Artículo de revista:

Castro Torres, A. F. (2024). "Revisiting North-South Disparities in Naming Practices: An Extension and Update". Social Currents, 0(0) (ISSN 2329-4973) https://doi.org/10.1177/23294965241297932

Revisiting North-South Disparities in Naming Practices: An Extension and Update.

Abstract

This research note presents evidence of persistent disparities in naming practices in social science research between the global North and South from 1960 to the present. The analysis relied on bibliometric records of 11.9 million articles sourced from the OpenAlex collection. By covering the 1960-2024 period and including six social science and humanities sub-disciplines –Economic, Geography, History, Political Sciences, Psychology, and Sociology– this study replicates and expands an existing study that used Scopus data (0.5 million articles). Results show that scientific articles about global North countries and populations are systematically less likely to mention the country name in their titles compared to global South. For example, compared to the US, articles about China are between 1.8 and 2.7 times more likely to include the country name in the title. Significant variations across sub-disciplines are observed, with Psychology displaying the lowest localization rates and greatest geographical gaps, and History showing seemingly converging trends over time across geographies. Continued research is needed to warrant diversity and inclusivity in social science research. A country-year dataset with country-mention counts in titles and abstracts is made available.

Revisiting North-South Disparities in Naming Practices: An Extension and Update.

INTRODUCTION

Written language plays a critical role for science progress and dissemination. Scientifi claims that fail to specify sufficiently the population, place, or time to which they refer may be harmful. For example, expressions such as boys are stronger than girls, or mosquitoes transmit viruses, may be true only in specific contexts- Not all boys are stronger than girls, and not all mosquitoes transmit viruses, indeed, only a tiny share of them do so (Wodak et al., 2015). This type of expression is referred to as generic language. In the social sciences, specifically, it is crucial to include geographical references to contextualize evidence and conclusions. Since titles play a central role in disseminating scientific findings, trends in naming locations in research article titles may reveal underlying social biases. Examining these trends could help identify potential pathways to address and reverse such biases.

To do so, this work expands and updates the evidence shown by Castro Torres & Alburez-Gutierrez (2022) regarding a sustained geographical gap in naming practices in social science articles. The authors reported the geographical distribution of articles based on the countries mentioned in abstracts, and the proportion of them that include a country name or demonym in the title. They termed this proportion *localization rate*. The authors conceptualized the lack of country mentions in titles as a form of generic language whereby unwarranted generalizability is ascribed to evidence produced in locations that do not require to be mentioned, namely, Western societies (DeJesus et al., 2019; Wodak et al.,

2015). The reverse is true for locations and populations for which evidence is deemed as not generalizable.

According to their results, articles on global North countries and populations -the US and West European countries– display the lowest *localization rate* for the entire period of analysis: 1990 to 2022. These naming gaps were robust to controlling for articles' number of authors, the length of the title, authors' location, and the year of publication. These results potentially imply an unjustified assertion of universality of global North studies, which could diminish recognition of studies focused on the global South. Together with other scholars, the authors argue that these trends are a consequence of Eurocentric biases in social science research (Alvares, 2011; Grosfoguel, 2013; Krause, 2016; Mignolo, 2014; Quijano, 2000).

In this study, I replicate and expand upon the work of Castro & Alburez-Gutierrez (2022) by broadening its geographical and historical scope and analyzing specific trends for six selected subdisciplines: Economics, Geography, History, Psychology, Political Sciences, and Sociology. I also assess the potential role of city names in naming practices. For this, I calculate the *localization rate* for articles that mention the three most populous cities in the US, the UK, China, and India. These are the top four contributing countries in terms of research articles across the examined subdisciplines.

These enhanced and disaggregated analyses are possible thanks to the large coverage and size of the OpenAlex collection compared to the Scopus data (Haupka et al., 2024; Priem et al., 2022). I utilize OpenAlex data spanning the period from 1960 to 2024, thereby

incorporating over 30 additional years of evidence with respect to Castro Torres & Alburez-Gutierrez' work.

This study offers a replication, extension, and update of biases in scientific naming practices across the Social Sciences and Humanities. The analysis of city names further supports the findings based on country names and demonyms, opening new avenues for future research on naming practices. These contributions underscore the power and potential of using open data sources for conducting science-of-science research.

MATERIAL AND METHODS

The results presented here are based on the OpenAlex collection. This collection offers three advantages. First, OpenAlex is Open Access, making this research and future research accessible to virtually any scholar worldwide.

Second, OpenAlex has a larger coverage in terms of articles' language and geographic location of authors compared to privative databases (Falagas et al., 2008; Priem et al., 2022). For example, according to Haupka et al. (2024, p. 6), the OpenAlex database included over 69 million journal articles published between 2012 and 2022, compared to approximately 56 million in Semantic Scholar, 32 million in Scopus, 24 million in Web of Science, and 13 million in PubMed. Additionally, at least 96% of the articles in the smaller databases are included in the OpenAlex collection, demonstrating the comprehensiveness of this data source (Haupka et al., 2024). This greater coverage makes new estimates more representative of global scholarly production than results from privative bibliometric

databases. In addition, larger sample size allows me to analyze trends for five Social Sciences subdisciplines and for History. For the sake of brevity, I refer to these six sub disciples as Social Sciences subdisciplines although History is typically classified within the Humanities.

Third, by using data from 1960 onwards, this study adds more than three decades of additional evidence. This is an important contribution even though full coverage for early years may not be guaranteed.

Using the openalexR R-package (Aria et al., 2024), I captured all articles pertaining to Economics, Geography, History, Psychology, Political Sciences, and Sociology with complete abstract information and published between January 1st, 1960 and July 31st, 2024. This procedure yielded a database of 38,876,201 articles. After removing duplicates and articles with titles with fewer than 15 characters, this sample was reduced to 38,626,855. Importantly, the share of articles with abstract information has increased steadily over time from only 35% in 1960 to over 70% from 2020 onwards.

Next, I used a list of 249 regular expressions for country names and demonyms provided by Castro Torres and Alburez-Gutierrez to remove articles with no country or demonym mentions (country mentions hereafter) in the abstract. This procedure yielded a working sample of 11,948,826 articles. To favor comparability and reduce the potential influence of space limitation in naming practices, I kept articles with at most two country mentions, i.e., 93% of articles of the working sample. This restriction yielded an analytical sample of 11,102,827 articles, which is over 22 times larger than the one used by Castro Torres and Alburez-Gutierrez (i.e., 0.5 million).

Using the analytical sample, I built a long-format dataset where each line corresponds to a unique country-mention in an articles' abstract (n = 12.6 million). I coded a target variable (Y) as one if the country-mention was present in the title and zero otherwise. Following Castro Torres & Alburez-Gutierrez (2022)'s methodology, I use this variable to compute country-level *localization rates* as the fraction of country-mentions in titles over country mentions in abstracts. Results are presented separately for the top four countries in terms of the number of articles (i.e., the US, the UK, China, and India) and for the rest of the countries grouped into the United Nations Sustainable Development Goals (UNSDG) world regions classification (United Nations, 2017).

As expressed in the following equation, I estimate relative localization risks for top three countries and other UNSDG world regions (X_1) using a Poisson model with the natural logarithm as the link function. The model specification includes controls for quartiles of the number of authors (X_2) –One, Two, Three or Four, more than four–, and titles' length in characters (X_3) –16-61, 62-85, 86-112, and 112+–, and articles' year of publication in two-year bins from 1960 to 2024 (X_4),

$$\log(\mathbb{E}[Y_{rijk}]) = \beta_0 + \sum_{r=1}^{R} \beta_{1r} X_{1r} + \sum_{i=1}^{I} \beta_{2i} X_{2i} + \sum_{j=1}^{J} \beta_{3j} X_{3j} + \sum_{k=1}^{K} \beta_{4k} X_{4k}$$

In this equation, the log(), E[] and β_0 are the natural logarithm, the expected value operator, and the models' intercept, respectively. The subscripts r, i, j, and k index the categories of the dependent variables as explained above with reference categories being the "United States of America" for X_1, and the first category for X_2 to X_4. The choice of this

specification favors compatibility with Castro Torres & Alburez-Gutierrez' results. The only variable I am not able to include in the model is authors' location because 4,300,541 articles, i.e., 39% of the analytical sample, did not include this information.

Finally, to partially assess the role of city names, I conducted query on OpenAlex abstracts using the names of the three most populous cities in the US (New York, Chicago, Los Angeles), the UK (London, Glasgow, Birmingham), China (Shanghai, Beijing, Chongqing), and India (Mumbai, Delhi, Bangalore). The results are grouped by country. I then counted the number of titles including these city names in the analytical sample and aggregated them by country. I took the ratio of these two aggregate counts as an estimate of the *localization rate* for city names.

RESULTS AND DISCUSSION

Historical Roots of North/South Naming Disparity Gaps

Figure 1 shows time trends in the *localization rate* for articles across the six selected subdisciplines from 1960 to July 31st, 2024. I group estimates in two-year bins to smooth trends. The total number of country-mentions and the percentage comprised by the top three countries for each discipline are displayed as legends.

Figure_1_here

Overall, localization rate trends align with the evidence presented by Castro Torres & Alburez-Gutierrez despite the discrepancies in terms of the numerical dominance of the US and UK in the share of articles.

Starting with the discrepancies, China is the country with the largest share of mentions in four out of six subdisciplines: Psychology (10%), Political Science (12%), Geography (14%), and Economics (14%). In Sociology, where the US is first with 9%, the difference between China and the US is below one percentage point. The UK only leads in History, and it does not appear in the top three in Geography. In this latter subdiscipline, China dominates with 14% and India takes the third position with 5%. The shares for the US and the UK may underestimated due to mentioning biases in abstracts as reported by Kahalon et al. (2021) for Psychology. However, the dominance of the two Anglo countries, along with two of the largest countries worldwide population-wise is consistent with existing results and it speaks to the greater scope of the OpenAlex data.

Moving on to the localization rates, the US displays the lowest *localization rate* together with the UK and other European countries; China and India display among the highest *localization rates*, only surpassed by Oceania and the Sub-Saharan Africa region. These trends support the idea that naming practices disparities stem from epistemological hegemony rather than numeric dominance. Countries with similar contributions to science show divergent *localization rates*, with the divide largely reflecting global North/South distinctions. This suggests structural inequalities in how research from different regions is referenced in scientific publications.

Figure 1 offers additional insights into the North/South disparities in naming practices. First, it confirms the historical persistence of *localization rate* gaps between the global North and South. There is a general flat pattern in *localization rates* up to the 2000s and recent years depict increases for articles about regions such as Sub-Saharan Africa, Latin

America and the Caribbean, and Europe and North America. The step increase in localization rates for the Sub-Saharan Africa region is striking and it was not present in Castro Torres & Alburez-Gutierrez's (2022) work. Likewise, an arguably out of the mainstream world region, Oceania (excluding Australia and New Zealand), displays expected trends: historically erratic due to small numbers of country mentions along with a high *localization rate* by the end of the study period. This region was left out by Castro Torres & Alburez-Gutierrez due to small sample size.

Second, the geographical gaps in Castro Torres & Alburez-Gutierrez' study are smaller than those shown here. This difference may be a consequence of the Scopus bias toward English literature which likely translates into a higher likelihood of US and UK mentions in titles, and fewer articles about global South countries. Thus, these larger and persisting gaps, suggest that Scopus-based results were slightly biased, although not entirely driven, by the selection of journals in the Scopus database. The argument on Eurocentrism as a root cause for naming practice disparities holds in a longer time series and over a more global dataset.

Third, differences across subdisciplines include localization rates' levels and trends, yet their overall range by geography is rather similar, particularly in recent years. For example, for the 2023-2024 period the range of the localization rates by geographies is of at least 40 percentage point in all subdisciplines except for History. In addition, there are significant differences in terms of *localization rates* levels and trends for Psychology and History. Articles in Psychology display the lowest *localization rate* with most of the data points below 0.5, and the US and UK hovering around 0.2 across the six-decade period. Articles

in History display a decreasing trend over time in *localization rates* and there are visual indications of convergence after 2005 (excluding Oceania). There are also slight increases in the *localization rate* for recent years in all regions and countries except the US. This trend in History articles is unique and may be due to the relatively small overlap of articles with other subdisciplines. In contrast, trends across all Social Science subdisciplines appear similar because they share many common articles.

Nuances in Naming Practices Gaps Across Subdisciplines

Figure 2 summarizes geographical gaps for each discipline by estimating articles' relative risk of localization compared to articles about the US (reference category). These relative risks are estimated by predicting a variable that takes the value of one if the title includes a country-mention and zero otherwise. Models control for the number of authors, titles' length, and the year of publication as described in the equation above.

Figure_2_here*

Figure 2 confirms the significance of geographical gaps in naming practices between the US and the rest of the countries, and between global North and South locations. Net of the number of authors, titles' length, and the year of publication, the US is the location with the lowest *localization rate* with relative risks for other countries and world regions above 1.5. The UK and other European and North American countries display relative risks below 1.8 in all studied subdisciplines.

Regarding gaps with global south locations, China and Sub-Saharan countries display the largest relative risks with values above 1.9. These figures imply that articles about these two locations are almost twice as likely to include the country study name in their title compared to articles about the US. Relative risks for articles about regions other than Sub-Saharan Africa range between 1.6 and 1.8. The Latin American and the Caribbean region stands out as a global South region with relatively low relative risks, particularly in Political Science.

Articles in History display smaller geographical gaps with relative risks below 2.2 and overlapping confidence intervals for China and global South regions, except for Sub-Saharan Africa and Oceania. These results are consistent with visual inspection of geographical convergence in the *localization rate* for recent years and it suggests a weaker hegemony of global North over the global South in terms of naming practices. Articles in Psychology display the opposite pattern with very large and sustained North/South geographical gaps suggesting strong global North and US hegemony. For example, the relative risk of localization for articles mentioning Sub-Saharan countries in their abstract compared to articles mentioning the US is over 3.3, meaning an over 300% likelihood of localization.

Assessing the Potential Role of City Names on Naming Practices

The 12-city names query yielded 1,158,754 abstracts. This figure is at least twleve times smaller than the 11.9 million abstracts with country mentions confirming that country names and demonyms are more frequent to refer to study locations in research articles.

The distribution of these articles across countries shows a strong dominance of US and UK cities, with half of the articles set in US locations (50%), almost one third in the UK (31%), and 15% and 4% in China and India, respectively. This results itself merits further research beyond the scope of this note.

More importantly for the purpose of this note, *localization rates* for articles set on cities replicate the global North/South patterns observed at the country level, although with smaller gaps. The US cities have the lowest proportion of localized papers at 28.9%, followed closely by UK cities at 29.1%, India at 33.7%, and Chinese cities at 41.1%. These fractions display a declining trend through the period for US and UK cities, and they are erratic for Indian and Chinese cities due to a reduced number of articles set in these locations from the 1960s to the late 1990s. After the 1990s, localization rates for Chinese and Indian cities become stable at above 30%. Year-specific results are available upon request.

The fact that these fractions do not account for country mentions in titles may explain the lower localization rates for China and India compared to their country-level rates. Meanwhile, the higher localization rates for US and UK cities relative to their respective countries suggest a stronger incentive to highlight specific study locations. Given the distribution of articles and the similarity of these localization rates, our results are likely robust to the inclusion of city names.

CONCLUSION

Using publicly available bibliometric data on 11.9 million scientific articles published between January 1st, 1960, and July 31st, 2024, I have found support for Castro Torres & Alburez-Gutierrez' argument on the centrality of Eurocentrism in social science and its consequence for naming practices. The results presented here add a long-term historical perspective and subdiscipline-specific nuances to scientific naming biases.

The historical persistence of naming practice gaps and the disciplinary nuances reported here speak of the complexity of this phenomena with articles in History displaying the lowest geographical gaps and articles in Psychology the largest. These results augment the significance of Castro Torres & Alburez-Gutierrez' recommendation for keeping diversity and inclusivity in social science research.

Evidence is always localized and its scientific validity and generalizability can only be enhanced by referring to the historical and geographical context. Omitting geographical context may have the opposite effect. As such, these historical persisting gaps require attention from the scientific community and scholars as authors, reviewers, and editors. Social scientists should carefully consider the wording of article titles to prevent perpetuating detrimental dynamics of intellectual dominance, which hinder inclusivity and may undermine the generalizability of scientific knowledge.

Our relatively small empirical exercise on city names opens an interesting avenue for refining research on naming practices. Cities represent more specific locations compared to countries, which may provide authors with greater incentives to include city names in titles. This is supported by the higher localization rates observed for U.S. and U.K. cities, as compared to their respective country mentions. In contrast, the results for Chinese and Indian cities suggest the opposite trend. To advance this research, a more refined methodology and possibly a different theoretical framework are necessary.

Data availability statement: A country-level dataset (667,484 rows and nine columns) with counts of country-mentions in titles (r_dtitle) and abstracts (r_dabstr), and the localization rate (loc_rate) by all independent variables, the United Nations Sustainable Goals regions including the top three countries, and the six sub disciplines is available at: (XXXX, removed to preserved anonymity. Data included in the submission)

REFERENCES

Alvares, C., 2011. A Critique of Eurocentric Social Science and the Question of Alternatives. Economic and Political Weekly 46, 72–81.

Aria, M., Le, T., Cuccurullo, C., Belfiore, A., Choe, J., 2024. openalexR: An R-Tool for Collecting Bibliometric Data from OpenAlex. The R Journal 15, 167–180. https://doi.org/10.32614/RJ-2023-089

Castro Torres, A.F., Alburez-Gutierrez, D., 2022. North and South: Naming practices and the hidden dimension of global disparities in knowledge production. Proc. Natl. Acad. Sci. U.S.A. 119, e2119373119. https://doi.org/10.1073/pnas.2119373119

DeJesus, J.M., Callanan, M.A., Solis, G., Gelman, S.A., 2019. Generic language in scientific communication. Proc Natl Acad Sci USA 116, 18370–18377. https://doi.org/10.1073/pnas.1817706116

Falagas, M.E., Pitsouni, E.I., Malietzis, G.A., Pappas, G., 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB j. 22, 338–342. https://doi.org/10.1096/fj.07-9492LSF

Grosfoguel, R., 2013. Racismo/sexismo epistémico, universidades occidentalizadas y los cuatro genocidios/epistemicidios del largo siglo XVI. Tabula rasa 31–58. https://doi.org/10.25058/20112742.153

Haupka, N., Culbert, J.H., Schniedermann, A., Jahn, N., Mayr, P., 2024. Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar. https://doi.org/10.48550/ARXIV.2406.15154

Kahalon, R., Klein, V., Ksenofontov, I., Ullrich, J., Wright, S.C., 2021. Mentioning the Sample's Country in the Article's Title Leads to Bias in Research Evaluation. Social Psychological and Personality Science 194855062110240. https://doi.org/10.1177/19485506211024036

Krause, M., 2016. 'Western hegemony' in the social sciences: fields and model systems: 'Western hegemony' in the social sciences: fields and model systems. The Sociological Review Monographs 64, 194–211. https://doi.org/10.1002/2059-7932.12008

Mignolo, W.D., 2014. Spirit out of bounds returns to the East: The closing of the social sciences and the opening of independent thoughts. Current Sociology 62, 584–602. https://doi.org/10.1177/0011392114524513

Priem, J., Piwowar, H., Orr, R., 2022. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/10.48550/ARXIV.2205.01833

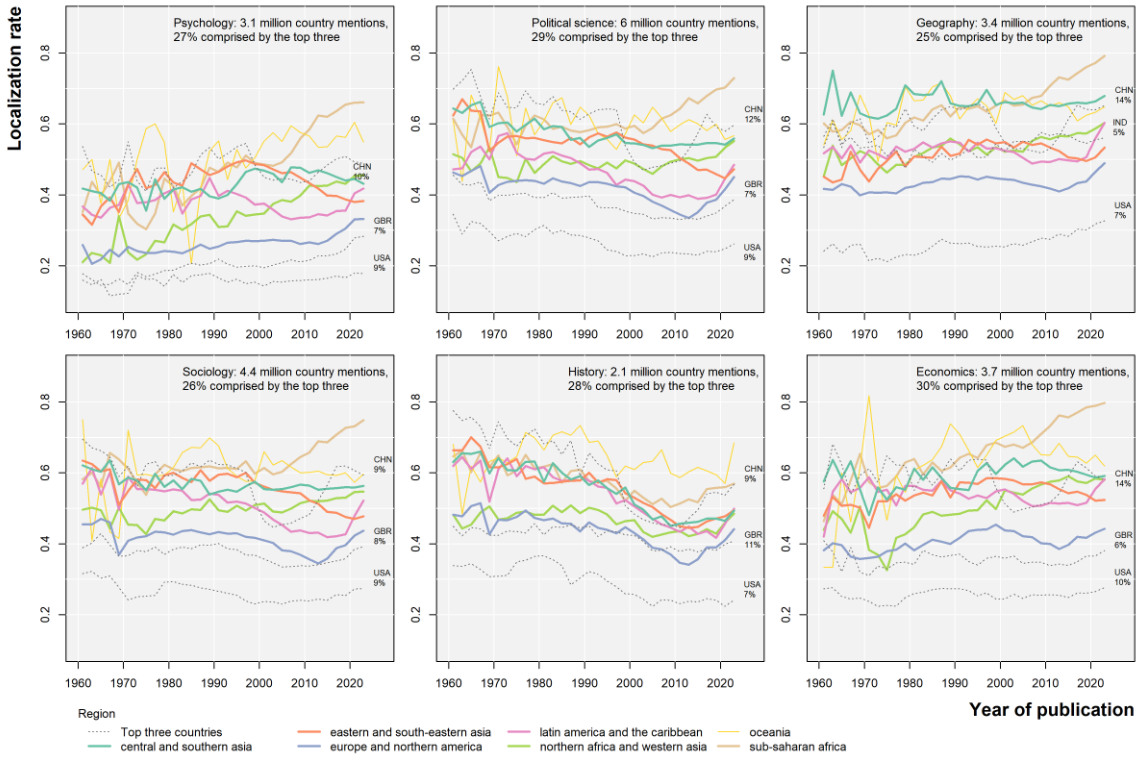
Quijano, A., 2000. Coloniality of Power and Eurocentrism in Latin America. International Sociology 15, 215–232. https://doi.org/10.1177/0268580900015002005

United Nations, 2017. United Nations Sustainable Development Goals [WWW Document]. Regional groupings used in Report and Statistical Annex. URL https://unstats.un.org/sdgs/indicators/regional-groups/

Wodak, D., Leslie, S.-J., Rhodes, M., 2015. What a Loaded Generalization: Generics and Social Cognition: What a Loaded Generalization. Philosophy Compass 10, 625–635. https://doi.org/10.1111/phc3.12250

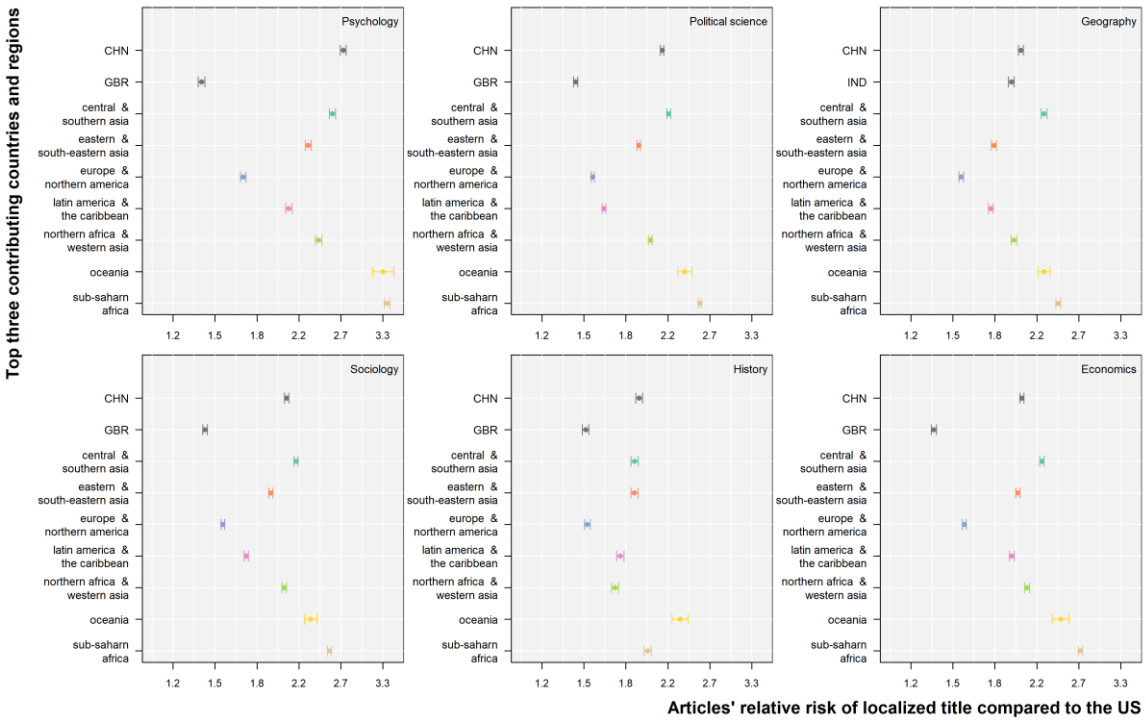
FIGURES

Figure 1. Time trends in the *localization rate* for the top three contributing countries and other countries grouped by United Nations Sustainable Development Goal world regions.



Note: The top three and regional categories are mutually exclusive. Oceania excludes Australia and New Zealand. These countries are grouped in the Europe and North America region.

Figure 2. Relative risk of localized titles for articles about China, the UK, and seven world regions compared to articles about the US. OpenAlex data 1960-2023.



Note: Oceania excludes Australia and New Zealand. These countries are grouped in the Europe and North America region.