. Total institutions and publications per classification scheme and document types

Classification scheme	Institutions		Records	
	All docs.	Citable only	All docs.	Citable only
WOS	8,434	7,875	63,908,002	51,399,082
KAKENL2	8,433	7,874	63,895,799	51,389,373
KAKENL3	8,432	7,874	63,895,799	51,389,373
UKREF21	8,432	7,874	63,895,799	51,389,373
RIS	8,432	7,874	63,895,799	51,389,373
OCDE	8,432	7,874	63,895,799	51,389,373
GIPP	8,432	7,874	63,895,799	51,389,373
FAPESP	8,432	7,874	63,895,799	51,389,373
CAPES9	8,432	7,874	63,895,799	51,389,373
CAPES49	8,432	7,874	63,895,799	51,389,373
ANVUR	8,431	7,874	63,888,303	51,389,373
UKREF14	8,414	7,865	63,671,768	51,241,279
CAPES121	8,380	7,837	63,011,376	50,799,943

SHANGHAI	8,231	7,737	59,830,551	48,881,578
CHINA BROAD	8,141	7,610	58,977,093	47,077,808
CHINA NARROW	8,141	7,610	58,977,087	47,077,802
TOPICSMACRO	7,931	7,717	53,728,425	49,126,966
TOPICSMESO	7,931	7,717	53,728,425	49,126,966
TOPICSMICRO	7,931	7,717	53,728,425	49,126,966
PL19	7,746	7,187	51,317,825	40,102,542
ESI	7,497	6,943	49,844,239	39,158,270
FOR2	6,560	5,959	36,711,808	27,927,846
FOR1	3,549	3,348	11,671,816	10,107,716

Next, we report the average estimated error for each classification resulting from considering either all document types or only citable documents (Table 4). As observed, the lowest relative error is reported for the H-Index and the total number of citations, in both case below 1% for all classification schemes. We observe error estimates ranging between 0.7 for the CNCI to 5.3% for the average percentile indicators. On the other extreme we find that number of documents is the indicator with the largest estimated error (7.4% on average). Furthermore, we observe different levels of variability in the error by indicator. While in the cases of times cited and H-Index these are below 0.1%, in the case of top 1% highly cited papers, we observe a variability of 8.3 between the largest estimated error (11% for TOPICSMICRO) and the lowest value (2.7% for TOPICSMACRO). A similar case we observe again in the number of publications, where there is a variability of up to 6.1% between FOR2 (9.6%) and the three TOPICS schemes (3.5%).

Table 4. Average relative error by classification scheme considering document type

Classification scheme	**Docs**	**+/-**	**Times Cited**	**+/-**	**CNCI**	**+/-**	**Top 1%**	**+/-**	**Top 10%**	**+/-**	**Avg. percentile**	**+/-**	**H-Index**	**+/-**
WOS	8.0	6.3	0.8	0.6	2.2	2.0	4.5	3.7	2.7	2.3	4.3	3.6	0.3	0.3
KAKENL2	8.0	6.3	0.8	0.6	3.0	3.0	4.9	4.1	2.7	2.3	4.2	3.6	0.3	0.3
KAKENL3	8.0	6. 3	0.8	0.6	2.8	2.7	4.5	3.8	2.7	2.3	4.2	3.6	0.3	0.3
UKREF21	8.0	6. 3	0.8	0.6	2.7	2.6	4.7	3.9	2.9	2.4	4.3	3.6	0.3	0.3
RIS	8.0	6. 3	0.8	0.6	3.3	3.2	5.0	4.3	2.7	2.2	4.2	3.6	0.3	0.3
OCDE	8.0	6. 3	0.8	0.6	2.8	2.7	4.8	4.0	2.9	2.4	4.3	3.6	0.3	0.3
GIPP	8.0	6. 3	0.8	0.6	3.0	2.9	5.0	4.2	2.9	2.3	4.4	3.7	0.3	0.3
FAPESP	8.0	6.3	0.8	0.6	2.9	2.8	4.8	3.9	2.8	2.4	4.3	3.6	0.3	0.3
CAPES9	8.0	6.3	0.8	0.6	3.0	2.9	5.1	4.3	2.9	2.4	4.3	3.6	0.3	0.3
CAPES49	8.0	6.3	0.8	0.6	2.7	2.6	4.6	3.9	2.7	2.3	4.3	3.6	0.3	0.3
ANWUR	8.0	6.3	0.8	0.6	3.0	2.9	4.8	4.0	2.9	2.5	4.3	3.6	0.3	0.3
UKREF14	8.0	6.3	0.8	0.6	3.0	2.6	4.6	3.9	2.9	2.4	4.3	3.6	0.3	0.3
CAPES121	7.9	6.3	0.8	0.6	3.0	2.1	4.5	3.7	2.7	2.3	4.3	3.6	0.3	0.3
SHANGHAI	7.7	6.0	0.8	0.6	3.0	2.5	4.5	3.8	2.7	2.2	3.9	3.3	0.3	0.3
CHINA BROAD	8.2	6.3	0.8	0.6	3.0	2.9	4.9	4.2	2.9	2.4	4.4	3.7	0.3	0.3
CHINA NARROW	8.2	6.3	0.8	0.6	2.7	2.6	4.6	3.7	3.1	2.6	4.5	3.7	0.3	0.3
TOPICSMACRO	3.5	2.6	0.8	0.6	1.0	0.9	2.7	2.4	1.3	1.0	0.7	0.5	0.3	0.3
TOPICSMESO	3.5	2.6	0.8	0.6	0.8	0.7	4.3	3.5	1.2	1.0	0.7	0.5	0.3	0.3
TOPICSMICRO	3.5	2.6	0.8	0.6	0.7	0.6	11.0	6.2	1.6	1.1	0.6	0.5	0.3	0.3
PL19	8.6	6.6	0.8	0.6	3.4	3.4	5.1	4.3	2.8	2.3	4.6	3.8	0.3	0.3

ESI	8.7	6.5	0.8	0.6	3.0	2.8	4.8	4.0	3.0	2.4	4.5	3.7	0.3	0.3
FOR2	9.6	7.0	0.8	0.6	2.8	2.7	4.8	3.9	2.8	2.2	5.3	4.0	0.3	0.3
FOR1	4.7	3.4	0.8	0.6	1.6	1.4	3.5	2.8	1.7	1.3	2.0	1.7	0.3	0.3

Given differences on the institutional coverage observed in Table 3, we present error estimates only considering institutions which are present in all schemes. To show how the exclusion of schemes with lower coverage affect error estimates, we present the world average relative error when considering the 23 classification schemes and considering only the top 13 with the largest institutional coverage (Table 5). Here we observe relatively small differences between focusing on all document types or citable documents. When considering the 23 classification schemes, we observe that the largest estimated errors are found for number of documents (12.2-12.8%) and top 1% most highly cited papers (10.0-11.2%). When reducing the number of classifications, we find how the error estimates are reduced drastically for size dependent indicators (docs, times cited and H-Index) while they are very similar for non-size dependent indicators (CNCI, Top 1%, Top 10% and avg. percentile). However, while in the case of the CNCI and the average percentile, the error is reduced for 13 classification schemes, it increases for the Top 1% and Top 10% indicators.

Table 5. World Percentage errors considering 23 and 13 classification schemes by document type selection

	23 schemes		13 schemes	
Indicators	**All docs**	**Citable only**	**All docs**	**Citable only**
Docs	12.8	12.2	0.2	0.2
+/-	1.9	1.9	0.2	0.2
Times Cited	9.2	9.2	0.2	0.2
+/-	0.4	0.4	0.2	0.2
CNCI	4.8	4.6	3.6	3.4
+/-	1.7	1.4	1.8	1.5
Top 1%	11.2	10.0	10.0	11.8
+/-	4.1	3.9	4.8	6.6
Top 10%	5.8	5.8	5.3	6.0
+/-	1.8	1.7	1.8	2.2
Avg. Percentile	3.1	2.4	2.2	2.3
+/-	1.1	0.7	0.7	0.7
H-Index	4.1	4.1	0.1	0.1
+/-	0.4	0.4	0.1	0.1

Field differences in error estimates for document types: Macro Topic and ESI fields

As a means to deepen on how the choice of including all document types or only those defined as citable, in tables 6 and 7 we look into differences in error estimates by field. To do so, we use TOPICSMACRO scheme formed by 10 major fields and the 22 ESI fields respectively.

In the case of the macro topics, we observe that the largest errors can be found in the fields of Arts & Humanities (between 0.6% for H-Index and 8.1% for number of documents), followed by Clinical & Life Sciences (between 0.4% for the H-Index and 6.0% for number of documents). Interestingly, the greatest variability in percentage error is found in the number of documents, for which Arts & Humanities is the field with the largest average error estimate, while Electrical Engineering, Electronics & Computer Science has an average percentage error of 1%. The rest of the patterns between indicators hold to what we observed in Table 4.

By using the citation topics, we considered that their calculation is only possible for publications with cited references. Therefore, all the document types usually non-including cited references (like e.g., meeting abstracts) are automatically excluded. That is the reason why the differences between using all document types and only citable items will be lower than expected when considering other classifications elaborated on journal level.

Table 6. Percentage errors from document type selection by Macro Topics

TOPICSMACRO	Docs	+/-	Times Cited	+/-	CNCI	+/-	Top 1%	+/-	Top 10%	+/-	Avg. percentile	+/-	H-Index	+/-
Agriculture, Environment & Ecology	2.3	1.1	1.0	0.5	0.7	0.5	2.1	1.7	0.9	0.7	0.4	0.3	0.297	0.3
Arts & Humanities	8.1	2.2	3.0	0.8	1.7	1.0	5.2	3.8	2.5	1.2	2.8	1.0	0.6	0.7
Chemistry	1.9	1.1	1.0	0.6	0.5	0.4	1.8	1.7	0.7	0.6	0.3	0.2	0.2	0.3
Clinical & Life Sciences	6.0	2.8	1.4	0.6	1.6	1.3	3.5	2.7	1.7	1.2	1.0	0.6	0.4	0.3
Earth Sciences	2.3	1.0	0.7	0.3	0.6	0.4	1.8	1.4	0.7	0.4	0.4	0.2	0.2	0.2
Electrical Engineering, Electronics & Computer Science	1.0	0.5	0.6	0.4	0.4	0.3	1.2	1.2	0.4	0.3	0.1	0.1	0.2	0.3
Engineering & Materials Science	1.0	0.6	0.4	0.3	0.4	0.3	1.6	1.5	0.6	0.5	0.2	0.1	0.1	0.2
Mathematics	1.3	0.7	0.7	0.5	0.5	0.4	1.6	1.4	0.5	0.4	0.2	0.2	0.3	0.4
Physics	1.4	0.6	0.5	0.3	0.5	0.3	1.3	1.1	0.5	0.4	0.2	0.2	0.2	0.2
Social Sciences	3.4	1.6	1.3	0.6	1.0	0.8	2.4	1.9	0.9	0.6	0.7	0.4	0.4	0.4

In the case of the ESI fields (Table 7) a different pattern is observed. Here it is the Multidisciplinary category the one accounting for the largest errors in all indicators with notable differences with respect to the rest of the categories. The exception is found in the average percentile with other fields exhibit greater percentage errors (e.g., Clinical Medicine, Social Sciences, general). The other exhibiting a large, estimated error is Clinical Medicine, where the error in terms of number of documents is just above 20%.

Table 7. Percentage errors from document type selection by ESI fields

ESI	Docs	+/-	Times Cited	+/-	CNCI	+/-	Top 1%	+/-	Top 10%	+/-	Avg. percentile	+/-	H-Index	+/-
Agricultural Sciences	3.5	2.7	0.4	0.3	1.8	1.5	3.0	2.3	1.3	1.0	1.8	1.7	0.1	0.2
Biology & Biochemistry	11.5	5.5	1.3	0.610	3.0	2.2	6.5	4.7	3.9	2.9	6.9	3.8	0.4	0.3
Chemistry	5.4	4.6	0.7	0.6	3.2	3.1	5.5	5.4	3.3	3.6	3.8	4.0	0.2	0.3
Clinical Medicine	20.5	5.5	1.8	0.6	6.5	4.6	8.0	4.7	6.5	3.5	10.6	3.2	0.4	0.3
Computer Science	3.1	1.2	0.8	0.5	1.4	1.1	3.0	2.2	1.1	0.8	0.8	0.4	0.3	0.3
Economics & Business	6.6	2.9	1.2	0.5	3.0	2.3	4.8	3.7	1.8	1.1	3.5	2.0	0.4	0.3
Engineering	1.9	1.1	0.5	0.3	1.7	1.5	2.9	2.6	1.3	1.0	0.5	0.4	0.1	0.2
Environment/Ecology	2.2	0.9	1.1	0.5	1.1	0.7	2.1	1.5	1.0	0.7	0.3	0.2	0.5	0.4
Geosciences	3.5	1.4	0.7	0.3	1.2	0.9	3.2	2.8	1.1	0.9	1.1	0.7	0.2	0.3
Immunology	13.7	3.1	1.8	0.5	3.1	2.1	5.5	3.4	4.6	2.3	6.9	2.2	0.5	0.3
Materials Science	1.1	0.6	0.3	0.2	1.1	1.0	2.3	2.4	0.8	0.6	0.3	0.2	0.1	0.1
Mathematics	1.3	0.6	0.3	0.2	0.6	0.5	1.9	1.6	0.6	0.4	0.5	0.3	0.1	0.2
Microbiology	4.1	1.3	1.5	0.7	1.1	0.8	3.1	2.3	1.3	0.8	1.2	0.7	0.4	0.3

Molecular Biology & Genetics	9.1	2.8	0.9	0.3	2.8	1.7	5.0	2.4	3.9	1.8	5.3	1.9	0.3	0.2
Multidisciplinary	23.0	22.6	3.1	1.9	12.6	7.0	19.9	15.1	13.1	9.4	8.8	9.8	1.1	0.8
Neuroscience & Behavior	14.6	3.8	1.2	0.4	3.4	2.5	5.4	3.3	3.6	3.0	8.4	2.4	0.2	0.2
Pharmacology & Toxicology	11.8	4.5	1.6	0.7	2.8	2.0	5.2	3.5	4.1	2.4	6.4	2.8	0.4	0.4
Physics	1.6	0.7	0.7	0.4	0.6	0.4	1.5	1.1	0.6	0.4	0.3	0.2	0.2	0.2
Plant & Animal Sciences	5.2	2.8	1.0	0.4	2.0	1.4	4.2	3.4	2.0	1.6	2.2	1.8	0.2	0.3
Psychiatry/Psychology	12.5	2.8	1.3	0.5	3.3	2.4	4.8	3.2	3.4	1.7	7.5	1.9	0.3	0.3
Social Sciences, general	15.1	5.6	1.5	0.5	4.1	3.0	5.8	3.9	4.1	2.4	9.8	4.6	0.4	0.3
Space Science	1.3	0.4	0.4	0.2	0.5	0.4	1.2	1.0	0.5	0.4	0.3	0.1	0.1	0.1

Regional differences in error estimates for classification schemes: United States vs. South America

Finally, we delve into regional differences in order to understand how homogeneous or heterogeneous is the effect of using different classification schemes in institutions located in different parts of the world. As an illustrative example, in Table 8 we report the average percentage errors of institutions located in the United States and in South America. Again, we report errors considering all document types or citable documents, as well as including the 23 classification schemes or only the 13 classifications with the largest institutional coverage. While the general pattern of errors is similar to that observed in table 5, we do observe differences between institutions of these two regions. Overall, we observe that differences of error are always lower than 1% with some exceptions. The largest difference is that observed for the Top 1% most highly cited publications, where differences of error are above 4% (favoring US institutions) when considering the 23 classification schemes. These differences are below 1% when considering only 13 schemes. The other exception is for all document types and Top 10%, where the difference of error is just above 1%, being larger for South American institutions.

Table 8. Percentage errors from classification schemes for institutions located in the United States vs. institutions located in South America

	UNITED STATES				**SOUTH AMERICA**			
	23 schemes		**13 schemes**		**23 schemes**		**13 schemes**	
Indicators	**All docs**	**Citable only**	**All docs**	**Citable only**	**All docs**	**Citable only**	**All docs**	**Citable only**
Docs	13.8	11.8	0.2	0.3	12.4	12.6	0.1	0.1
+/-	1.6	1.5	0.2	0.2	1.0	1.3	0.1	0.1
Times Cited	9.3	9.3	0.3	0.3	9.3	9.2	0.2	0.2
+/-	0.3	0.3	0.2	0.2	0.3	0.3	0.1	0.1
CNCI	5.4	4.9	4.5	3.6	5.1	5.0	2.0	2.0
+/-	1.9	1.4	2.3	1.6	1.2	1.4	0.8	0.6
Top 1%	8.6	7.6	9.2	8.8	13.2	12.2	9.3	9.7
+/-	2.7	2.4	4.3	4.1	3.6	3.4	3.8	4.1
Top 10%	5.0	5.2	5.1	5.2	6.0	6.0	5.0	5.3
+/-	1.2	1.3	1.9	1.8	1.2	1.3	1.4	1.5
Avg. Percentile	4.1	2.0	1.9	2.0	3.2	2.9	2.4	2.5
+/-	1.1	0.5	0.5	0.5	0.8	0.6	0.5	0.5

H-Index	4.3	4.3	0.1	0.1	4.0	4.0	0.1	0.1
+/-	0.3	0.3	0.1	0.1	0.3	0.3	0.1	0.1

Conclusions & Discussion

In bibliometric practices as well as in many bibliometric studies, several choices always have to be made which naturally can seriously affect the results obtained and their interpretations. In this article, we have assessed the magnitude of these errors occurring in two recurring situations: 1) considering the document type and 2) considering different classification schemes. Moreover, we think that a science that is mainly based on statistics should indicate the validity of the retrieved values, both of its numbers and of its decimals, in order to reveal their significance. The purpose of this paper is to reveal bibliometricians which errors will be derived from their document type and classification schemes decisions.

In the first case, we have seen how the choice of what type of documents should be included in the analysis, can influence not only the indicators of publication activity, where they are manifest, but also those of impact and especially those of normalized impact, such as the CNCI, and the Top 1% and Top 10%, which are the most commonly used in bibliometric practices (Moed 2017).

Generally, in this case, the decision falls between choosing the "citable publications" or all types of documents that are available. This difference was already introduced by Garfield, when he introduced his measure of the Journal Impact Factor by considering only the citable items (articles, reviews and proceedings) in the denominator, while in the numerator including the citations to all document types.

Our study shows that this decision can severely distort the results in the Essential Science Indicators (ESI) Category "Multidisciplinary". As it is well known, this category appears as well in the ESI classification scheme as in the Journal Citation Reports, while in InCites all these publications are reallocated to categories that are more precise. Anyhow, our results show the big differences bibliometricians may confront with dealing with journals and journal impact measures assigned to this category.

Other categories are also affected by the document type decision, especially *Clinical Medicine, Immunology* and *Pharmacology & Toxicology*, where percentage errors higher than 4% are reported for the calculation of the top 10% most cited, that is commonly used as a measure for academic excellence. These are mainly due to the effect of the *Meeting Abstracts*, as it has already been often reported (Gorraiz et al. 2016). This also corroborated by the results obtained when using the Citation Topics – Macrotopics. In this case, the percentage errors are considerably diminished, because Meeting Abstracts lack of references and cannot be considered in this scheme. *Editorial Materials* and *Book Chapters* are the other document types responsible for the differences in the impact measures, also in other categories like *Social Sciences, general*, *Biology & Biochemistry* and *Psychiatry/Psychology*. Interestingly. *Physics*, *Space Physics* and *Mathematics* are almost not affected, and the only consideration of citable items is a sound decision.

Therefore, in our analyses we are considering errors due to three different decision-making processes: 1) to select a classification scheme, 2) to select a classification schema based on journal level versus one based on document level; and 3) to use a classification schema on different aggregation levels. And for the bibliometrician community it is crucial to be able to estimate the effect that these decisions are expected to have on the values calculated for their impact measures.

In this our first error analysis study, we have considered all the classification schemes available nowadays in InCites and calculated the errors due to their selection. In a first instance, we wanted to extend our work to the classifications available in SciVal and compare our results. But to our great surprise, we discovered that the selection of the classification did not change the impact results at the working aggregation level (meso-level) in that analytical tool, so they could not be incorporated into our study.

Our results show the alterations that the impact measures would undergo in the case of using another classification scheme. Of the four normalized impact indicators, the top 1% shows the largest margin of error ranging from 3 to 17%. This is mainly due to the short number of publications contributing to this percentile. The measure of excellence (Top 10% most cited) fluctuates between 5 and 8%, and the average percentile varying between approximately 1.5 and 3%. The value of the CNCI can vary from 1.5 to 5%.

Differences in the values of the H-index are only reported when including classification schemes reducing considerably the number of publications considered (10 from 23 Schemes). Furthermore, topological factors, would only increase slightly these errors or deviations due to the selection of a classification scheme. In a study case, the differences between North-American and South-American institutions, the discrepancies between the percentage errors were only slightly higher for the percentile indicators for the South-American organizations when considering only the citable items. Finally, our results show that the decision made by for using all types of documents or only the citable ones hardly alter the divergences resulting from the use of one or the other classification system (see table 5) at the meso level.

Limitations and further research

All the analyses in this study have been carried out at the meso-level, which is the most significant for this purpose. All analyses have been performed for the period 1980-2022 to increase the significance of the results. For shorter periods, the errors, especially for indicators that are not standardized, such as the number of publications and the number of citations, would naturally be much higher and should therefore be recalculated for each specific situation in further studies

At the macro level very similar results are expected, while at the micro level, i.e., in the evaluation of individuals, the analysis is much more problematic due to the small number of publications available, and the great diversity of the cases and criteria to be considered (such as gender, age of career, etc.). For example, Åström, Hammarfelt and Hansson (2017) discuss how scientific publications can be categorized in different fields depending on the unit of assessment being evaluated: the publication, the individual or the institution. They found variations in terms of purpose of categorization as well as purpose of evaluation, i.e., the definition and function of the publications depending on whether it is situated in a context of scholarly communication or a context of research evaluation. The raising questions such as on what levels the distinctions are made, and in terms of on what principles the categories are being defined. The varying functions of the boundary object becomes critical when contextualized within the concept of infrastructures (publication databases, citation indices, evaluation systems and classification systems). Therefore, it is always advisable to perform the measurements on a case-by-case basis and with different data sources and purpose of use.

One of the most subtle and critical problems of bibliometrics is classifications. As is well known, there is no standard, and each database and even each nation or continent uses its own schema for their evaluation systems. That is why Clarivate, in a big effort, tried to collect the most used ones on different aggregation levels (macro, meso and micro) in its flagship product "InCites" and to use them for the calculation of the most common bibliometric indicators (see Table 1). Besides, classification systems are usually created at the level of journals (Pudovkin & Garfield, 2002) but also at the paper level (Waltman & van Eck 2012; Rivest et al 2021). Comparisons of these two levels of aggregations, journal classification versus paper classification using the same classification scheme and the same dataset revealed that almost half of the papers could be misclassified in journal classification systems (Shu et al 2019). When comparing rankings of the most productive institutions and authors, classification of papers has less influence on rankings at the institutional level than at the individual level (Shu et al 2020), which has implications for bibliometric evaluation. At this point, it is important to emphasize that InCites also includes other classifications based not only on journal level (e.g. ESI and WoS subject categorization) but also most recent ones based on document level like the classification system builds on "Citation Topics", algorithmically derived citation clusters using an algorithm developed by CWTS, Leiden[1]. Further studies can bring light on differences using journal or article level when calculating errors.

Another limitation of this study is that we have performed all the analyses on a single data source, the Web of Science Core Collection. Therefore, it will be necessary to perform future analyses comparing the results in different sources, such as WoS CC, Scopus, Dimensions and even other Open sources, such as OpenAlex, Crossref or Lens. We are fully convinced that these studies will be of great help to all those involved in providing bibliometric services to be able to argue, justify and foresee the effects of the decisions they have had to make in carrying out their analyses.

Conflict of interest: Nicolas Robinson-Garcia is Associate Editor of Scientometrics.

Acknowledgments: The authors acknowledge funding from the Spanish Ministry of Science (COMPARE project PID2020-117007RA-I00; and RESPONSIBLE project PID2021-128429NB-I00). Nicolas Robinson-Garcia is funded by a Ramón y Cajal grant from the Spanish Ministry of Science and Innovation (REF: RYC2019-027886-I).

References

Åström, F., Hammarfelt, B., & Hansson, J. (2017). Scientific publications as boundary objects: theorising the intersection of classification and research evaluation. *Information Research, 22*(1), CoLIS paper 1623. http://InformationR.net/ir/22-1/colis/colis1623.html

Australian Bureau of Statistics. (2008, March 31). Australian and New Zealand Standard Research Classification (ANZSRC). c=AU; o=Commonwealth of Australia; ou=Australian Bureau of Statistics. https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1297.02008?OpenDocument. Accessed 30 November 2023

Bastedo, M. N., & Bowman, N. A. (2010). U.S. News & World Report College Rankings: Modeling Institutional Effects on Organizational Reputation. *American Journal of Education, 116*(2), 163–183. https://doi.org/10.1086/649436

[1] Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

Bensman, S. J. (2007). Garfield and the impact factor. *Annual Review of Information Science and Technology*, *41*(1), 93–155. https://doi.org/10.1002/aris.2007.1440410110
BIPM, IEC, IFCC, ILAC, ISO, IUPAC, et al. (2008). Evaluation of measurement data — Supplement 2 to the "Guide to the expression of uncertainty in measurement" — Extension to any number of output quantities. https://www.bipm.org/documents/20126/2071204/JCGM_102_2011_E.pdf/6a3281aa-1397-d703-d7a1-a8d58c9bf2a5
Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and Update of a Classification System: The UCSD Map of Science. *PLOS ONE*, *7*(7), e39464. https://doi.org/10.1371/journal.pone.0039464
Cox, A., Gadd, E., Petersohn, S., & Sbaffi, L. (2019). Competencies for bibliometrics. *Journal of Librarianship and Information Science*, *51*(3), 746–762. https://doi.org/10.1177/0961000617728111
Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, *113*(1), 219–236. https://doi.org/10.1007/s11192-017-2483-y
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. *Scientometrics*, *102*(3), 2181–2186. https://doi.org/10.1007/s11192-014-1503-4
Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, *10*(4), 933–953. https://doi.org/10.1016/j.joi.2016.07.003
Gadd, E. (2020). University rankings need a rethink. *Nature*, *587*(7835), 523–523. https://doi.org/10.1038/d41586-020-03312-2
Gadd, E., Holmes, R., & Shearer, J. (2021). Developing a Method for Evaluating Global University Rankings, *3*(1), 2. https://doi.org/10.29024/sar.31
Gómez, I., Bordons, M., Fernández, M. T., & Méndez, A. (1996). Coping with the problem of subject classification diversity. *Scientometrics*, *35*(2), 223–235. https://doi.org/10.1007/BF02018480
Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369–383. https://doi.org/10.1016/j.joi.2014.01.011
Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10/f8d2bp
Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science*, *34*(5), 715–725. https://doi.org/10.1177/0165551507086991
Gorraiz, J., Wieland, M., Ulrych, U., & Gumpenberger, C. (2020). De Profundis: A Decade of Bibliometric Services Under Scrutiny. In C. Daraio & W. Glänzel (Eds.), *Evaluative Informetrics: The Art of Metrics-Based Research Assessment : Festschrift in Honour of Henk F. Moed* (pp. 233–260). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-47665-6_11
Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative Analysis of the Bibliographic Data Sources Dimensions and Scopus: An Approach at the Country and Institutional Levels. *Frontiers in Research Metrics and Analytics*, *5*. https://www.frontiersin.org/articles/10.3389/frma.2020.593494. Accessed 5 December 2023

Gumpenberger, C., Wieland, M., & Gorraiz, J. (2012). Bibliometric practices and activities at the University of Vienna. *Library Management*, *33*(3), 174–183. https://doi.org/10.1108/01435121211217199

Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, *127*(5), 2683–2745. https://doi.org/10.1007/s11192-022-04289-7

Hammarfelt, B., & Rushforth, A. D. (2017). Indicators as judgment devices: An empirical study of citizen bibliometrics in research evaluation. *Research Evaluation*, *26*(3), 169–180. https://doi.org/10.1093/reseval/rvx018

Haunschild, R., & Bornmann, L. (2022). 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? *Scientometrics*, *127*(8), 4419–4438. https://doi.org/10.1007/s11192-022-04446-y

Heneberg, P. (2014). Parallel worlds of citable documents and others: Inflated commissioned opinion articles enhance scientometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 635–643. https://doi.org/10.1002/asi.22997

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193–215.

Janssens, F., Zhang, L., Moor, B. D., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702. https://doi.org/10.1016/j.ipm.2009.06.003

Leydesdorff, L., Wouters, P., & Bornmann, L. (2016). Professional and citizen bibliometrics: complementarities and ambivalences in the development and use of indicators—a state-of-the-art report. *Scientometrics*, *109*(3), 2129–2150. https://doi.org/10.1007/s11192-016-2150-8

Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics*, *1*(2), 145–154. https://doi.org/10.1016/j.joi.2006.09.007

Minguillo, D. (2010). Toward a new way of mapping scientific fields: Authors' competence for publishing in scholarly journals. *Journal of the American Society for Information Science and Technology*, *61*(4), 772–786. https://doi.org/10.1002/asi.21282

Moed, H. F., & Van Leeuwen, Th. N. (1995). Improving the accuracy of institute for scientific information's journal impact factors. *Journal of the American Society for Information Science*, *46*(6), 461–467. https://doi.org/10.1002/(SICI)1097-4571(199507)46:6<461::AID-ASI5>3.0.CO;2-G

Moed, Henk F. (2008). UK Research Assessment Exercises: Informed judgments on research quality or quantity? *Scientometrics*, *74*(1), 153–161. https://doi.org/10.1007/s11192-008-0108-1

Moed, Henk F. (2017). *Applied Evaluative Informetrics*. Cham: Springer.

Muñoz-Écija, T., Vargas-Quesada, B., & Chinchilla Rodríguez, Z. (2019). Coping with methods for delineating emerging fields: Nanoscience and nanotechnology as a case study. *Journal of Informetrics*, *13*(4), 100976. https://doi.org/10.1016/j.joi.2019.100976

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2018). The impact of classification systems in the evaluation of the research performance of the Leiden Ranking universities. *Journal of the Association for Information Science and Technology*, *69*(8), 1046–1053. https://doi.org/10.1002/asi.24017

Rafols, I., Ciarli, T., & Chavarro, D. (2019). Under-reporting research relevant to local needs in the South: Database biases in rice research. In R. Arvanitis & D. O'Brien (Eds.), *The Transformation of Research in the South Policies and outcomes*. Éditions des archives contemporaines. https://digital.csic.es/handle/10261/226953. Accessed 28 November 2023

Ràfols, I., Molas-Gallart, J., Chavarro, D. A., & Robinson-Garcia, N. (2016). *On the Dominance of Quantitative Evaluation in 'Peripheral' Countries: Auditing Research with Technologies of Distance* (SSRN Scholarly Paper No. ID 2818335). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2818335. Accessed 7 January 2019

Robinson-Garcia, N., & Calero-Medina, C. (2014). What do university rankings by fields rank? Exploring discrepancies between the organizational structure of universities and bibliometric classifications. *Scientometrics*, *98*(3), 1955–1970. https://doi.org/10.1007/s11192-013-1157-7

Robinson-Garcia, N., Torres-Salinas, D., Vargas-Quesada, B., Chinchilla-Rodríguez, Z., & Gorraiz, J. (2023). Errors of measurement in scientometrics: Identification and calculation of systematic errors. In *Proceedings of ISSI 2023 – the 19th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 2, pp. 387–393). Presented at the ISSI 2023, Bloomington, IN (United States). https://doi.org/10.5281/zenodo.8428899

Robinson-Garcia, N., Van Leeuwen, Th. N., & Torres-Salinas, D. (2020). Measuring Open Access uptake: Data sources, expectations, and misconceptions. *Scholarly Assessment Reports*. https://doi.org/10.5281/zenodo.4071143

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: the case of oncology journals. *Scientometrics*, *82*(3), 567–580. https://doi.org/10.1007/s11192-010-0172-1

Scuro, S. R. (2004). Introduction to error theory. *Visual Physics Laboratory, Texas A&M University, College Station, TX, 77843*. http://web.ist.utl.pt/~mcasquilho/compute/errtheory/,basics/ScuroErrTheo.pdf. Accessed 28 November 2023

Selivanova, I. V., Kosyakov, D. V., & Guskov, A. E. (2019). The Impact of Errors in the Scopus Database on the Research Assessment. *Scientific and Technical Information Processing*, *46*(3), 204–212. https://doi.org/10.3103/S0147688219030109

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005

Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of Science, Scopus, and Dimensions as varying resonance chambers. *Scientometrics*, *127*(5), 2413–2431. https://doi.org/10.1007/s11192-022-04309-6

Sugimoto, C. R., & Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, *71*(4), 775–794. https://doi.org/10.1108/JD-06-2014-0082

Thijs, B., Zhang, L., & Glänzel, W. (2015). Bibliographic coupling and hierarchical clustering for the validation and improvement of subject-classification schemes. *Scientometrics*, *105*(3), 1453–1467. https://doi.org/10.1007/s11192-015-1641-3

Torres-Salinas, D., Arroyo-Machado, W., & Robinson-Garcia, N. (2023). Bibliometric denialism. *Scientometrics*, *128*(9), 5357–5359. https://doi.org/10.1007/s11192-023-04787-2

van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Raan, A. F. J. V. (2001). Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. *Scientometrics*, *51*(1), 335–346. https://doi.org/10.1023/A:1010549719484

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies*, 1–22. https://doi.org/10.1162/qss_a_00112

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., et al. (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. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wang, Q., & Waltman, L. (2016). Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus. *Journal of Informetrics*, *10*(2), 347–364. https://doi.org/10.1016/j.joi.2016.02.003