Journal of Informetrics 20 (2026) 101810

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Research paper

Measuring the effect of structural differences between Web of Science and Scopus on research impact assessment

Sándor Soós [a,b,*], Anna Kiss [a,c], Zsófia Viktória Vida [a]

[a] Department of Science Policy and Scientometrics, Library and Information Centre of the Hungarian Academy of Sciences, Budapest, Hungary
[b] Faculty of Education and Psychology, ELTE Eötvös Loránd University, Budapest, Hungary
[c] Pro-Sharp Research and Innovation Centre, Budapest, Hungary

ARTICLE INFO

Keywords:
Research impact assessment
Novel metric
Citation impact
Country-level publication output
Database-specific structural features
Web of Science, Scopus

ABSTRACT

Our work elaborates on the study of how the choice of databases affects the measurement of scientific impact. We address a gap in the analysis of the "crown indicator", the Normalized Citation Score (NCS), when applied in practical settings in different database contexts. It is shown that database-specific structural features have a profound effect on the differences in NCS values across databases, and that the benchmark component of the metric can be used as a proxy for quantifying the effect of such differences ("benchmark effect"). The primary aim of our analysis was the measurement of the effect of benchmark differences on impact scores at various levels of assessment. To that end, we developed and theoretically validated a novel metric for quantifying the effect of benchmark differences on NCS discrepancies across databases (Citation-Benchmark Differentials Ratio, CBDR). We applied the CBDR metric on the five-year country-level publication output of Hungary, retrieved from WoS and Scopus. Measurement results were subjected to statistical analyses to reveal the potential benchmark effect in various aggregations of the publication record, especially by research fields, research organizations (HEIs), journal quartiles, and publication years. Results show that a considerable fraction of the country-level output is affected by the benchmark effect, and no substantial differences emerge between the units of assessment in any aggregation under study. It is demonstrated that database-specific citation counts and journal coverage only partially account for the differences between WoS- and Scopus-based normalized citation scores.

1. Introduction

The "crown indicator" for the bibliometric measurement of academic or scientific impact, the Mean Normalized Citation Score or the MNCS measure (Waltman et al., 2011) not only has preserved its "royal" position in the bibliometric community and the literature on impact measurement, despite reoccurring criticisms and proposed alternatives (Abramo & D'Angelo, 2016a,b), but also gained strong popularity outside the scientific discourse in the praxis of research assessment. Outstanding examples of its spread are the big commercial services, namely the research intelligence tools provided by Clarivate and Elsevier on the basis of their respective databases, the Web of Science and Scopus. Both tools, InCites from Clarivate and SciVal from Scopus, offer the NCS measure for various units of analysis (usually in the aggregated form, as the Mean Normalized Citation Score, MNCS) under different names. In the

* Corresponding author.
E-mail address: soos.sandor@konyvtar.mta.hu (S. Soós).

https://doi.org/10.1016/j.joi.2026.101810
Received 19 August 2025; Received in revised form 9 April 2026; Accepted 14 April 2026
Available online 25 April 2026
1751-1577/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Scopus—SciVal system, the measure is called the *Field Weighted Citation Index* (FWCI), while in the WoS—InCites system it is the *Category Normalized Citation Index* (CNCI), both labels emphasizing the most relevant feature of the metric (hereafter: impact metric) of controlling for research fields' different citation densities in the measurement of citation impact. The practical significance of this metric is clearly underscored by the fact that it is employed in the most prominent university rankings: in the Times Higher Education (THE) ranking methodology, citation impact is covered by the FWCI metric based on Scopus data. Given the nominal weight of this metric in the ranking score and the findings of the related empirical studies (Robinson-Garcia, 2019; Tóth et al., 2024), FWCI is one of the strongest, if not the strongest predictor of ranking positions.

Although the very same NCS measure is applied for both the WoS-based and the Scopus-based services, hence the technical definition and calculation of its variants are basically the same, the actual values of the impact metric for individual papers and aggregates are expected to differ in the two respective databases. This expectation is primarily based on the differing database contexts, i.e., different data fed into the calculation of the metric, being a natural consequence of, e.g., different journal coverage in WoS vs. Scopus, inducing database-specific citation environments. Comparisons between the size and structure of database coverage are traditional exercises in the literature. A recent large-scale comparative study (Visser, Van Eck & Waltman, 2021) analyzed the structure of the content of four central databases (WoS, Scopus, Dimensions, Microsoft Academic). Beyond overall publication coverage, structural dimensions of database content included the distribution of coverage among publication years, disciplines, document type, number of references, and number of citations per document. On one hand, a close structural similarity between WoS and Scopus content was found along these dimensions, at least relative to combinations with the other two databases. On the other hand, the differences found in disciplinary structure, especially in publication coverage at the lower region of the citation distribution, certainly indicate a differing citation environment. Altering the representation of disciplines and the citation distributions per database implies altering the potential number of citations and the reference set for any publication, which naturally affects its NCS value at various levels (see the analysis of factors affecting the NCS value below).

The database as a citation environment was also addressed by the analysis of Thelwall (2018) in a comparison between Scopus and Dimensions. In this case, the effect of the structural differences on citation counts was investigated using a smaller scale, a specific subject category in the database (Food science). Results in this case showed a high correlation between citation counts based upon the two different sources (Scopus and Dimensions), what's more, between Scopus citations and the RCR (Relative Citation Ratio) measure in Dimensions, the latter being a normalized impact measure similar to the NCS value. However, the high level of agreement was most likely attributable to the sample covering a single and homogenous subject category, indicated by the high cross-database correlation even between the raw and normalized citation measures (ibid., p.433). The relevance of sampling and sampling units is especially underscored by another comparative study, which directly investigated how the robustness of university rankings may be affected by the choice of the citation database used in the ranking (Huang et al., 2020). In particular, the study investigated the potential differences between WoS, Scopus, and Microsoft Academic in the case of the output of 15 selected universities in order to gauge the sensitivity of the ranking (of these unis) to database choice. The results clearly showed that differences in output size, composition, and citation counts, even OA-levels are dramatically shifted the rank positions of these institutions when altering the underlying database. Common to all the previous approaches is that the effect of database choice on impact measurement is only investigated through comparing citation counts.

Fewer studies focus on analyzing the effects of the choice between databases on more complex citation measures: a more frequent theme has been the choice between the crown indicator and some alternative metric, both applied on the same database, such as the Relative Citation Ratio or RCR (Purkayastha et al., 2019) or the Fractional Scientific Strength or FSS (Abramo and D'Angelo, 2016c). Among the works addressing database choice, Pech and Delgado (2020) provided a comparison of the (raw) citation percentile values across research fields between WoS and Scopus, based on a common research categorization scheme to ensure commensurability. Closer to our present focus is the work of Stahlschmidt and Stephen (2022), who directly addressed the difference between the database-specific Normalized Citation Scores (NCS) of the country-level output of Germany, that is, compared the CNCI and FWCI values for the fraction of the German publication output that was covered both in WoS and Scopus (and also in the Dimensions database).

These studies also represent a clear relevance for research assessment by demonstrating various effects, or even biases introduced by the choice of databases, on the measurement of research performance at various levels. An outstanding example of such potential biases is the choice between university rankings, each of which depends on a selected database to provide the bibliometric indicators for the ranking composite.

1.1. The network of effects of database organization on the crown indicator

Our present work elaborates on the study of how the choice of databases affects the measurement of scientific impact and, consequently, the assessment of research performance. In particular, we address a gap in the analysis of the "crown indicator", the Normalized Citation Score (NCS), when applied in practical settings in different (database) contexts, through such metrics as the CNCI or the FWCI. The starting point of this analysis is the fact that the NCS is a relatively complex measure in the following sense. For a given item or publication, it compares the citation count of the item to a benchmark value, which is the average citation count of all papers in the same research field, publication year, and of the same document type, for the purposes of normalization and commensurability. Hence, an item's score is a function of various factors that directly and/or indirectly shape the metric's outcome, where indirect effects are manifested in the interplay between these factors. For a database setting like the WoS or Scopus, and especially when accounting for NCS differences across databases, a minimal model of the network of such factors is presented in Fig. 1. The model shows the database features along with the pathways that influence the numerator of the NCS metric, i.e., the citation count

of an item, and those for the denominator of NCS, viz. the benchmark value. According to this model, the citation count component is mainly affected by database journal and item coverage, while the benchmark component is the net result of the interplay between various database features. Database coverage is one such feature, but also its categorization into research areas usually specific to the respective database. The journal composition of those categories determines the "reference set" inducing the benchmark value, the average citation count for that set. Coverage and categorization, therefore, interact in forming the benchmark value. Even more involved is the influence of item assignments to research categories. In the first place, items are naturally assigned to content-wise differing categories (hence different reference sets) across databases, given the database-specific category systems. Secondly, most items are assigned to more than one category (Wang & Waltman, 2016), which can be assumed to amplify the effect of the database-specificity of research categorization on benchmark formation. According to currently available online sources provided for Scopus/SciVal and WoS/InCites, the multiplicity of categories is considered via using multiple field averages for calculating the benchmark value. Both the CNCI and the FWCI combine these via the harmonic mean function to come up with a benchmark with a well-balanced contribution of the categories composing the reference set(s).

1.2. The benchmark component as a summary of database organization

In sum, the model conveys that impact measurement through the NCS metric is severely affected by a series of interrelated factors that are characteristic of the database supporting the measurement or assessment. We may call these factors and the related differences between databases as "structural features" or "structural differences". The main aim of the present research is to provide a systematic study of the effect of these structural differences between WoS and Scopus on impact measurement across these two databases. Though several studies addressed various aspects of this phenomenon, such as the implications of differing or "mis-" categorizations of research output on the outcomes of assessment (cf. Bartol et al., 2016; Robinson-García & Calero-Medina, 2014), our attempt goes further in order to explicitly quantify the contribution of the structural features to the differences in impact assessment through the NCS metric. Instead of the direct estimation of the role of individual factors, we capitalize on the fact that these effects are summated in the benchmark value, i.e., in the denominator of the NCS metric. Therefore, when it comes to the role of database choice, it is more effective to investigate the extent to what extent benchmark differences account for the differences in the impact metric between databases. Previous studies have investigated the sensitivity of the NCS metric to various methodological aspects in constructing the denominator or benchmark component of the measure. Smolinsky (2016) demonstrated the mathematical properties of the crown indicator with respect to the formulation of its denominator as the expected number of citations. Ruiz-Castillo and Waltman (2015) analyzed the potential differences in MNCS values when the metric is applied along different algorithmic field categorizations of the same dataset, providing different benchmarks for the calculation of the NCS measure at the level of publications. The focus of their study was the difference in the granularity of the categorization schemes. Haunschild, Daniels and Bornmann (2022) addressed the method of categorization as a potential factor affecting NCS scores, testing citation relations, topical similarities, intellectual relatedness, and journal or journal-based categories behind the formation of citation benchmarks. Common in these studies was that they varied the categorization schemes but kept the dataset and the database, i.e., the "universe of discourse" constant. In contrast, our goal is to go a step further and consider the effect of alternating the database context as well.

We therefore use benchmarks as a proxy for measuring the effect of structural features of a database (WoS or Scopus, in this case). The theoretical validation of using the benchmark component as an aggregate of structural effects lies in the model outlined in the previous section. As this model demonstrates the pathways through which the identified factors affect both the citation count component and the benchmark component of the NCS metric, we shall focus on the latter, i.e., the pathways accounted for by the denominator in the calculation scheme. In particular, the denominator/benchmark is the average number of citations received by documents in the reference set of the target document. This value, or, rather, the citation distribution represented by the benchmark is therefore determined by 1) the size and 2) the composition of database coverage in terms of indexed journals and disciplines, age, and

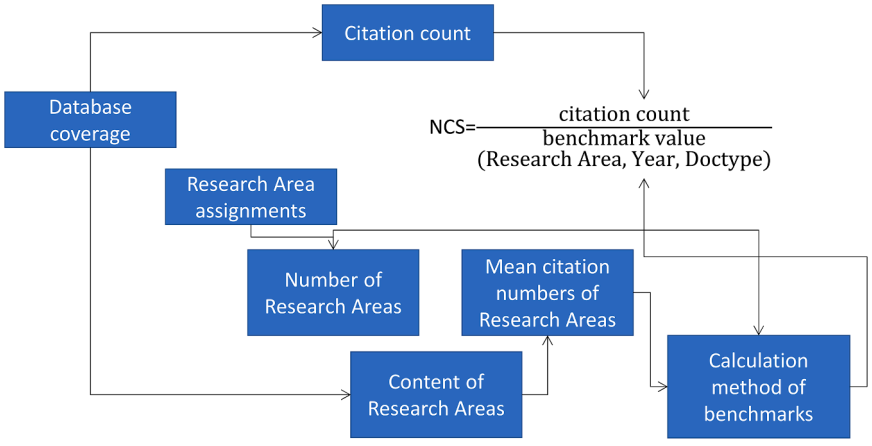

Fig. 1. A model of database-related factors affecting the impact metric (Normalized Citation Score).

document type as the database-specific citation environment of the reference set. Furthermore, since the reference set is supposed to convey an age cohort of publications belonging to the same research field, its citation distribution (hence, the benchmark value) is directly determined by 3) the field categorization of database content, mostly journals. As a further complication, the reference set can be a composite of the citation distributions of various field categories in the case of multidisciplinary assignments. The field composite is defined by the calculation scheme for the benchmark value applied in either the CNCI or the FWCI definition (see the comparison of the two versions below). As such, this factor does not strictly represent a database feature; nevertheless, both versions of the impact metric are the "native" elements of the corresponding database contexts, so that we can legitimately deem 4) the method of combining research fields for multidisciplinary reference sets as the next database-specific factor. In short, the link between structural features and the benchmark value is the recognition of how these structural features shape the reference set for the target publication, based on which set the benchmark value is discerned.

In addition to quantifying this effect, by the very same token, we can characterize the relative effect of citation counts (and their differences) on the impact metric, as compared to the effect of structural features, i.e., benchmarks. In this way, we will be able to show the sensitivity of the NCS impact metric to its basic components, the number of citations of an item, and the associated benchmark value across the databases under study. To put it another way, using the language of research assessment, we may be able to show whether publications are more (less) impactful in WoS or Scopus, either because more (less) citations could be collected from the respective database context, or because of the structural differences between WoS and Scopus (or both). Based on the above discussion, the study addresses two main research questions.

RQ1: What is the extent to which the structural differences between WoS and Scopus affect academic impact measurement? In other words, given the above operationalizations, what is the exact contribution of benchmark differences, as compared to citation count differences, across these databases in altering the NCS impact metric, resulting in differing CNCI and FWCI scores?

RQ2: Are the groups of publications defined along typical factors in research assessment, such as the units of evaluation (institutions, research fields), annual cohorts, or papers in various journal quartiles, differentially affected by the structural differences instantiated in benchmark differences between the databases? To what extent can these factors account for the variation of the structural effect (which we call the "benchmark effect") within (and beyond) the sample?

Our research design to gauge the effect and factors of structural differences was as follows: 1) we developed a metric with the purpose of identifying and quantifying the relative effect of benchmark differences on the Normalized Citation Score (NCS) discrepancies across databases for publications occurring in each respective database, which we called the *Citation-Benchmark Differentials Ratio (CBDR)*. We applied the measure on a large-scale publication record, representing a five-year country-level publication output, retrieved from (the overlapping fraction of) WoS and Scopus. Measurement results were subjected to a series of statistical analyses to reveal the overall effect of structural differences on the country-level output, and the extent to which this effect varies across the groups of publications discussed above, thereby investigating the related relevance of grouping factors often employed in research assessment.

2. Materials and methods

In order to gauge the relative effect of benchmark differences on the NCS differences of database items in WoS and Scopus, we developed a specific measure leveraging the analytic relationship between the NCS impact metric and its components as detailed below.

2.1. The measurement of the effect of benchmark differences

In order to detect the contribution of benchmark differences to the impact discrepancies, we utilized the simple mathematical relations between the benchmarks and the NCS measure. Let *c1* and *c2* be the citation counts for paper P in WoS and Scopus, respectively. Furthermore, let *b1* and *b2* be the respective benchmarks pertinent to P in the two databases (i.e., the field- and year-specific citation averages in the field(s) or research categories to which P is being assigned). The ratio of database-specific impact scores can be expressed as

$$NCS_1 / NCS_2 = (c1/b1) / (c2/b2) = (c1/c2) * (b2/b1) \quad (1)$$

where $NCS_{i=(1,2)}$ is the impact score obtained from the respective database *i=(1,2)*, Eq. (1) conveys that the impact discrepancy for P is directly proportional to both the ratio between its respective citation counts (citation ratio) and the ratio between the respective benchmarks (benchmark ratio) for P in the two databases. This formulation also shows that identical impact scores (i.e., when their ratio equals 1) can only occur when the benchmark values and the citation count values are the same across databases (both ratios equal 1), or when these values perfectly compensate each other to yield the same impact scores. This latter situation can happen when one of the two ratios is the reciprocal of the other one (e.g., the citation ratio = 5 and the benchmark ratio = 1/5, resulting in an impact ratio = 5 × 1/5 = 1).

More importantly, due to the direct proportionality of the impact ratio and the benchmark ratio to the citation count ratio, the greater quantity of the latter two will increase the impact ratio with a percentage that exceeds the smaller quantity (ratio). We might say that, in this case, this greater quantity has a higher influence on the impact ratio (the discrepancy of impact metrics between databases), which influence can be directly captured and quantified by comparing the values of the citation ratio and the benchmark ratio. This allows us to formulate such a comparative measure. In particular, this enables us to capture publication items with this condition, where benchmark differences have more weight in citation score discrepancies than citation scores themselves, we

introduced a measure labelled as "Citation-Benchmark Differentials Ratio" (CBDR) for a given paper P as follows:

Let's denote c1/c2 with *c-ratio*, and b2/b1 with *b-ratio*. Then, we can define the so-called *Citation-Bechmark Differentials Ratio* as

$$\text{CBDR} = \text{c-ratio}^* \,/\, \text{b-ratio}^*. \tag{2}$$

where c-ratio* = 1/c-ratio, if (c-ratio ≤ 1 and b-ratio > 1) and c-ratio otherwise b-ratio* = 1/b-ratio, if (c-ratio ≥ 1 and b-ratio < 1), and b-ratio otherwise.

The rationale for introducing these modified versions of the original ratios is as follows. The CBDR measure is supposed to compare the contribution of the two ratios to the impact score ratio based on their relative size. However, in several cases, one ratio exceeds the other only due to the order of database comparison (database 1 compared to database 2), although the other ratio contributes more to the outcome: this can happen when the two ratios are above and below 1, respectively. Consider, for example, a c-ratio = 1, and a b-ratio = 0.7. Clearly, in this case, the b-ratio decreases the outcome (the impact score ratio) with 30% (1 × 0.7 = 0.7), while the c-ratio alone causes no impact discrepancy: still, technically, the c-ratio > b-ratio, which suggests a greater contribution of the c-ratio. To avoid this bias, in such cases, the ratios are compared along their "absolute value", taking the reciprocal of the below-one-valued ratio. In our example, instead of the raw b-ratio, b-ratio* as its reciprocal (1/b-ratio) is taken, yielding a b-ratio*=1.43. Hence, the CBDR = 1/1.43 = 0.7, which captures the 30% excess of the benchmark ratio above the citation ratio, also capturing how much more it contributes to the outcome relative to the other measure (the c-ratio).

The CBDR measure in Eq. 2 is essentially the ratio between the differential of citation counts (c1/c2) and of benchmark values (b2/b1) with some necessary technical modifications. As such, the measure has the following mathematical properties relevant to our research questions:

$$\text{If CBDR} = 1 \text{ then b-ratio}^* = \text{c-ratio}^*. \tag{3}$$

$$\text{Iff CBDR} > 1 \text{ then b-ratio}^* < \text{c-ratio}^*. \tag{4}$$

$$\text{Iff CBDR} < 1 \text{ then b-ratio}^* > \text{c-ratio}^*. \tag{5}$$

These properties qualify the CBDR measure as a conceptually valid tool for detecting the effect of benchmark differences on impact measure discrepancies (as compared to differences in citation counts) between databases. In particular, (only) if the CBDR value of a publication is lower than 1, the "absolute" ratio of the benchmarks (b-ratio*) exceeds the "absolute" ratio of the respective citation counts (c-ratio*), which conveys that the contribution of differing benchmarks to citation impact is larger than that of differences in citation counts (as displayed in Eq. (1)). The size of this quantity is indicative of the size of the effect: the lower the CBDR value is below 1, the larger the effect of benchmark differences on the impact measure.

2.2. The bibliometric sample

In order to test the effect of benchmark differences on impact measurement, two matched bibliometric samples were taken from the Web of Science databases and the Scopus database, respectively. For the purposes of demonstrating direct implications for research evaluation practices, the matched samples were designed to delineate the publication output of a selected country in a recent and relatively long, five-year period. The country-based sample was motivated by the fact that it is a typical unit of evaluation with large-scale data for our analysis, especially when taken from a wide-enough publication window. Accordingly, bibliographic data was retrieved via the search query scheme "*Country of affiliation* = Hungary AND *Publication Year* = (2018—2022). Using the corresponding query in WoS resulted in a total of $n = 66{,}995$ publications, comprising the first sample (WoS sample). From Scopus, the corresponding query returned $n = 69{,}563$ publications, out of which the second sample was formed (Scopus sample).

As the next step of data processing, the two samples were matched based on their overlap. More precisely, for the comparative analysis of database-specific parameters (impact values, citation counts, and citation benchmarks), we have identified the publications that occurred in both samples, taken from WoS and Scopus, respectively. For matching identical papers in the two datasets, we used the DOI number. The matching procedure resulted in a final sample of n = 64,961 papers submitted to subsequent analysis.

2.3. Model calibration

In order to apply the CBDR index in the measurement of benchmark effects, we collected the corresponding benchmarks and impact metrics for the WoS and the Scopus sample from the respective databases. The data was retrieved from the InCites analytic service, part of the WoS platform, and the SciVal service accompanying the Scopus database, respectively. According to the terminology used in these services, the normalized citation score (NCS) measure, as reported in the Web of Science context, was represented via the *Category Normalized Citation Index* (CNCI) values collected for our WoS sample from InCites. The same measure in the Scopus context is dubbed the *Field Weighted Citation Score* (FWCS) within SciVal, providing data for our Scopus sample. Although the two metrics are conceptually the same and play the role of database-specific versions of the crown indicator, at least one technical difference in their calculation should be noted: the FWCI uses a differing citation window both in the citation count and the benchmark component as compared to CNCI. In particular, the FWCI limits the counting of citations for each component in 3 years after publication, resulting in a 4-year citation window (Scelles & Teixeira da Silva, 2025), whereas the CNCI does not have this limitation, implying a citation window from publication up to the date of calculation. However, given the circumstances of data collection for our study, this discrepancy does not seriously impacted our comparisons between the two metric: due to the 5-year publication window applied in our

analysis and the data collection conducted in the 5th year of this window, the actual citation windows under study almost fully overlapped for obtaining values of the CNCI and the FWCI, so that the validity of the comparison is not violated by this feature.

Crucial for our comparison were the retrieval of the respective citation benchmark data for WoS and Scopus: we used the *Category expected citations* field in InCites to add benchmark data to WoS publications, and the *Field citation average* field in SciVal for Scopus publications (both values are provided for each sample item in the respective services). To validate "corporate" source benchmark data, we also produced a calculated version of the benchmark variable via dividing the citation counts by the CNCI and FWCI values item-wise for WoS and Scopus, respectively. Database-specific citation counts were also harvested from InCites and SciVal, respectively. Given that the CBDR measure, as defined above, is not applicable for publications that are uncited in any of the two databases (since the citation ratio would contain a zero denominator), we slightly modified citation counts to overcome this problem. In particular, zero citations were replaced with a value of 0.05, which is below the natural minimum of the citation count metric (i.e., 1), therefore preserving the meaning of uncitedness but still allowing the CBDR measure to be meaningfully computed for the item. We also conducted a sensitivity analysis of the CBDR metric relative to the choice of the replacement value. Given a series of alternative values representing the (0,1) interval, the pointwise difference of the original and the modified CBDR distribution was calculated. Results showed that altering the replacement value does not affect the CBDR, as no difference could be detected throughout the distribution for any alternative value (at percentiles 5, 10, 25, 50, 75, 90, the difference was equal to 0). The only exceptions were the extreme values in the distribution (percentiles 0 and 100), but those were basically outliers not affecting the robustness of the conclusions.

2.4. Data analysis

As a preliminary analysis of the discrepancies between the WoS and Scopus contexts, the citation impact scores were compared between WoS and Scopus for our sample, using (1) descriptive statistics of the respective distributions and 2) non-parametric ANOVA (Kruskal-Wallis) tests of the potential CBDR differences between relevant groups of sample publications with a special focus on effect size measurement. As the core analysis of our study, the CBDR measure was applied on the final (country-level) sample to detect and characterize the effect of benchmark differences on impact measurement. The resulted distribution of the CBDR measure was analyzed according to different levels and types of aggregations, (1) at the sample (i.e. country) level, and comparing the output of (2) organizations (universities), (3) research fields, (3) publication years, (4) journal quartiles and (5) publications categorized into different numbers of research fields. Descriptive statistical analysis and ANOVA (Kruskal-Wallis) tests of the CBDR differences between the groups according to each factor were conducted.

3. Results

3.1. Preliminary descriptive analysis: comparison of citation impact values between WoS and Scopus

The preliminary analysis of citation impact differences was based on computing the ratio of the normalized citation score (NCS) for the country-level output reported in *SciVal* and *InCites*, respectively. Three parameters of the distribution of the paper-level NCS ratios, the mean, the median, and the 3rd quartile values, are reported. Also, in our present and subsequent analysis, only those items were taken into account where the reported citation metrics actually differed between the two databases (i.e., where CNCI $\neq$ FWCI – the reason being that we were interested in how much benchmark differences account for impact differences, once the latter exist). This subsample amounted to $n = 43{,}700$ items, about 70% of the full sample.

At the county-level, a mean ratio of 1.98 has been found, conveying a 98% higher citation impact in the Scopus context on average. The median and the 3rd quartile values were, on the other hand, 1.14 and 1.6, respectively, jointly signalling a skewed distribution that moderates the citation (metric) advantage for Scopus, but still preserves it for a large amount of publications (at least 25% of the country output, based on the 3rd quartile value).

A more detailed picture was obtained based on two aggregation schemes of interest in research assessment: (1) research fields (using the scheme for the categorization of the country-level sample into research fields provided in SciVal, especially for its relevance

Table 1
Ratio of the normalized citation score (NCS) of publications in Scopus vs. WoS by research field.

Research field	p	mean	median	q3	min	max
Arts and Humanities	515	**1.61**	1.11	**1.93**	0.01	18.15
Social Sciences	1269	**1.47**	1.13	**1.62**	0.01	17.18
Physical Sciences	11,695	**1.47**	1.19	**1.61**	0.00	82.93
Education	192	**1.42**	1.05	**1.60**	0.10	12.80
Law	83	**1.40**	1.14	**1.53**	0.14	6.86
Computer Science	1616	**1.36**	1.13	**1.51**	0.04	14.70
Engineering and Technology	5179	**1.34**	1.12	**1.46**	0.06	66.05
Life Sciences	7492	**1.34**	1.12	**1.52**	0.02	28.27
Clinical, pre-clinical and health	10,172	**1.31**	1.08	**1.47**	0.05	35.91
Business and Economics	805	**1.29**	1.01	**1.52**	0.09	10.99
Psychology	653	**1.23**	0.99	**1.35**	0.09	7.11

for university rankings) and (2) universities (HEIs).

In the case of research fields (Table 1), the median value ranges between approx 1—1.2, while the mean conveys a more dramatic ratio (range: 1.2—1.6), notably with small differences between research fields. Also, for each field, the mean is much closer to the 3rd quartile value (range: 1.3—1.9) than to the median, reflecting a skewed distribution. In sum, at least 25% of the papers in each field are substantially inflated in terms of academic impact (NCS) when shifting from the context of WoS to that of Scopus, which is not a negligible quantity given the size of analyzed research fields (number of publications, column *p* in Table 1). The degree of this inflation shows little variation between fields, although on the top-ranked field (Arts and Humanities) impact is multiplied by almost a factor of 2 (3rd quartile value = 1.9, but also the Physical Sciences exhibits a value of 1.6), while the lowest-ranked, Psychology, shows an inflation factor of around 1.4. Still, as most ratio values are above 1.5, papers in this 25% having an impact value in WoS that amounts only to half-world average (from NCS_{WoS}=0.6) will be positioned very close or above the world average (NCS=1) in Scopus (with $NCS_{Scopus} \geq 0.9$).

Note: the abbreviations used in this table refer to the following university names: SZTE: University of Szeged; DE: University of Debrecen; PTE: University of Pécs; SE: Semmelweis University; ELTE: Eötvös Loránd University; BME: Budapest University of Technology and Economics; MATE: Hungarian University of Agriculture and Life Sciences; PE: University of Pannonia; ME: University of Miskolc; OE: Obuda University; BCE: Corvinus University of Budapest; ÁE: University of Veterinary Medicine Budapest; SZIE: Széchenyi István University; PPKE: Pázmány Péter Catholic University; KRE: Károli Gáspár University of the Reformed Church in Hungary

A very similar landscape unfolds regarding the case of Hungarian universities (or HEIs, see Table 2). Only HEIs with a publication output of $p \geq 100$ between 2018 and 2022 are included. Again, each HEI exhibits substantially differing WoS- and Scopus-based NCI values, respectively, in at least 25% of their publication output (given the 3rd quartile boundary values reported in Table 2). However, just like in the case of research fields, these values (i.e., the boundary values) show little difference between HEIs, in most cases being close to 1.5. These slight differences still induce some correlation with size: HEIs show a weak-to-medium negative (rank) correlation between the number of publications and, respectively, the mean (Spearman's $\rho \approx -0.3$) and 3rd quartile values (Spearman's $\rho \approx -0.5$). These relationships are demonstrated in Fig. 2 by plotting these values against the rank of universities in the size distribution of HEIs (ordered decreasingly, from the largest to the smallest HEI). According to this result, as size of publication output decreases, HEIs' impact assessment becomes more prone to the differences between WoS and Scopus (in terms of citation impact measures).

3.2. Application of the CBDR: measuring the effect of benchmark differences on impact metrics discrepancies

After determining the discrepancies between WoS and Scopus impact metrics values, the CBDR measure was applied to the sample to address the main question of the present study regarding the contribution of citation benchmark differences to this finding. Computing the CBDR upon the whole sample resulted in normally distributed values (Fig. 3), with a mean of M = 0.9 (after removing outliers), a median of Mdn = 0.9 and a 3rd quartile value of 1.05, implying that, on average, benchmark differences have a somewhat larger effect on impact metrics (their differences) than citation count differences between the two databases, as both M(CBDR) < 1 and Mdn (CBDR) < 1. The pooled effect, however, appears to be moderate as the average CBDR value is still quite close to 1. Nonetheless, from a research assessment perspective, it is notable that at least half of the sample output is subject to the "benchmark effect", which is underscored by the fact that 50% of the papers are below the mean value (0.9) in terms of the CBDR measure.

In order to further investigate the effect of benchmarks on impact metrics, in our subsequent analysis, we were focusing on the fraction of the sample for which CBDR<1 was found, i.e., where benchmark difference had a larger effect on impact metrics than citation difference. For convenience, we refer to the effect of these differences as the "benchmark effect", and the "citation effect", respectively. This fraction amounted to 66% (n = 26985) of the sample (part of the sample, where the impact metrics difference was detected). The mean and the median CBDR were M = 0.76 and Mdn = 0.8 respectively, both expressing a moderate but clear dominance of the benchmark effect over the citation effect. Even more critical for research assessment than the pooled effect, we

Table 2
Ratio of the normalized citation score (NCS) of publications in Scopus vs. WoS by universities (HEIs).

HEI	p	mean	median	q3	min	max
SZTE	12,231	**1.23**	1.05	**1.37**	0.01	66.05
DE	10,513	**1.35**	1.11	**1.51**	0.04	82.93
PTE	8808	**1.22**	1.04	**1.39**	0.08	19.52
SE	8195	**1.22**	1.02	**1.36**	0.02	29.04
ELTE	7700	**1.38**	1.11	**1.55**	0.03	82.93
BME	5774	**1.38**	1.15	**1.50**	0.00	22.58
MATE	3184	**1.40**	1.17	**1.59**	0.03	40.93
PE	1495	**1.30**	1.11	**1.55**	0.04	9.38
ME	1157	**1.34**	1.12	**1.53**	0.09	10.47
OE	989	**1.46**	1.22	**1.65**	0.04	18.35
BCE	917	**1.32**	1.04	**1.53**	0.11	8.43
ÁE	706	**1.28**	1.14	**1.51**	0.15	6.75
SZIE	527	**1.43**	1.12	**1.56**	0.13	11.28
PPKE	510	**1.31**	1.04	**1.42**	0.10	13.21
KRE	90	**1.35**	1.10	**1.55**	0.20	5.64

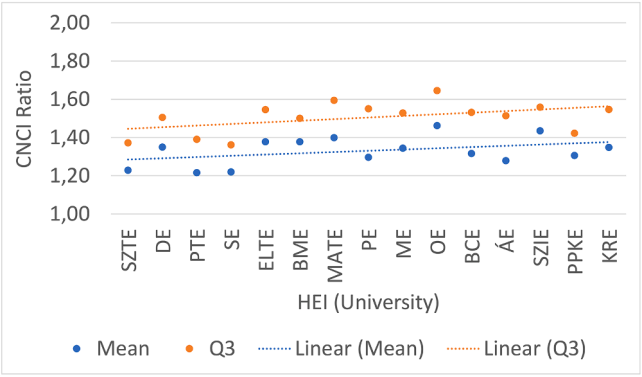

Fig. 2. The relationship between HEI output size and the NCS ratio (Scopus vs. WoS). Output size increases to the right.

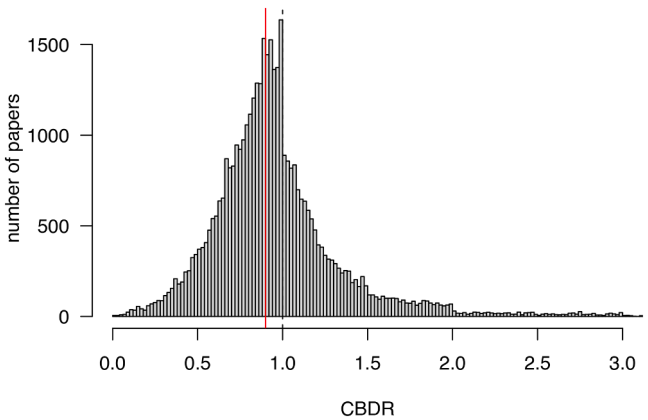

Fig. 3. Distribution of the CBDR measure over the full bibliometric sample.

addressed the relevant groupings of the sample introduced above. Consequently, the CBDR was applied to detect how benchmark differences affect (1) research fields and (2) universities (HEIs). Furthermore, we also investigated the relationship between the CBDR value and (3) the journal quartile, (4) the age of items (publication years), and (5) the database-specific multidisciplinarity of items, i. e., the number of research fields associated with publications in the respective databases.

3.3. Effect on impact metrics 1: research fields

The benchmark effect detected by research fields is reported in Fig. 4A and Fig. 4B The share of publications subject to the benchmark effect ranged between approx 60% and 70% throughout research fields, with the Humanities, Life Sciences, and Clinical Sciences having the highest (~70%) and Computer Science having the lowest (~60%) share; however, the distribution was close to being even with small differences in a 10 percent point range. The mean CBDR value calculated for research fields spanned from about 0.8 (Engineering, Clinical medicine, Physical Sciences) to 0.6 (Humanities), signalling a clear benchmark effect for each field (Fig. 4B). We can observe that the fields with the highest/lowest share of affected output exhibit the larger/smaller influence of benchmark differences (in terms of mean CBDR), respectively, although no substantive positive correlation between the two measures could be found o($\rho = -0.2$).

The between-field differences according to the Kruskal-Wallis test were, in general, statistically significant (existent) but practically insignificant: the measured effect size of research fields on the CBDR values was small ($p < 0.01$, $\eta^2(H) = 0.03$). The overlapping confidence intervals around the mean values (Fig. 4B) suggest that there are significant (but small) differences between, but not within the main groups of research fields (SS, H, and STEM).

3.4. Effect on impact metrics 2: universities (HEIs)

The benchmark effect differences among universities (HEIs) are demonstrated in Fig. 5A and B The amount of university output prone to the effect was found to range between approx 60% and 70%, but with the majority of HEIs having a value of close to 60%. The mean CBDR of the affected output varied between 0.7 and 0.8 among HEIs, again, conveying a systematic influence of the database differences under study. The between-university CBDR differences according to the Kruskal-Wallis test were still statistically

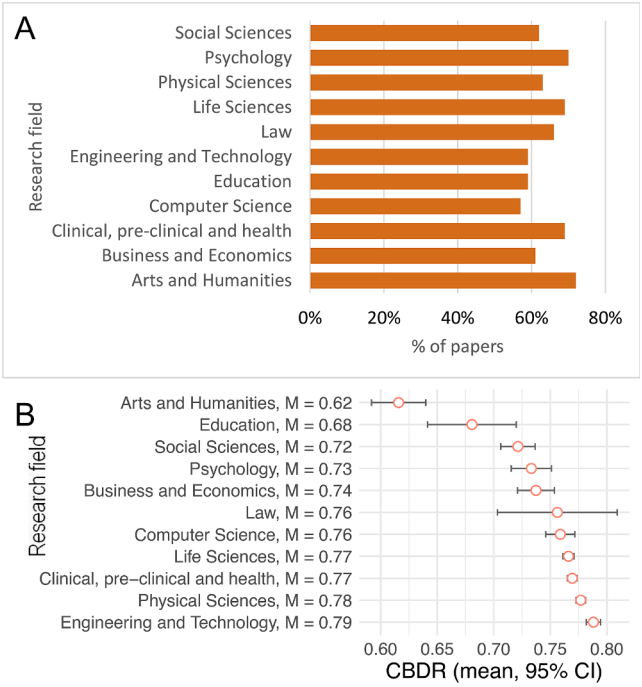

Fig. 4. A: The share of publications with the benchmark effect by research fields B: The mean CBDR value of publications by research fields (with a 95% confidence interval).

significant and practically negligible: the measured effect size of the university factor on the CBDR values was extremely small ($p < 0.01$, $\eta^2(H) = 4 \times 10^{-4}$). According to Fig. 5B, even the significance of the overall (small) differences was mainly attributable to certain HEI pairs, beyond a large sample size, such as *OE*, having a somewhat higher value than *BCE* with non-overlapping CIs (HEIs are ordered according to the size of their output). As seen above in the CNCI comparison, we observe a slight tendency of the CBDR mean increasing with HEI (output) size, suggesting that the benchmark effect somewhat decreases with scale (the Spearman correlation between size and mean CBDR was found to be $\rho = 0.6$)

Note: the abbreviations used in this Figure. refer to the following university names: SZTE: University of Szeged; DE: University of Debrecen; PTE: University of Pécs; SE: Semmelweis University; ELTE: Eötvös Loránd University; BME: Budapest University of Technology and Economics; MATE: Hungarian University of Agriculture and Life Sciences; PE: University of Pannonia; ME: University of Miskolc; OE: Obuda University; BCE: Corvinus University of Budapest; ÁE: University of Veterinary Medicine Budapest; SZIE: Széchenyi István University; PPKE: Pázmány Péter Catholic University; KRE: Károli Gáspár University of the Reformed Church in Hungary

3.5. *Effect on impact metrics 3: journal quartile*

Whether and to what extent the journal quartile makes a difference in the benchmark effect inflicted on country-level output is demonstrated in Fig. 6A and B Quartiles are represented in terms of inverted percentile ranges (that is, the range 0–25 corresponds to Q1, 25–50 to Q2, 50–75 to Q3, and 75–100 to Q4). Findings show that, similarly to the previous groupings, the share of affected papers was quite uniform, ranging between 65 and 70%. The average CBDR values were still lying between 0.6–0.8, evidencing the presence of the benchmark effect throughout journal quartiles as well, and the Kruskal–Wallis test also resulted in statistically significant, but practically insignificant differences with very low effect size, attributing a low explanatory power to journal quartile on the benchmark effect (($p < 0.01$, $\eta^2(H) = 0.007$). Furthermore, according to Fig. 6B, the mean CBDR for the 1st quartile was slightly (in terms of difference) but significantly (beyond confidence intervals) higher than that of the 2nd quartile, which, in turn, was also significantly higher than that of the 3rd and the 4th quartiles (the latter two not differing significantly). These differences signal that Q1 and Q2 publications can be expected to be somewhat less prone to the benchmark effect (in fact, Q1 types even less than Q2 types) than Q3 or Q4 papers compared to any of the other top categories.

3.6. *Effect on impact metrics 4: publication years*

Considering the year of publication, age cohorts of the country-level scholarly output also show, with little variation, high percentages of publications subject to the benchmark effect (Fig. 7A): the share of publications with CBDR < 1 varies between 60 and 70% in the period 2018–2022. The mean CBDR value ranges from about 0.7 to 0.8, evidencing a moderate but systematic effect of

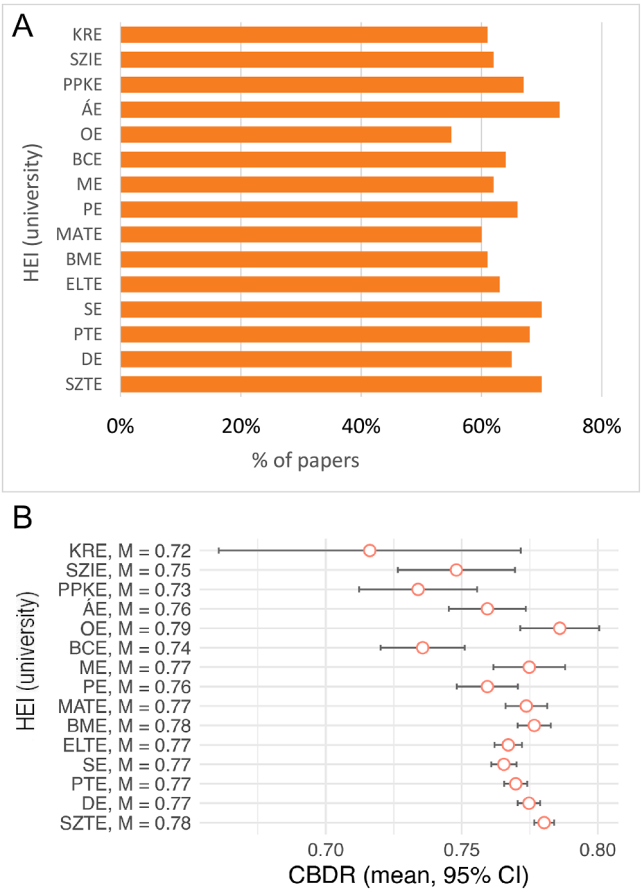

Fig. 5. A: The share of publications with the benchmark effect by universities (HEIs) B: The mean CBDR value of publications by HEIs (with a 95% confidence interval).

benchmark differences on impact metrics. The distribution of these values over the years under study, however, shows a tendency (Fig. 7B): we can observe a monotonic increase in the mean CBDR from 2018 towards 2021, suggesting that more recent cohorts tend to be less prone to the benchmark effect, than older publications (an interesting exception is the closing year, 2022, where the mean CBDR drops below the starting level in 2018). A plausible explanation may relate this tendency to the parallel effect of the length of the citation window on both paper citation counts and benchmark values (we make an attempt to explain this tendency in more detail under the *Discussion* section). Regarding the effect of this factor (year of publication) on CBDR, the Kruskal-Wallis test still resulted in a statistically significant but small effect ($p < 0.01$, $\eta^2(H) = 0.05$), which, however, is the highest shown by the factors under study, and also quite close to the range of medium effect sizes (considered to range from 0.06–0.14). Fig. 7B also shows that, besides the significance of the overall effect, most publication years also differ significantly (along the mean CBDR) in a pairwise comparison as well.

3.7. Effect on impact metrics 5: multidisciplinarity

Since the benchmark effect under study fundamentally originates from the differences in the contexts of research categorization between citation indexes (or databases), it is natural to inquire whether single vs. multiple assignments to research fields or areas make a difference to the benchmark effect (or CBDR values). In other words, the question is whether publications belonging (being assigned to) to a single research field show different levels of the benchmark effect than those belonging to two, three, or more fields, and which can be labelled as multidisciplinary publications. As a first attempt to address this question, in this study, we used the university ranking scheme to categorize the sample output, which categorization is provided in the *SciVal* database (as we did in the analysis of differences among research fields).

The results show very small differences in the percentage of publications subject to the benchmark effect among four types of items: those assigned (1) to a single research field, (2) to two research fields, (3) to three research fields, and (4) to more than three research fields (Fig. 8A). In each case, 60—70% of the items are affected. Interestingly, however, somewhat counterintuitively, single-field items are the most affected (68%), while multiple-field items with more than three fields are the least affected (61%). The average CBDR value ranges from 0.7 to 0.8, again, suggesting a general but moderate benchmark effect over these types (Fig. 8B). More notable is a clear positive correlation between the number of research fields and the mean CBDR value: the more fields are associated with the

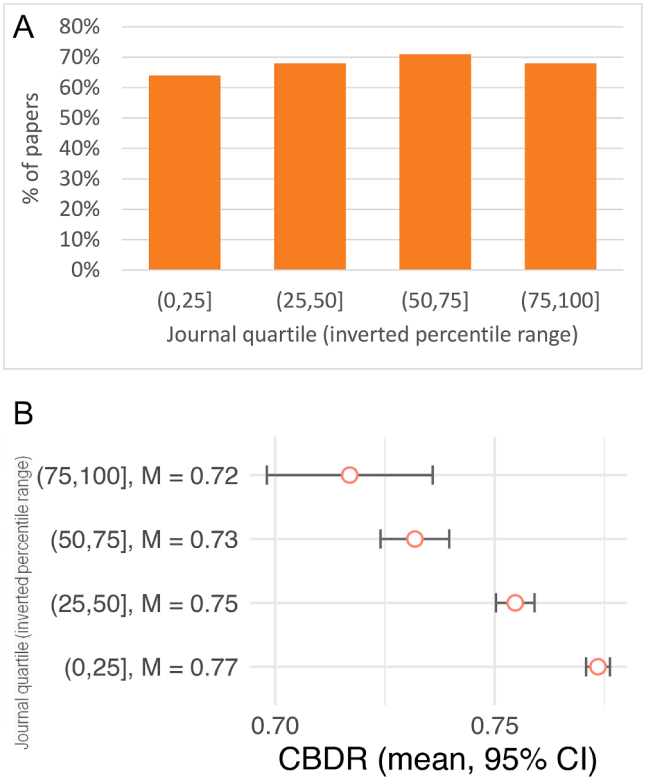

Fig. 6. A: The share of publications with the benchmark effect by journal quartiles B: The mean CBDR value of publications by journal quartiles (with a 95% confidence interval). Quartiles are represented as inverted percentile ranges (0–25: Q1, 25–50: Q2, 50–75: Q3, 75–100: Q4).

publication, the higher the expected CBDR value is, i.e., the less prone the publication seems to be to the benchmark effect. It can be hypothesized that multiple assignment somewhat compensates for the differences in the assignment contexts.

As to what extent multidisciplinarity (number of fields assigned) accounts for the variation of CBDR values within the sample, the Kruskal-Wallis test signifies a low explanatory power (significant but small effect size, $p < 0.01$, $\eta^2(H) = 0.001$). Despite its low contribution to CBDR variation, the pairwise differences between assignment types are clearly significant (Fig. 8B).

4. Discussion

The starting point for our research on the contribution of benchmark differences across databases to impact measurement was the assumption that those impact differences do exist and have an influence on research assessment through the choice of databases. Such differences have already been reported in the literature (cf. Stahlschmidt & Stephen, 2022), and our findings reinforced previous results: our country-level sample showed a detectable difference between WoS-based and Scopus-based citation impact. In particular, in 70% of the sample, a difference in the MNCS measure was detected, which amounted to a 98% higher value, on average, in Scopus. Though this advantage was overestimated due to the highly skewed distribution of differences, the median and 3rd quartile values still confirmed that this advantage is above 14% and 60% for one-half and one-quarter of the corresponding output, respectively. Moreover, neither research fields nor research institutions – as the two aggregation schemes perhaps most relevant for research assessment – made a substantial difference: at least 25% of the papers in each field and HEI was substantially inflated in terms of academic impact (NCS) when shifting from the context of WoS to that of Scopus, which is not a negligable quantity given the size of the output (especially in the content of research assessment).

Having said that, on the existence of impact differences, our research questions rather concerned the causes and factors of those differences, in particular, the role of benchmark differences in the formation of the MNCS measure. Recall that benchmark differences convey various structural differences between databases, so that the "benchmark effect" is essentially a proxy to gauge the influence of those structural differences on impact measurement. The variation of the CBDR measure introduced for the detection of this effect was investigated systematically on the sample in its various aggregations, also to reveal the factors that most plausibly affect the severity of the benchmark effect.

Although the pooled effect of benchmark differences (average CBDR on the entire sample under study) was found to be moderate (CBDR value close to 1), it was also found to be affecting one-half of the country-level output (with differing impacts along databases). Even more striking was the result that the benchmark effect followed a very similar pattern in each aggregation: about 60–70% of the output per all research fields, HEIs, publication years, journal quartiles and disciplinarity categories fell into the group of affected

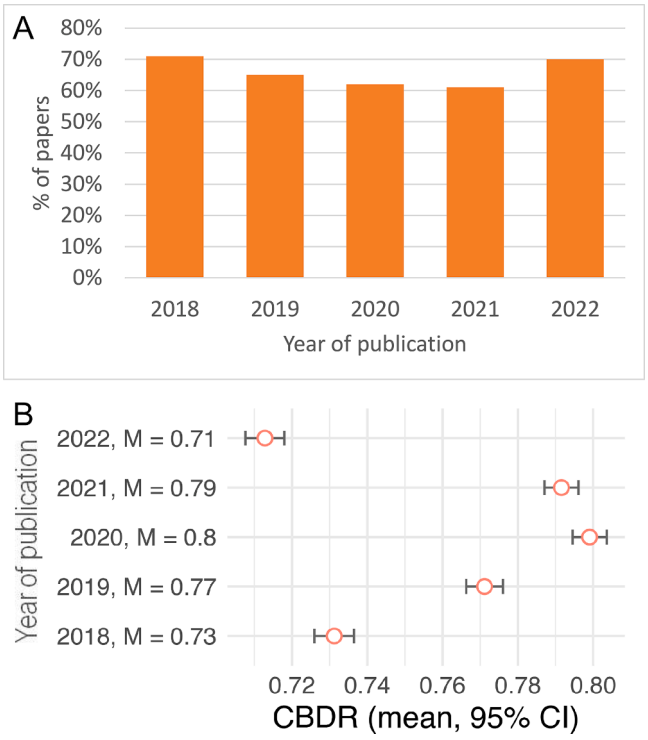

Fig. 7. A: The share of publications with the benchmark effect by publication year B: The mean CBDR value of publications by publication year (with a 95% confidence interval).

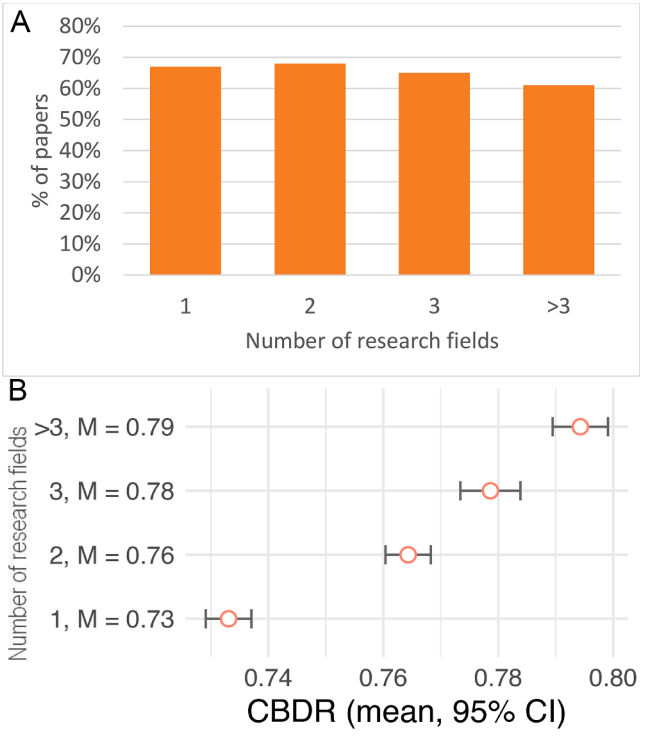

Fig. 8. A: The share of publications with the benchmark effect by the number of research fields B: The mean CBDR value of publications by the number of research fields (with a 95% confidence interval).

papers, with an almost universal mean CBDR range between 0.7—0.8, so that the benchmark difference contributed 20—30% higher to the impact difference than the difference between citation counts in the respective databases. This quasi-universal pattern also led to the (statistical) inference that no real differences existed between units of analysis (and assessment), especially between research fields and HEIs. In other words, the factors most frequently considered in research assessment had no substantive effect (or explanatory power) on the CBDR value of publications. It should be noted that these results were obtained for that part of the country-level publication output where any level of variability could be detected in the NCS value across the two databases (i.e., where CNCI $\neq$ FWCI holds), and the severity of the benchmark effect was studied among the inflicted papers (where CBDR < 1).

Despite the low effect sizes, it was still possible to set up a relevance ranking of the factors (or aggregation schemes) based on the actual effect size value. In particular, we found that HEIs (universities) and journal quartiles exhibited the smallest differences in terms of the benchmark effect (CBDR). One magnitude higher (though still technically small) was the effect of multidisciplinarity, research fields, and publication year, in this order. In fact, the effect of publication years turned out to be quite close to a medium effect size, hence it was worth considering the differences between units (annual output in each year), but also in the case of fields and multidisciplinarity categories as well. Indeed, the tendencies somewhat conflicted with our expectations. Regarding research fields, the largest benchmark effect was exhibited by the "hardest" and "softest" areas, such as the Clinical Sciences and the Humanities, respectively, although we had expected the STEM areas to be more resistant to database switching (in benchmark issues) given the sharper separation of research fields therein (Leydesdorff, Hammarfelt & Salah, 2011). Similarly, our expectation on multidisciplinarity dictated that papers with multiple field assignments would induce greater benchmark differences (regarding these assignments) between WoS and Scopus, due to the structural discrepancies between the two, especially in research categorization. The opposite trend that emerged, i.e., the benchmark effect, showed a reversed relationship with the number of fields: a higher number of field assignments resulted in lower benchmark effects. This gave room for the interesting attempt to explain the trend with a potential compensatory potential of multiple assignments, that is, by assuming that single-field items depend only on one assignment, while for multiple-field items the benchmark differences can "average out" in the interplay of various categories. The relatively most influential factor, the year of publication, also calls for an explanation, as a clear tendency was found with diminishing benchmark effects with increasing recency. We proposed an explanation that when citations start to accrue, the database differences are much less significant both at the paper level and at the level of the research field – that is, the citation ratio and the benchmark ratio are more similar for recent cohorts of papers (hence the CBDR value is closer to 1). With time and the accumulation of citations, the database-specific citation environment becomes more salient, which manifests itself in both the citation counts of papers and the benchmark value, ultimately revealing the characteristic discrepancies between the databases – such as the benchmark effect. Therefore, older papers with a wider citation window exhibit higher levels of the benchmark effect than recent ones – in fact, the minimal three-year citation window, which is often advocated in the context of bibliometric assessments, being the minimum for a responsible citation analysis was reinforced by these results as well. In our case, the annual mean CBDR value for 2018 and 2019 can be considered to meet this requirement. Indeed, after 2019, the CBDR values are invariably higher (aside from last year's outlier value) than in the preceding years, suggesting that characteristic citation levels and metrics can be found where the minimal three-year window is available (i.e., in our example before 2020).

Details aside, perhaps the most striking, overall finding resulted from our study was the evidence that the impact differences for assessment units, or, more precisely, the differences in virtually identical impact metrics, such as the so-called CNCI (WoS) or FWCI (Scopus) that result from the choice of the unerlying database does not exclusively depend on differing citation counts and journal coverage in these databases. More notably, impact metrics are somewhat more sensitive to altering the benchmark component of the underlying measure (NCS) when switching from WoS to Scopus or vice versa, which we call the "benchmark effect". That is, a higher sensitivity can be seen to benchmark differences than to citation count differences (implying that even though papers may have the same citation count in both databases, they will nevertheless exhibit differing database-specific impact). Even more notably, the benchmark differences are the net result of a handful of structural differences between the databases, which include journal coverage, journal/document composition in terms of disciplines, document types, language, publication years, the precision and completeness of references and citations, the field categorization scheme(s) and the actual assignment of documents to categories, the handling of multi-category documents and the benchmark calculation formula (see the *Introduction* section). That is to say, the structural differences between databases seem to have a more profound effect on impact measurement than the number of citations collected in either WoS or Scopus.

Our results are also well in accord with the conclusions of Stahlschmidt and Stephen (2022). In the case of the German publication output compared across the top databases, this study found a relatively small difference between the citation counts of German publications in WoS and Scopus, respectively, but systematically higher normalized impact values in Scopus as compared to WoS in every sector subjected to analysis. The authors also explicitly link this discrepancy to the structural differences of the two databases ("We see […] the macro effect of the databases' structural differences via the differences in normalized citation impact", ibid., p. 2426). Our study can be conceived as a natural extension of this investigation, offering a measurement of these structural differences and their contribution to this discrepancy (applied, in this scenario, to the Hungarian publication output). We can also utilize the metaphor invoked by the authors describing the situation as the databases being "resonance chambers", inducing a unique context for the assessment of publications. An important corollary of their work, reinforced by our results, is that referring to different citation counts and coverage (as being interrelated) across databases is also not valid (or at least sufficient) to account for the impact differences. However, as with the popular concept of Kuhnian incommensurability of scientific paradigms, we can still find room for comparing the impacts across these different contexts: it is exactly the comparison (and characterization) of the contexts against which academic impact is evaluated, along with the comparison of impact metrics themselves. Although it is a more labourous exercise, and would require the exploration of the databases-specific reference sets (e.g. the number and composition of the journals comprising

these sets in both cases), it could shed light on the differing, database-specific "concepts" of research fields, which, to recall Kuhn's famous taxonomic incommensurability thesis (Sankey, 1998), are "extensionally different" categories.

4.1. Practical implications

Based on these findings, we can now more explicitly formulate how, and in what sense, the proposed CBDR measure can inform practical research assessment at least on three levels:

(1) In the era of the commercialization of research assessment tools and a recurring scepticism towards quantitative evaluation, an ever-increasing need is present in academia and science policy for responsible (uses of) assessment metrics (clearly embodied in the most recent international initiative called *Coalition for the Advancement of Research Assessment* or *CoARA*). Responsible use, in turn, calls for the transparency of those metrics, especially the more complex ones. At this general level, the contribution of the CBDR measure is to reveal latent factors – in the form of the benchmark effect – that influence the behavior of the most popular such metric, the "crown" indicator. It also provides a means to explain and quantify this behavior based on the identified factors.
(2) At the level of assessment exercises, when selecting the basis from the supply of potential data sources (e.g., WoS vs. Scopus or both), it is typically a crucial decision, as it provides a tool for a very detailed evaluation of the reliability of academic impact measurement (NCS) over different data sources. For example, as the CBDR can quantify and compare the weight of salient (citation count) differences and of latent structural differences in the outcome measure, it can prevent false impressions on reliability fuelled by (often accidental) observations like "papers in both databases receive the same, or very similar, number of citations". More generally, the measure can be a useful tool in evaluating the implications of, or inform the selection of databases.
(3) Yet another level concerns the management of scientific information, libraries, and other institutional actors responsible for channelling the services of database providers in academia. Negotiating the optimal service portfolios involves certain strategies on the part of the providers to demonstrate the advantages of the promoted service. A typical strategy in the case of citation databases is to highlight the size and the scale of the database, such as the coverage, the total number of journals and publications, or the (total) number of citations offered by the database. In such a scenario, it is imperative that information specialists be equipped with instruments such as the CBDR, which provides evidence that these extensive numbers do not necessarily result in a higher performance value for their institution, at least relative to other sources (Again, the effect of citations can be suppressed by the benchmark effect as witnessed by the CBDR measure.)

In our view, similar considerations apply to the relevance of the CBDR metric for the discourse on university rankings. These considerations are well in accord with the conclusions of Huang et al. (2020), whose study (discussed in the Introduction section) demonstrated the poor robustness of the impact ranking of sample universities when altering the underlying data sources (replacing Scopus with either the Web of Science or the Microsoft Academic). In our case, the CBDR can be used as a tool to study and understand the robustness of rankings with respect to impact measurement, or, since ranking services rely on a selected data source, the implications of the choice among these services. It should also be emphasized that a responsible assessment exercise should be "service agnostic" in the sense of overcoming the biases of individual rankings, such as the database bias in the case of impact assessment. (hence not relying on a single ranking system). The CBDR approach can be utilized to quantify important aspects of those biases for the impact metric, which plays a crucial role and bears a significant effect on ranking scores.

5. Conclusion

This study introduced and validated the CBDR metric, providing a novel method to quantify how database-specific benchmarks cause discrepancies in NCS. Our analysis of Hungary's publication output in WoS and Scopus demonstrates that a significant fraction of country-level data is affected by the benchmark effect. These discrepancies remain consistent across research fields, institutions, journal quartiles, and publication years, suggesting a general structural bias. Furthermore, we demonstrated that differences in coverage and citation counts only partially explain NCS variations between databases. Ultimately, the CBDR metric offers a practical tool for more transparent cross-database evaluations, alerting evaluators that the choice of database inherently shapes normalized results regardless of the unit of assessment.

Our study and results also touch on an important and mostly neglected aspect of the discourse on research evaluation and assessment. A constant agenda and a very popular narrative, especially in the praxis and policy-oriented approaches of evaluation, is that assessment is in need of novel instruments, indicators, and metrics in order to address the more complex and valid mapping of scholarly performance, given its diverse and multifaceted nature. We can, of course, fully agree with such tenets; however, as it unfolds from our investigation, we should also draw attention to the more profound understanding of existing or "traditional" metrics in the disposal of scientometrics. As it turns out, there is still room for investigating those tools in theoretical and practical settings, as they can inform the responsible use of bibliometric indicators by practitioners.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Sándor Soós: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Anna Kiss:** Writing – review & editing, Visualization, Project administration, Data curation. **Zsófia Viktória Vida:** Writing – review & editing, Visualization.

Declaration of competing interest

The authors have nothing to declare.

Acknowledgements

The authors would like to thank the Hungarian Academy of Sciences for its institutional support throughout the research process.

References

Abramo, G., & D'Angelo, C. A (2016a). A farewell to the MNCS and like size-independent indicators. *Journal of Informetrics, 10*(2), 646–651.
Abramo, G., & D'Angelo, C. A (2016b). A farewell to the MNCS and like size-independent indicators: Rejoinder. *Journal of Informetrics, 10*(2), 679–683.
Abramo, G., & D'Angelo, C. A (2016c). A comparison of university performance scores and ranks by MNCS and FSS. *Journal of Informetrics, 10*(4), 889–901.
Bartol, T., Budimir, G., Juznic, P., & Stopar, K. (2016). Mapping and classification of agriculture in Web of Science: Other subject categories and research fields may benefit. *Scientometrics, 109*(2), 979–996.
Haunschild, R., Daniels, A. D., & Bornmann, L. (2022). Scores of a specific field-normalized indicator calculated with different approaches of field-categorization: Are the scores different or similar? *Journal of Informetrics, 16*(1), Article 101241.
Huang, C. K., Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. *Quantitative Science Studies, 1*(2), 445–478.
Leydesdorff, L., Hammarfelt, B., & Salah, A. (2011). The structure of the Arts & Humanities Citation Index: A mapping on the basis of aggregated citations among 1157 journals. *Journal of the American Society for Information Science and Technology, 62*(12), 2414–2426.
Pech, G., & Delgado, C. (2020). Assessing the publication impact using citation data from both Scopus and WoS databases: An approach validated in 15 research fields. *Scientometrics, 125*(2), 909–924.
Purkayastha, A., Palmaro, E., Falk-Krzesinski, H. J., & Baas, J. (2019). Comparison of two article-level, field-independent citation metrics: Field-weighted citation impact (FWCI) and relative citation ratio (RCR). *Journal of informetrics, 13*(2), 635–642.
Robinson-García, 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.
Robinson-Garcia, N., Torres-Salinas, D., Herrera-Viedma, E., & Docampo, D. (2019). Mining university rankings: Publication output and citation impact as their basis. *Research Evaluation, 28*(3), 232–240.
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.
Sankey, H. (1998). Taxonomic incommensurability. *International Studies in the Philosophy of Science, 12*(1), 7–16. https://doi.org/10.1080/02698599808573578
Scelles, N., & Teixeira da Silva, J. A. (2025). Making the impact of publications within a field comparable by improving the field-weighted citation impact (FWCI): The case of sport management. *Scientometrics, 130*(3), 1571–1586.
Smolinsky, L. (2016). Expected number of citations and the crown indicator. *Journal of Informetrics, 10*(1), 43–47.
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.
Tóth, B., Motahari-Nezhad, H., Horseman, N., Berek, L., Kovács, L., Hölgyesi, Á., Péntek, M., Mirjalili, S., Gulácsi, L., & Zrubka, Z. (2024). Ranking resilience: Assessing the impact of scientific performance and the expansion of the Times Higher Education Word University Rankings on the position of Czech, Hungarian, Polish, and Slovak universities. *Scientometrics, 129*(3), 1739–1770.
Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics, 12*(2), 430–435.
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, 2*(1), 20–41.
Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics, 87*(3), 467–481.
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.