E-ISSN 1808-5245

Self-citation in countries: comparisons between the global south and north using the OpenAlex Database

Sandro Barbosa de Oliveira [I]

https://orcid.org/0000-0001-5395-2492

João de Melo Maricato [I]

https://orcid.org/0000-0001-9162-6866

[I] Universidade de Brasília, Brasília, DF, Brazil

Abstract: This paper aims to examine self-citation rates in developed and developing countries, as well as across the Global North and South, and explore their relationship with scientific productivity. Methods: The study utilized data from the OpenAlex database, encompassing 12.3 million articles from 50 countries published between 2020 and 2023. Self-citation rates were calculated at the country and socioeconomic region levels, based on the proportion of articles citing works originating from the same country. Results: Developing countries exhibited higher adjusted self-citation rates and stronger correlations between scientific output and self-citation. In contrast, developed countries showed weaker correlations between productivity and citations, with their citations being less focused on domestic works. While Global South countries account for 42.6% of total scientific output, their contributions represent only 10.4% of references in publications from developed countries, reflecting a concentration of citations among Global North nations. Conclusion: Despite the increasing contribution of developing countries to global scientific output and their growing representation in the OpenAlex database, these countries display high self-citation rates. Moreover, citations from developed countries remain largely concentrated within their own group, underscoring persistent inequalities in the flow of knowledge between the Global North and South.

Keywords: citation analysis; country self-citation; scientific insularism

1 Introduction

Self-citation occurs when an author references their own previously published work in a subsequent article. This practice may be motivated by various factors, including the need to continue a line of research, enhance the visibility and impact of one's studies, and strengthen the coherence and credibility of scientific contributions (Bonzi; Snyder, 1991). However, self-citation can also be strategically employed to artificially inflate personal impact metrics, such as total citation counts and the h-index (Poirrier; Moreno; Huerta-Cánepa, 2021).

The impact of self-citation on scientometric indicators is an important issue, as this practice can introduce biases into scientific research evaluation methods (West; McIlwaine, 2002). Indicators such as the h-index and journal Impact Factor are widely used to assess the quality and influence of scientific outputs; however, self-citation can artificially inflate these metrics, leading to distorted evaluations of researchers' true impact. Foley and Della Sala (2010) further highlight that self-citation can exaggerate the perceived scientific significance of an article or researcher, leading to ongoing discussions within the scientific community about excluding self-citations from citation metrics. Excessive self-citation is currently a significant concern, extensively documented in the scientific literature (Szomszor; Pendlebury; Adams, 2020).

Similarly, self-citation can be observed at various levels of aggregation, including institutions, fields of knowledge, and geographic regions. Szomszor, Pendlebury, and Adams (2020) summarize studies examining self-citation across multiple contexts: individual researchers (especially in relation to the h-index), scientific journals (particularly concerning manipulation of Impact Factors), institutions, and national contexts. Furthermore, research indicates significant variation in self-citation rates among different disciplines, with fields such as computer science and engineering typically exhibiting higher rates compared to social sciences and humanities (Snyder; Bonzi, 1998). These variations may reflect differences in publication practices, innovation cycles, and structural characteristics of the respective scientific communities. Regardless of these contextual differences, self-citation can significantly influence critical decisions related to research funding, academic career progression, and international research collaborations, thereby maintaining inequalities in the recognition and evaluation of scientific research.

Biases associated with self-citation can be further amplified by the geographic origin of articles, as research from certain regions may have a higher tendency toward self-citation practices (Pasterkamp *et al*., 2007). National self-citation, also known as scientific insularism (Ladle; Todd; Malhado, 2012) or domestic citation (Lancho Barrantes *et al.*, 2011), is extensively discussed in scientific literature. Shehatta and Al-Rubaish (2019) assert that self-citation significantly influences a country's perceived academic performance, recommending that self-citations be excluded from bibliometric indicators to achieve a more accurate representation of publication impact at the national level. Conversely, Bardeesi *et al*. (2021) found in their study that self-citation did not significantly affect country rankings in the SCImago Journal & Country Rank (SJR).

Research indicates significant variation in self-citation rates among countries, with developing nations often showing higher rates compared to developed countries (Baccini; Petrovich, 2023; Bakare; Lewison, 2017; Shehatta; Al-Rubaish, 2019). Several factors may contribute to this pattern, such as efforts to enhance the visibility and international impact of local research in a global academic landscape largely dominated by developed countries. For example, despite Brazil accounting for approximately one-third of Latin America's internationally indexed publications, its overall citation performance remains notably lower than that of developed countries (Guimarães, 2004; Packer, 2011).

Ladle, Todd, and Malhado (2012), using data from the Scopus database, classified the phenomenon of self-citation among countries as scientific insularism and proposed a metric based on national self-citation rates adjusted for total scientific output. Their study, covering data from 1996 to 2010, identified high levels of insularity in developing countries such as Brazil, Russia, India, China, and Iran. Among countries with more than 10,000 articles published during the evaluated period, Indonesia, the United Arab Emirates, Ireland, Switzerland, and Austria exhibited the lowest levels of scientific insularism.

Additional studies have utilized various databases to assess self-citation practices. An Iranian study examined the scientific output of 238 countries, classifying them based on self-citation rates using the Journal Citation Reports (JCR) database (Yaminfirooz; Tirgar, 2019). Szomszor, Pendlebury, and Adams (2020) utilized the 2019 Web of Science list of highly cited researchers, highlighting the significant influence certain scientists exert through frequently cited publications. Taşkın *et al*. (2021) analyzed journal self-citation patterns and their impact on journal impact factors using data from the JCR database. Additionally, Baccini, and Petrovich (2023) investigated trends in self-citation across 50 countries from 1996 to 2019, drawing data from the Scopus database.

Multiple factors influence citation and self-citation rates at various aggregation levels, including language and country-specific behaviors. Yitzhaki (1988) conducted an analysis of citation behavior across different languages within the academic context, observing that English-speaking scientists often utilize few materials published in foreign languages, predominantly favoring English-language works. This behavior indicates a language barrier that influences the dissemination and global accessibility of scientific knowledge.

Other country-focused studies have demonstrated that authors from the United States and the United Kingdom publishing in health journals tend to disproportionately cite research originating from their own countries compared to their actual national scientific output. In contrast, these authors cited significantly fewer works from non-American and non-British countries relative to their scientific production (Campbell, 1990). Furthermore, it has been noted that publications resulting from international collaborations tend to have higher self-citation rates compared to those authored within a single nation (Adams, 2013).

Additionally, researchers might preferentially cite national studies due to their relevance in addressing local developmental issues or because these studies examine topics with specific geographic, historical, political, or sociological significance. Poor referencing practices and an excessively nationalistic approach to education within academic institutions may further reinforce this citation pattern (Ladle; Todd; Malhado, 2012).

The aim of this article is to analyze self-citation rates across different countries, specifically distinguishing between developed and developing nations (Global North and South), using a large and publicly accessible database. The study investigates correlations between self-citation rates and scientific productivity indicators at various levels, addressing the following research questions: Are self-citation rates higher in developing countries compared to developed ones when assessed through a more inclusive database? To what extent do countries from the Global North cite research from the Global South? Lastly, is there a discernible tendency toward self-citation within these economic groupings?

2 Material and methods

The data was obtained from the OpenAlex database via the Innovation Systems, Strategies, and Policies Program (InSySPo - Unicamp) at the Instituto de Geociências da Unicamp (https://www.ige.unicamp.br/insyspo/).

OpenAlex is a freely accessible and open-source alternative to traditional scientific databases, notable for its broader coverage, which includes publications in multiple languages and extensive representation of research from the Global South (OpenAlex, 2024). Unlike more restrictive databases such as Scopus and Web of Science, which systematically underrepresent certain disciplines and geographic regions, OpenAlex bypasses licensing barriers, providing a reliable resource for comprehensive, country-level analyses (Alperin *et al*., 2024; Culbert *et al*., 2024).

Using Google Cloud BigQuery, we selected all articles from the OpenAlex database published between 2020 and 2023 that included information on their cited references. The inclusion criteria required articles to specify the country affiliation of the first author. Only articles published by the 50 most productive countries during this period were included in the analysis. Figure 1 illustrates the sampling procedure and inclusion criteria.

Country-level self-citations can be calculated using narrower or broader definitions, depending on how citations involving international publications are treated. Baccini and Petrovich (2023) adopted a broader approach, classifying a citation as a country self-citation whenever there is an overlap between the authors' country affiliations in both the citing and cited publications. In other words, a country self-citation is identified whenever at least one country affiliation is common to both the cited and citing articles.

In this study, we identified the country affiliation based solely on the first author for both the published works and cited references. This choice was motivated by the need to manage the substantial volume of data available in OpenAlex, considering computational limitations. Although this approach represents a limitation, as it disregards international collaborations and may introduce some bias into the results, the impact of this bias is likely reduced since only one-fifth of the articles included in the study involve international collaboration. Additionally, due to the complexities involved in analyzing second-order self-citations, which consider all authors other than the first, we opted for techniques relying exclusively on the first author and their affiliation (Carley; Porter; Youtie, 2019).

The ranking of the 50 most productive countries was determined based on the country affiliation of the first author, with the most productive country ranked first and the least productive ranked 50th. Similarly, the ranking of the most cited/referenced countries in each country's publications also considered the first author's country affiliation, assigning position 1 to the most frequently cited country and continuing sequentially. Countries appearing in the citation ranking but not included among the 50 most productive were assigned position 51 in the production ranking to facilitate correlation analyses.

The Self-Citation Rate (SCR) for each country was calculated by dividing the total number of self-citations – defined as references to publications authored by individuals from the same country as the first author – by the total number of cited references, expressed as a percentage. The Adjusted Self-Citation Rate (ASCR) was determined by subtracting from the country's SCR the percentage corresponding to its contribution to total scientific output during the analyzed period (Ladle; Todd; Malhado, 2012).

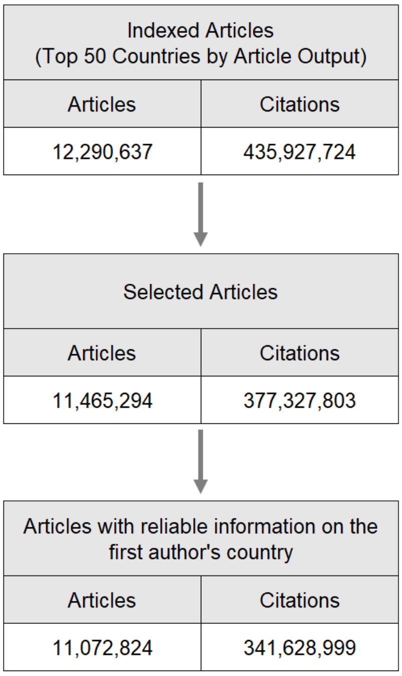
The citation conformity index was determined using Spearman's rank correlation coefficient (rho_{raw}), supplemented by weighted correlation analyses as outlined by Sanatgar, Dolati, and Amini (2020). The weighting criteria applied in the calculations are detailed below:

a) weighted by the number of articles produced ($rho_{weighted\ 1}$): this measure weights the correlation coefficient based on the total number of articles published by the cited country. As a result, citations to countries with higher scientific output have a greater influence on the correlation, reducing the effect of minor mismatches between citation frequency and publication volume;
b) weighted by the publication ranking position ($rho_{weighted\ 2}$): in this approach, weights are assigned according to each country's position in the publication ranking, regardless of the number of articles produced. Citations that closely mirror the ranking of scientific productivity yield a stronger correlation. Unlike the previous method, this weighting focuses solely on ranking alignment, minimizing the impact of small deviations without accounting for production volume;
c) penalized by self-citation ($rho_{weighted\ 3}$): this method assigns double weight to the citing country's position in the publication ranking. A greater discrepancy between a country's rank in production and in citation frequency results in a lower correlation coefficient. The aim is to penalize excessive self-citation, especially when it is not proportionate to the country's actual contribution to global scientific output.

The socio-economic classification of countries into "developed" and "developing" was based on information from the International Monetary Fund (IMF), which considers indicators such as Gross Domestic Product (GDP) per capita, the Human Development Index (HDI), and other macroeconomic factors.

This classification aimed to reflect the prevailing economic and social conditions in each country analyzed. Information on the official languages of the countries was retrieved from Wikipedia (Wikipedia, c2024). The categorization of countries as part of the Global North or Global South followed the framework proposed by Gomes, Maricato, and Costa (2024). Boxes 1 to 3 present the distribution of countries by socio-economic status, Global North/South classification, and official language, respectively. All analyses were conducted using R software, version 4.4.0.

Figure 1 - Sample selection flowchart



Source: Research data.

Figure 1 description: Flowchart illustrating the selection process of articles and their references, beginning with an initial set of over 12 million articles and 435 million references, and narrowing down to approximately 11 million articles and 341 million references. The inclusion criterion was the availability of consistent information on the first author's country of affiliation.

Chart 1 - Country distribution based on the "Developed" and "Developing" classification by the International Monetary Fund (IMF)

Developed		
Australia	Greece	Portugal
Austria	Hungary	Singapore
Belgium	Ireland	South Korea
Canada	Israel	Spain
Czech Republic	Italy	Sweden
Denmark	Japan	Switzerland
Finland	Netherlands	Taiwan
France	Norway	United Kingdom
Germany	Poland	United States
Developing		
Argentina	Indonesia	Russia
Bangladesh	Iran	Saudi Arabia
Brazil	Iraq	South Africa
Chile	Malaysia	Thailand
China	Mexico	Turkey
Colombia	Nigeria	Ukraine
Egypt	Pakistan	Vietnam
India	Romania	

Source: Research data.

Chart 2 - Country distribution based on the Global North/Global South classification (Gomes, Maricato, and Costa, 2024)

Global North		
Australia	Greece	Portugal
Austria	Hungary	Romania
Belgium	Ireland	Singapore
Canada	Israel	South Korea
Czech Republic	Italy	Spain
Denmark	Japan	Sweden
Finland	Netherlands	Switzerland
France	Norway	United Kingdom
Germany	Poland	United States
Global South		
Argentina	Indonesia	Saudi Arabia

Bangladesh	Iran	South Africa
Brazil	Iraq	Taiwan*
Chile	Malaysia	Thailand
China	Mexico	Turkey*
Colombia	Nigeria	Ukraine
Egypt	Pakistan	Vietnam
India	Russia*	

Source: Research data.

* Under discussion.

Chart 3 - Country distribution based on official languages according to Wikipedia (c2024)

English Language			
Australia	Pakistan		
Canada	Singapore		
India	South Africa		
Ireland	United Kingdom		
Nigeria	United States		
Non-English Language			
Argentina	Egypt	Italy	Saudi Arabia
Austria	Finland	Japan	South Korea
Bangladesh	France	Malaysia	Spain
Belgium	Germany	Mexico	Sweden
Brazil	Greece	Netherlands	Switzerland
Chile	Hungary	Norway	Taiwan
China	Indonesia	Poland	Thailand
Colombia	Iran	Portugal	Turkey
Czech Republic	Iraq	Romania	Ukraine
Denmark	Israel	Russia	Vietnam

Source: Research data.

3 Results

A total of 11,072,824 articles were selected from the 50 countries with the highest number of publications between 2020 and 2023. These articles cited a total of 341,628,999 references, resulting in an average of 31 references per publication. Table 1 presents the 50 countries with the highest Adjusted Self-Citation Rates

(ASCR), along with their respective rankings based on the number of articles published. Countries such as Indonesia (ASCR = 52.1%), Ukraine (ASCR = 39.3%), Iraq (ASCR = 30.0%), and Nigeria (ASCR = 27.3%) exhibit notably high self-citation rates, despite holding lower positions in the publication ranking (positions 17, 36, 46, and 42, respectively). This suggests that, although their overall scientific output is relatively modest, authors in these countries make extensive use of national research as references.

On the other hand, countries such as China (ASCR = 39.7%), Brazil (ASCR = 32.8%), and India (ASCR = 28.6%) exhibit high self-citation rates while also holding prominent positions in the scientific publication ranking – first, ninth, and third, respectively. This indicates that, although they frequently cite their own national research, these countries make substantial contributions to global scientific output, securing a prominent role in the international knowledge production landscape.

Table 1 shows that among the 20 countries with the highest Adjusted Self-Citation Rates (ASCR), 18 (90%) belong to the Global South. Conversely, of the 20 countries with the lowest ASCR, 19 (95%) are classified as part of the Global North. The only exception is Taiwan, whose classification as either Global North or South remains a subject of debate (Andersson; Klinthäll, 2012).

Table 1 - Raw (SCR) and adjusted (ASCR) self-citation rate of the fifty countries with the most scientific publications in the OpenAlex database (2020-2023), classified according to the Global North and South criteria

Country	Self-Citation Rate (SCR)	% of Publications	Adjusted Self-Citation Rate (ASCR)	Publication Ranking
Indonesia[2]	53.4	1.4	52.1	17
Russia[2]	46.5	2.8	43.7	8
China[2]	63.8	24.1	39.7	1
Ukraine[2]	39.7	0.5	39.3	36
Brazil[2]	35.5	2.7	32.8	9
Iraq[2]	30.3	0.3	30.0	46
India[2]	34.1	5.5	28.6	3
Nigeria[2]	27.7	0.3	27.3	42

Country	Self-Citation Rate (SCR)	% of Publications	Adjusted Self-Citation Rate (ASCR)	Publication Ranking
Iran[2]	26.3	2.0	24.3	16
Pakistan[2]	24.0	0.7	23.3	26
Egypt[2]	23.9	0.7	23.2	23
Bangladesh[2]	22.8	0.2	22.6	50
Poland[1]	23.1	1.2	21.8	18
Colombia[2]	21.9	0.3	21.6	43
Turkey[2]	23.1	2.1	21.0	14
South Africa[2]	20.8	0.5	20.3	30
Mexico[2]	20.0	0.7	19.3	25
Argentina[2]	19.4	0.3	19.1	44
Vietnam[2]	18.7	0.3	18.4	47
Romania[1]	18.7	0.4	18.4	41
Japan[1]	21.9	3.7	18.3	5
Italy[1]	20.7	2.9	17.8	7
Chile[2]	17.4	0.3	17.2	48
Saudi Arabia[2]	17.2	0.7	16.6	27
Thailand[2]	16.5	0.5	16.0	37
Czech Republic[1]	15.8	0.4	15.4	38
United States[1]	31.6	16.4	15.3	2
Spain[1]	17.3	2.3	15.0	12
Malaysia[2]	15.6	0.7	14.9	24
Hungary[1]	14.9	0.3	14.6	49
South Korea[1]	16.7	2.5	14.3	10
Germany[1]	17.4	3.6	13.7	6
Portugal[1]	14.2	0.6	13.6	28
Australia[1]	15.2	2.1	13.1	15
Greece[1]	12.8	0.5	12.3	34
Taiwan[2]	12.5	1.0	11.5	20
France[1]	13.6	2.2	11.4	13
Norway[1]	11.7	0.5	11.3	33
Canada[1]	12.9	2.4	10.6	11
Finland[1]	10.9	0.4	10.5	40
United Kingdom[1]	14.0	3.9	10.2	4
Denmark[1]	10.6	0.5	10.1	31
Sweden[1]	10.5	0.8	9.8	22
Israel[1]	10.1	0.5	9.6	32
Austria[1]	9.0	0.5	8.5	35
Netherlands[1]	9.6	1.1	8.5	19

Country	Self-Citation Rate (SCR)	% of Publications	Adjusted Self-Citation Rate (ASCR)	Publication Ranking
Ireland[1]	8.6	0.3	8.3	45
Belgium[1]	8.2	0.6	7.6	29
Switzerland[1]	8.0	0.8	7.2	21
Singapore[1]	6.0	0.4	5.6	39

Source: Research data.

Country classification: [1] Global North; [2] Global South.

Overall, there is a moderate to strong correlation between the citation ranking and the ranking based on the total number of papers produced – that is, countries tend to cite publications from other countries in proportion to their prominence in global scientific output. Notably, countries with high self-citation rates also exhibit the strongest alignment between citation and production rankings. This is particularly evident in Brazil ($rho_{\text{weighted 3}} = 0.82$), China ($rho_{\text{weighted 3}} = 0.83$), India ($rho_{\text{weighted 3}} = 0.89$), and Russia ($rho_{\text{weighted 3}} = 0.90$), as shown in Table 2.

On the other hand, countries with low self-citation rates – such as the Netherlands, Ireland, and Sweden – tend to cite countries that are somewhat less aligned with the global ranking of scientific productivity, resulting in lower correlation values. The weighted correlation penalized for self-citation ($rho_{\text{weighted 3}}$) in these countries ranged from 0.63 to 0.70 (Table 2). This suggests a tendency to overlook contributions from countries with a significant share in global scientific output.

Table 2 also reveals that the majority of countries (20 out of 23) with a weighted correlation ($rho_{\text{weighted 3}}$) below 0.80 – between the citation ranking and the ranking of total scientific output – belong to the Global North. Only seven Global North countries exhibited a self-citation-penalized weighted correlation equal to or above 0.80: Czech Republic, Greece, South Korea, Spain, Italy, Romania, and Poland. This suggests that authors from Global North countries do not consistently reference researchers from the most prolific nations in terms of scientific production. In contrast, Global South countries tend to cite authors from the highest-producing countries, even while maintaining elevated self-citation rates.

Table 2 - Correlations* between the ranking of cited countries and the ranking of scientific output, ordered by self-citation-weighted correlation (2020-2023)

Country	rho_{raw}	$rho_{weighted\ 1}$	$rho_{weighted\ 2}$	$rho_{weighted\ 3}$
Netherlands[1]	0.63	0.67	0.71	0.63
Ireland[1]	0.66	0.76	0.78	0.69
Sweden[1]	0.70	0.67	0.71	0.70
Australia[1]	0.70	0.72	0.77	0.70
Norway[1]	0.69	0.62	0.67	0.70
Denmark[1]	0.69	0.67	0.71	0.71
Finland[1]	0.69	0.68	0.72	0.71
Switzerland[1]	0.71	0.68	0.72	0.71
United Kingdom[1]	0.73	0.71	0.76	0.72
Chile[2]	0.71	0.81	0.84	0.74
Belgium[1]	0.73	0.69	0.72	0.74
Singapore[1]	0.72	0.82	0.85	0.74
Germany[1]	0.75	0.71	0.75	0.75
Austria[1]	0.73	0.69	0.73	0.75
United States[1]	0.76	0.76	0.80	0.75
Hungary[1]	0.72	0.81	0.85	0.75
Japan[1]	0.77	0.78	0.82	0.76
Canada[1]	0.77	0.72	0.76	0.77
Colombia[2]	0.74	0.83	0.86	0.77
Israel[1]	0.76	0.70	0.74	0.78
France[1]	0.78	0.67	0.73	0.78
Argentina[2]	0.75	0.80	0.84	0.78
Portugal[1]	0.78	0.80	0.84	0.79
Czech Republic[1]	0.77	0.81	0.85	0.80
Greece[1]	0.78	0.81	0.84	0.80
Nigeria[2]	0.78	0.82	0.87	0.81
South Korea[1]	0.81	0.84	0.88	0.81
South Africa[2]	0.79	0.82	0.87	0.81
Taiwan[2]	0.81	0.84	0.88	0.82
Brazil[2]	0.82	0.78	0.84	0.82
Spain[1]	0.82	0.72	0.78	0.82
Saudi Arabia[2]	0.82	0.84	0.89	0.83
China[2]	0.83	0.95	0.95	0.83
Italy[1]	0.83	0.73	0.77	0.83
Iraq[2]	0.80	0.83	0.89	0.84
Mexico[2]	0.83	0.83	0.87	0.84

Country	*rho* $_{raw}$	*rho* $_{weighted\ 1}$	*rho* $_{weighted\ 2}$	*rho* $_{weighted\ 3}$
Vietnam[2]	0.81	0.84	0.90	0.84
Bangladesh[2]	0.81	0.85	0.90	0.85
Romania[1]	0.83	0.85	0.89	0.86
Ukraine[2]	0.83	0.85	0.89	0.86
Malaysia[2]	0.86	0.84	0.89	0.86
Thailand[2]	0.85	0.86	0.91	0.87
Pakistan[2]	0.86	0.84	0.90	0.87
Poland[1]	0.87	0.81	0.86	0.87
Indonesia[2]	0.87	0.79	0.86	0.87
Turkey[2]	0.88	0.80	0.87	0.88
Iran[2]	0.88	0.85	0.91	0.88
Egypt[2]	0.88	0.85	0.91	0.89
India[2]	0.90	0.82	0.87	0.89
Russia[2]	0.90	0.81	0.86	0.90

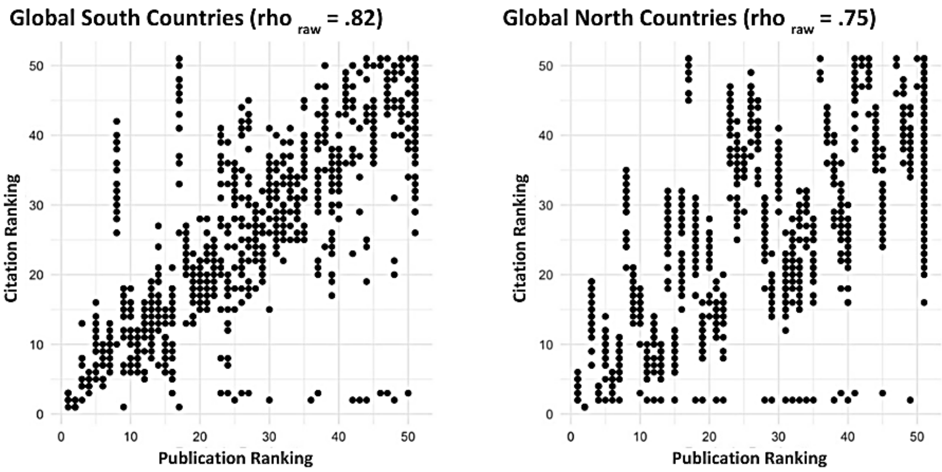
Source: Research data.

Country classification: [1] Global North; [2] Global South

* Spearman's correlation (*rho* $_{raw}$); correlation weighted by the number of articles produced (*rho* $_{weighted\ 1}$); correlation weighted by the publication ranking (*rho* $_{weighted\ 2}$); and correlation penalized for self-citation (*rho* $_{weighted\ 3}$).

This pattern is illustrated in Figure 2, which demonstrates that the citation ranking is more closely aligned with the production ranking in Global South countries than in those of the Global North. Using the raw correlation between scientific production and citation rankings (*rho* $_{raw}$), the plot for Global South countries shows points distributed along a trajectory close to the bisector of the X and Y axes, indicating a proportional relationship between scientific output and citation behavior. In contrast, the plot for Global North countries displays a more scattered distribution, lacking a clear pattern or alignment, which reflects a weaker correlation between production and citation rankings.

Figure 2 - Correlation Between Publication and Citation Rankings: Global North and Global South Countries (2020-2023)

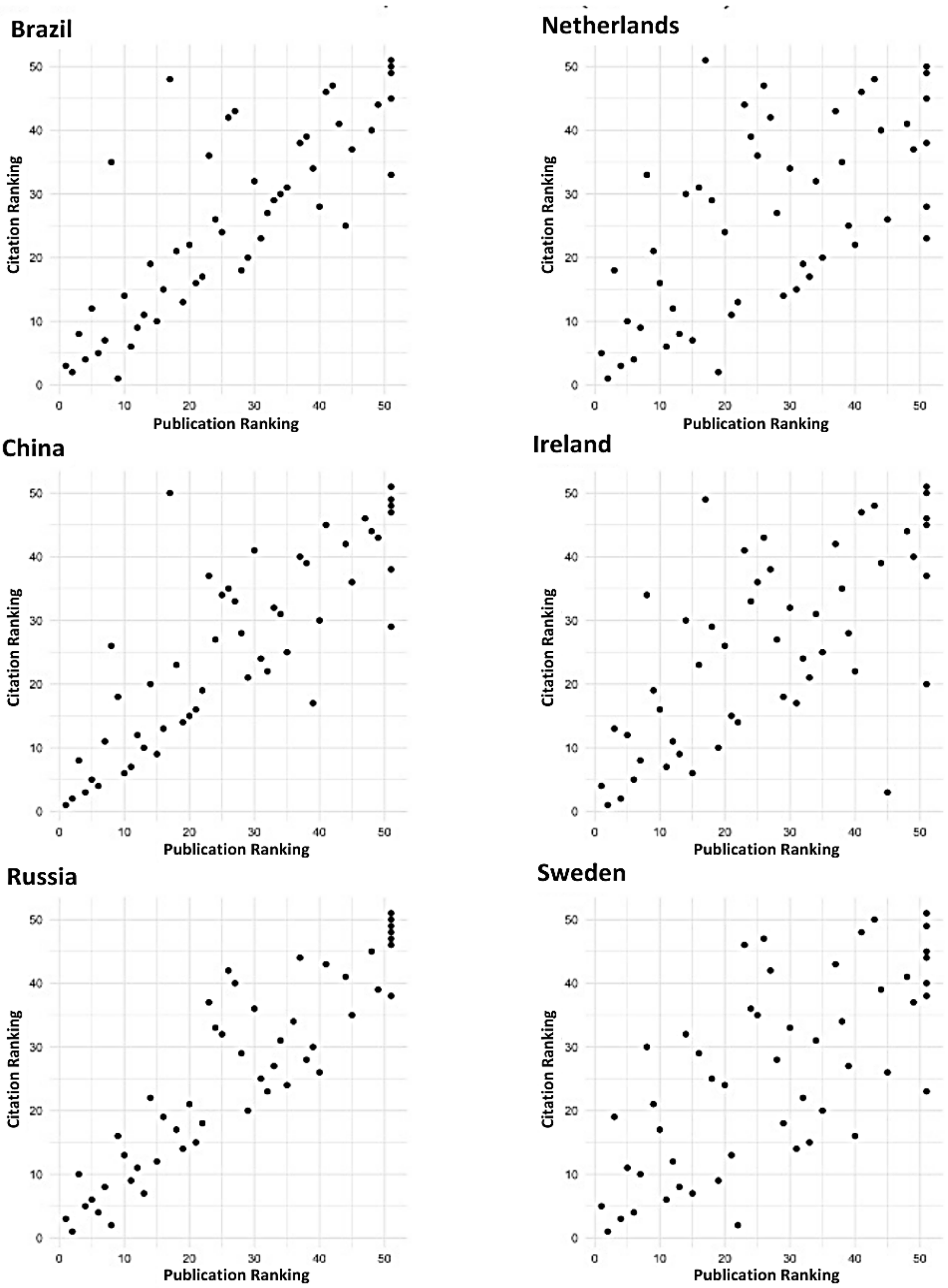


Source: Research data.

Figure 2 Description: This figure presents two scatter plots. In both diagrams, the X-axis represents each country's position in the article publication ranking, while the Y-axis reflects its position in the citation ranking. The left plot illustrates the Global South, where data points are more closely clustered along the diagonal (bisector), indicating a stronger correlation between publication volume and citation frequency. In contrast, the right plot represents the Global North, with points more widely dispersed, suggesting a weaker alignment between publication and citation rankings.

As an example, Brazil, China, and Russia tend to cite authors from countries that make significant contributions to global scientific output, as illustrated in Figure 3. Visually, the data points for these countries are more closely aligned with the bisector of the X and Y axes, indicating a stronger correlation between publication and citation rankings. In contrast, the data points for the Netherlands, Ireland, and Sweden appear more dispersed, suggesting a weaker alignment. This visual analysis reinforces the observation that a stronger correlation between publication and citation rankings is more evident in countries with higher self-citation rates.

Figure 3 - Examples of Correlation Between Publication and Citation Rankings: Brazil, Netherlands, China, Ireland, Russia, and Sweden (2020-2023)



Source: Research data.

Figure 3 Description: This figure displays six scatter plots, each with the X-axis representing a country's position in the article publication ranking and the Y-axis representing its position in the citation ranking. The plots are arranged in two columns. The three diagrams on the left illustrate the behavior of Global South countries (Brazil, China, and Russia), where the data points tend to align more closely with the axis bisector. In contrast, the three diagrams on the right represent Global North countries (Netherlands, Ireland, and Sweden), where the points are more widely dispersed, indicating weaker alignment between publication and citation rankings.

When evaluating the socio-economic characteristics of countries, it becomes evident that Global North nations exhibit the lowest self-citation rates. In contrast, countries in the Global South – despite having higher self-citation rates – demonstrate stronger alignment between their citation patterns and the article publication ranking ($rho_{\text{weighted 3}} = 0.84$). Similar patterns are observed when comparing Developed versus Developing countries, as well as English-speaking versus non-English-speaking nations (Table 3).

Table 3 - Adjusted Self-Citation Rate and Correlations Between Citation and Publication Rankings by Socio-Economic Characteristics of Countries (2020-2023)

Socio-economic characteristics	N	Median		
		Adjusted Self-Citation Rate (ASCR)	rho_{raw}	$rho_{weighted\ 3}$
Global Classification				
Global South	20	22.9	0.83	0.84
Global North	27	11.4	0.76	0.75
Russia, Turkey, and Taiwan	3	21.0	0.88	0.88
Economic Development Status				
Developing	23	22.6	0.83	0.85
Developed	27	11.4	0.76	0.75
Official language				
English	10	14.2	0.77	0.76
Non-English	40	16.3	0.81	0.81

Source: Research data.

This divergence between the ranking of the most productive countries and the ranking of the most cited countries in publications from the Global North is confirmed by the data presented in Table 4. While the Global North was responsible for 52% of total scientific output, only 12% of its citations referenced work from other regions. In contrast, the Global South, which accounted for 43% of global output, cited research from its own region in 44% of cases and from the Global North in 54%.

These data reveal an imbalance in citation distribution and knowledge exchange across global regions. The Global North appears less influenced by the scientific contributions of the Global South, whereas the reverse is not true. The Global South demonstrates greater interdependence, referencing both its own work and that of the Global North in a more balanced proportion.

Table 4 - Publication and Citation Distribution by Geopolitical and Socio-Economic Groupings (2020–2023)

Region of Publication Origin	Publications		Region of Citation Origin	Citations	
	N	%		N	%
Global North	5,703,339	51.5%	Global North	155,707,334	88.3%
			Global South	18,260,119	10.4%
			Russia, Turkey and Taiwan	2,426,213	1.4%
Global South	4,714,068	42.6%	Global North	77,927,530	54.0%
			Global South	62,888,383	43.6%
			Russia, Turkey and Taiwan	3,497,050	2.4%
Russia, Turkey, and Taiwan	655,417	5.9%	Global North	8,483,443	61.9%
			Global South	2,866,052	20.9%
			Russia, Turkey and Taiwan	2,344,889	17.1%

Source: Research data.

4 Discussion

This study analyzed self-citation rates in developed and developing countries, examining the relationship between self-citation and each country's share in international scientific output, both overall and in relation to socio-economic characteristics.

Self-citation is a concerning phenomenon due to its potential to distort bibliometric indicators. Mutti (2023) introduced the term citeflation to describe the practice of excessive citation – not necessarily based on intellectual merit or scientific contribution – but rather as a strategy to boost visibility and impact metrics, ultimately skewing the perceived value of research or institutions. In this study, the concept of self-citation was applied at the national level to investigate how it manifests within the broader global scientific landscape.

Using data from the OpenAlex database, our findings align with existing literature highlighting high self-citation rates in countries such as Indonesia, China, Russia, Ukraine, Brazil, India, Nigeria, Iran, Iraq, and Egypt. Baccini and Petrovich (2023), in their analysis of self-citation trends across 50 countries from

1996 to 2019, observed a general decline in self-citation rates -both at the national level and among individual researchers – with the exception of Indonesia, Ukraine, and Russia, which displayed trends contrary to those of the other countries.

Yaminfirooz and Tirgar (2019) found that although Iran ranks sixth in scientific production in Asia, it holds the second-highest self-citation rate on the continent alongside India, Malaysia, and Pakistan. In their study on what they termed "scientific insularism", Ladle, Todd, and Malhado (2012) also highlighted elevated self-citation rates in developing countries such as Brazil, Russia, India, and China.

However, our study shows that although developing countries such as Brazil, China, and India occupy prominent positions in article publication rankings, their work is seldom cited by developed nations. The findings indicate that while developing countries actively engage with and cite scientific output from the Global North, this exchange is not reciprocated. Ladle, Todd, and Malhado (2012) pointed to the use of non-English languages in many developing countries as a key barrier contributing to scientific insularity. Similarly, as early as the 2000s, Guimarães (2004) noted that the significant growth in Brazilian scientific output over recent decades had not been accompanied by comparable performance in qualitative indicators, such as the percentage of cited articles and impact index.

Thus, although developing countries have expanded their participation in global scientific production – as evidenced by data from OpenAlex – citations from developed countries remain limited. This trend reinforces the persistent inequality in the flow of scientific knowledge between the Global North and South. For example, the performance of Brazilian journals, measured by the average number of citations per article, remains lower than that of journals from developed countries. This disparity can be attributed to factors such as language barriers and limited international collaboration (Packer, 2011).

Vasconcelos (2008), in her doctoral thesis, had already identified the language barrier as a significant challenge. Her work emphasized that science functions within a multicultural and multilingual environment, which strongly influences the processes of research communication and publication. This dynamic has created a notable intersection between scientometrics and linguistics, particularly due to the dominance of the English language in major scientific databases. This presents a disadvantage for countries whose official language is not English – a common feature of many Global South nations. Vasconcelos (2008) also noted that, in addition to the language barrier, funding for translation and publication services in these countries is often scarce. Consequently, despite the substantial growth in publications by researchers from the Global South, linguistic constraints may limit the visibility and citation of their work by scholars in the Global North.

Overall, we observed a moderate to strong correlation between the total number of citations and the total number of articles produced ($0.63 \leq rho_{\text{weighted 3}} \leq 0.90$) among the 50 countries analyzed in this study. Similar findings were reported by Bardeesi *et al*. (2021), who found that the self-citation rates of countries in the field of clinical neurology (1996-2019) were significantly correlated with the total number of citable documents, total citations, citation rankings, and global citation standings. Notably, in our study, the strongest correlations were found among developing countries. This suggests that the scientific output of these nations tends to rely more heavily on the work produced by the most prolific countries, closely reflecting their representativeness in terms of publication volume.

Although our study indicates that the percentage of self-citation in developing countries with substantial contributions to international scientific output is not inherently problematic, it remains important to consider the broader implications of self-citation on bibliometric indicators. Shehatta and Al-Rubaish (2019) emphasize that country-level self-citation significantly influences both the total number of citations and the average citations per article, potentially distorting the perceived impact of research. This underscores the need for adjusted metrics to provide a more accurate and equitable evaluation of global scientific output.

On the other hand, Bardeesi *et al*. (2021) demonstrated that country-level self-citation likely does not have a significant impact on the scientific ranking of the top 50 countries in the field of clinical neurology. In other words, their findings did not support arguments for excluding or adjusting self-citations in citation-based metrics.

To more accurately assess self-citation rates, it is essential to consider differences across scientific disciplines. Szomszor, Pendlebury, and Adams (2020) reported substantial variability among fields, noting particularly high self-citation rates in Mathematics and Space Sciences compared to other areas. Similarly, Giordano *et al*. (2021) found that the Revista Brasileira de Ortopedia has a low self-citation rate, suggesting that Brazilian orthopedists rarely cite peers publishing in the same journal. Based on these findings, the authors emphasize the importance of adopting strategies to enhance the journal's impact factor on the global stage.

This study has some limitations. First, the decision to define a researcher's nationality based on the affiliation of the first author may obscure self-citation at the country level in the context of international collaboration. However, this potential bias is likely minimized, as only about one-fifth of the articles analyzed involved co-authorship between authors from different countries. Another limitation concerns the use of an open-access database. Nevertheless, our findings align with those of previous studies that relied on more established bibliometric sources. Regarding scientific output, it is important to note that this study did not assess the quality or impact factor of the journals in which Global South countries publish. In other words, while these countries contribute significantly to global scientific production, their output may be concentrated in journals with lower scientific impact. However, Moher *et al*. (2017), in an analysis of nearly 2,000 biomedical articles published in approximately 200 journals classified as predatory, found that predatory publishing is not limited to the Global South. In fact, half of the articles published in these journals had authors affiliated with middle- and high-income countries, including the United States and Japan. Finally, this study focused exclusively on the 50 most productive countries in the

OpenAlex database between 2020 and 2023. Therefore, future research including a broader range of countries is needed to expand the analysis of the self-citation phenomenon in a more comprehensive global context.

The results of this study not only support existing literature on the prevalence of high self-citation rates in developing countries but also underscore a significant imbalance in the global exchange of scientific knowledge. While some potential causes for this disparity have been discussed, further research is essential to more deeply investigate the barriers that prevent the Global North from referencing scientific output from the Global South in their own research.

5 Conclusion

This study explores the complex relationship between self-citation, scientific output, and the unequal flow of knowledge between developed and developing countries. It is the first to employ the free and open-access OpenAlex database for this purpose, yielding results consistent with those of previous studies based on traditional bibliometric sources.

While OpenAlex offers a comprehensive and accessible alternative, all prior studies referenced in this article have relied on more established databases such as Scopus, Web of Science, and SCImago. The alignment between our findings and those of earlier research reinforces the reliability and robustness of data retrieved from OpenAlex.

Despite the growing participation of developing countries in global scientific production, citations by developed countries remain limited – highlighting a persistent imbalance in the international exchange of scientific knowledge. It is therefore essential to identify the factors that hinder the inclusion of Global South research in the citation practices of developed countries. Doing so would help inform strategies aimed at promoting greater integration and equity in global scientific discourse.

Acknowledgments

The authors would like to acknowledge the participation of Alysson Mazoni for his technical support in accessing and working with the OpenAlex database.

References

ADAMS, J. The fourth age of research. **Nature**, New York, v. 497, n. 7451, p. 557-560, 2013. Available at: https://doi.org/10.1038/497557a. Access: Dec.18 2024.

ALPERIN, J. P.; PORTENOY, J.; DEMES, K.; LARIVIÈRE, V.; HAUSTEIN, S. An analysis of the suitability of OpenAlex for bibliometric analyses. **arXiv preprint**, Ithaca, v. 1, 26 Apr. 2024. Available at: https://doi.org/10.48550/arXiv.2404.17663. Access: Dec.18 2024.

ANDERSSON, M; KLINTHÄLL, M. The opening of the North-South divide: cumulative causation, household income disparity and the regional bonus in Taiwan 1976-2005. **Structural Change and Economic Dynamics**, Amsterdam, v. 23, n. 2, p. 170-179, 2012. Available at: https://doi.org/10.1016/j.strueco.2012.02.001. Access: Dec.18 2024.

BACCINI, A.; PETROVICH, E. A global exploratory comparison of country self-citations 1996-2019. **PLoS ONE**, California, v. 18, n. 12, p. 1-33, 2023. Available at: https://doi.org/10.1371/journal.pone.0294669. Access: Dec.18 2024.

BAKARE, V.; LEWISON, G. Country over-citation ratios. **Scientometrics**, New York, v. 113, n. 2, p. 1199-1207, 2017. Available at: https://doi.org/10.1007/s11192-017-2490-z. Access: Dec.18 2024.

BARDEESI, A. M.; JAMJOOM, A. A. B.; SHARAB, M. A.; JAMJOOM, A. B. Impact of country self citation on the ranking of the top 50 countries in clinical neurology. **eNeurologicalSci**, Amasterdam, v. 23, p. 100333, 2021. Available at: https://doi.org/10.1016/j.ensci.2021.100333. Access: Dec.18 2024.

BONZI, S.; SNYDER, H. W. Motivations for citation: A comparison of self citation and citation to others. **Scientometrics**, New York, v. 21, n. 2, p. 245-254, 1991. Available at: https://doi.org/10.1007/BF02017571. Access: Dec.18 2024.

CARLEY, S. F.; PORTER, A. L.; YOUTIE, J. L. A multi-match approach to the author uncertainty problem. **Journal of Data and Information Science**, Boston, v. 4, n. 2, p. 1-18, 2019. Available at: https://doi.org/10.2478/jdis-2019-0006. Access: Dec.18 2024.

CAMPBELL, F. M. National bias: a comparison of citation practices by health professionals. **Bulletin of the Medical Library Association**, Chicago, v. 78, n. 4, p. 376, 1990.

CULBERT, J.; HOBERT, A.; JAHN, N.; HAUPKA, N.; SCHMIDT, M.; DONNER, P.; MAYR, P. Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. **arXiv preprint**, Ithaca, v. 3, 1 Nov. 2024. Available at: https://doi.org/10.48550/arXiv.2401.16359. Access: Dec.18 2024.

FOLEY, J. A.; DELLA SALA, S. The impact of self-citation. **Cortex**, Edinburgh, v. 46, n. 6, p. 802-810, 2010. Available at: https://doi.org/ 10.1016/j.cortex.2010.01.004. Access: Dec.18 2024.

GIORDANO, V.; LYRA, J.; BONADIMAN, J. A.; LECH, O. Brazilian authors don't cite brazilian authors: nothing has changed since 1994. **Revista Brasileira de Ortopedia**, São Paulo, v. 56, n. 2, p. 154, 2021. Available at: https://doi.org/10.1055/s-0041-1728702. Access: Dec.18 2024.

GOMES, G. B.; MARICATO, J. M.; COSTA, M. P. Transição de periódicos para o acesso aberto com Article Processing Charges (APC): implicações no contexto do sul global. *In*: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 9., 2024. **Anais** [...]. Brasília: UNB/IBICT, 2024. p. 1-9.

GUIMARÃES, J. A. A pesquisa médica e biomédica no Brasil: comparações com o desempenho científico brasileiro e mundial. **Ciência & Saúde Coletiva**, Rio de Janeiro, v. 9, p. 303-327, 2004. Available at: https://doi.org/10.1590/S1413-81232004000200009. Access: Dec.18 2024.

LADLE, R. J.; TODD, P. A.; MALHADO, A. C. M. Assessing insularity in global science. **Scientometrics**, New York, v. 93, n. 3, p. 745-750, 2012. Available at: https://doi.org/10.1007/s11192-012-0703-z. Access: Dec.18 2024.

LANCHO BARRANTES, B. S. *et al.* Citation flows in the zones of influence of scientific collaborations. **Journal of the American Society for Information Science and Technology**, New Jersey, v. 63, n. 3, p. 481-489, 2011. Available at: https://doi.org/10.1002/asi.21682. Access: Dec.18 2024.

MOHER, D. *et al*. Stop this waste of people, animals and money. **Nature**, New York, v. 549, p. 23-25, 2017. Available at: https://doi.org/10.1038/549023a. Access: Dec.18 2024.

MUTTI, A. Pride and concern for bibliometric achievements: deserved results or result of cites inflation? **La Medicina del Lavoro**, Parma, v. 114, n. 4, p. 1-2, 2023. Available at: https://doi.org/10.1038/549023a. Access: Dec.18 2024.

OPENALEX. **The open catalog to the global research system**. [*S.l.*: *s.n.*], 2024.

PACKER, A. L. Os periódicos brasileiros e a comunicação da pesquisa nacional. **Revista USP**, São Paulo, n. 89, p. 26-61, 2011. Available at: https://doi.org/10.11606/issn.2316-9036.v0i89p26-61. Access: Dec.18 2024.

PASTERKAMP, G.; ROTMANS, J.; DE KLEIJN, D.; BORST, C. Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. **Scientometrics**, New York, v. 70, n. 1, p. 153-165, 2007. Available at: https://doi.org/10.1007/s11192-007-0109-5. Access: Dec.18 2024.

POIRRIER, M.; MORENO, S.; HUERTA-CÁNEPA, G. Robust h-index. **Scientometrics**, New York, v. 126, n. 3, p. 1969-1981, 2021. Available at: https://doi.org/10.1007/s11192-020-03857-z. Access: Dec.18 2024.

SANATGAR, M; DOLATI, A; AMINI, M. A General class of weighted rank correlation measures. **arXiv preprint**, Ithaca, 21 Jan. 2020. Available at: https://doi.org/10.48550/arXiv.2001.07298 -z. Access: Dec.18 2024.

SHEHATTA, I.; AL-RUBAISH, A. M. Impact of country self-citations on bibliometric indicators and ranking of most productive countries. **Scientometrics**, New York, v. 120, n. 2, p. 775-791, 2019. Available at: https://doi.org/10.1007/s11192-019-03139-3. Access: Dec.18 2024.

SNYDER, H.; BONZI, S. Patterns of self-citation across disciplines (1980-1989). **Journal of Information Science**, Thousand Oaks, v. 24, n. 6, p. 431-435, 1998. Available at: https://doi.org/10.1177/016555159802400606. Access: Dec.18 2024.

SZOMSZOR, M.; PENDLEBURY, D. A.; ADAMS, J. How much is too much? The difference between research influence and self-citation excess. **Scientometrics**, New York, v. 123, n. 2, p. 1119-1147, 2020. Available at: https://doi.org/10.1007/s11192-020-03417-5. Access: Dec.18 2024.

TAŞKIN, Z.; DOĞAN, G.; KULCZYCKI, E.; ZUCCALA, A. A. Self-citation patterns of journals indexed in the journal citation reports. **Journal of Informetrics**, Amsterdam, v. 15, n. 4, p. 101221, 2021. Available at: https://doi.org/10.1016/j.joi.2021.101221. Access: Dec.18 2024.

VASCONCELOS, S. M. R. **Ciência no Brasil: uma abordagem cienciométrica e lingüística**. 2008. Tese (Doutorado em Bioquímica) - Programa de Pós-Graduação em Química Biológica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

WEST, R.; MCILWAINE, A. What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality. **Addiction**, New Jersey, v. 97, n. 5, p. 501-504, 2002. Available at: https://doi.org/10.1046/j.1360-0443.2002.00104.x. Access: Dec.18 2024.

WIKIPÉDIA, A ENCICLOPÉDIA LIVRE. San Francisco: Wikimedia Foundation, c2024.

YAMINFIROOZ, M.; TIRGAR, A. Self-citation in Iran in comparison with other countries. **Acta Informatica Medica**, Bosnia, v. 27, n. 4, p. 259, 2019. Available at: https://doi.org/ 10.5455/aim.2019.27.259-262. Access: Dec.18 2024.

YITZHAKI, M. The language barrier in the humanities: Measures of language self-citation and self-derivation-the case of biblical studies. **Informetrics**, Amsterdam, v. 87/88, p. 301-313, 1988.

Autocitação em nível de países: comparações entre o sul e o norte global utilizando a base OpenAlex

Resumo: Este estudo tem como objetivo analisar as taxas de autocitação em países desenvolvidos e em desenvolvimento, do Norte e Sul Global, investigando sua relação com a produtividade científica. Método: Foram utilizados dados da base OpenAlex, com 12,3 milhões de artigos de 50 países publicados entre 2020 e 2023. A taxa de autocitação foi calculada em nível de país e região socioeconômica, com base na proporção de artigos que referenciam trabalhos do mesmo país. Resultado: Países em desenvolvimento apresentaram maiores taxas de autocitação ajustada e maior correlação entre produção e autocitação. Em contrapartida, os países desenvolvidos mostraram correlações mais fracas entre produção e citação, ainda que suas citações sejam menos centradas no próprio país. Apesar de os países do Sul Global representarem 42,6% da produção científica total, sua participação nas referências dos países desenvolvidos é de apenas 10,4%, refletindo a concentração das citações entre os países do Norte. Conclusão: A ocorrência de altas taxas de autocitação nos países em desenvolvimento é acompanhada pelo aumento da participação desses países na produção científica global. Ainda assim, as citações provenientes de países desenvolvidos permanecem predominantemente concentradas entre eles, evidenciando uma desigualdade persistente no fluxo de conhecimento entre o Norte e o Sul Global.

Palavras-chave: análise de citação; autocitação de países; insularidade científica; produção científica; desigualdades na ciência

Authorship and Responsibility Statement

Conception and design of the study: Sandro Barbosa de Oliveira, João de Melo
Data collection: Sandro Barbosa de Oliveira
Data analysis and interpretation: Sandro Barbosa de Oliveira, João de Melo
Writing: Sandro Barbosa de Oliveira, João de Melo
Critical review of the manuscript: Sandro Barbosa de Oliveira, João de Melo

Data availability statement

The dataset supporting the results of this study can be requested from the authors (see corresponding author).

Corresponding authorship

Sandro Barbosa de Oliveira

sandrobarboliveira@gmail.com

Editor-in-chief

Thiago Henrique Bragato Barros

How to cite

OLIVEIRA, Sandro Barbosa; MARICATO, João de Melo. Self-citation in countries: comparisons between the global south and north using the OpenAlex Database. **Em Questão**, Porto Alegre, v. 31, e-144735, 2025. https://doi.org/10.1590/1808-5245.31.144735

Open peer review:

https://doi.org/10.1590/1808-5245.31.144735A

Receveid: 12/30/2024
Accepted: 04/08/2025