Received: 15 July 2022 | Revised: 17 September 2022 | Accepted: 23 October 2022

DOI: 10.1002/asi.24723

RESEARCH ARTICLE

JASIST WILEY

Gender and country biases in Wikipedia citations to scholarly publications

Xiang Zheng[1] | Jiajing Chen[1,2] | Erjia Yan[3] | Chaoqun Ni[1]

[1]Information School, University of Wisconsin-Madison, Madison, Wisconsin, USA

[2]Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, New York, USA

[3]College of Computing & Informatics, Drexel University, Philadelphia, Pennsylvania, USA

Correspondence
Chaoqun Ni, Information School, University of Wisconsin-Madison, Madison, WI 53706, USA.
Email: chaoqun.ni@wisc.edu

Abstract

Ensuring Wikipedia cites scholarly publications based on quality and relevancy without biases is critical to credible and fair knowledge dissemination. We investigate gender- and country-based biases in Wikipedia citation practices using linked data from the Web of Science and a Wikipedia citation dataset. Using coarsened exact matching, we show that publications by women are cited less by Wikipedia than expected, and publications by women are less likely to be cited than those by men. Scholarly publications by authors affiliated with non-Anglosphere countries are also disadvantaged in getting cited by Wikipedia, compared with those by authors affiliated with Anglosphere countries. The level of gender- or country-based inequalities varies by research field, and the gender-country intersectional bias is prominent in math-intensive STEM fields. To ensure the credibility and equality of knowledge presentation, Wikipedia should consider strategies and guidelines to cite scholarly publications independent of the gender and country of authors.

1 | INTRODUCTION

Wikipedia may be the largest and most accessible knowledge source in the contemporary world (Mesgari et al., 2015). Between 2019 and 2021, English Wikipedia, the largest Wikipedia edition, had 40,000 active page editors and 9.5 billion page views from 844 million distinct devices per month (Wikimedia Foundation, 2022). Wikipedia users, including students, educators, researchers, professionals, and the general public, often deem Wikipedia a credible knowledge source, given its comprehensiveness, accessibility, readability, and currency (Eijkman, 2010; Mesgari et al., 2015; Okoli et al., 2014; Thelwall & Kousha, 2016). Popular search engines retrieve information from Wikipedia as authoritative answers to search queries and use Wikipedia as a knowledge base (Ford et al., 2015). Wikipedia has also become a primary online source for health and legal information for the public (Laurent & Vickers, 2009; Okoli et al., 2014). Usage of Wikipedia in scholarly research (Tomaszewski & MacDonald, 2016), K-12 education (Hew & Cheung, 2009), and patenting (Orduna-Malea et al., 2017) is rising. Thus, Wikipedia is widely seen as an authoritative platform for knowledge and facts, which has significant power in representing and disseminating knowledge to the public and professional communities.

Wikipedia's "authoritativeness" is primarily built upon its performative behavior of citing credible sources

Abbreviations: BHS, Biomedical and Health Sciences; CEM, coarsened exact matching; CWTS, Center for Science and Technology Studies at Leiden University; DOI, digital object identifier; LES, Life and Earth Sciences; MCS, Mathematics and Computer Science; OA, open access; PSE, Physical Sciences and Engineering; SSH, Social Sciences and Humanities; WoS, Web of Science.

(Ford et al., 2015; Priem, 2014), including scholarly publications. Scholarly publications are estimated to account for about 8.3% of all references cited by Wikipedia (Singh et al., 2021). Thanks to the contemporary peer review system, scholarly publications have become one predominant way to present and disseminate research discoveries with high credibility, reliability, authority, and public trust (Kelly et al., 2014). Wikipedia citations to scholarly publications partially form the foundation of Wikipedia's content. With its worldwide social impact and broad audience, Wikipedia also acts as a platform for communicating and dispersing scientific discoveries from academia to the public domain (Jemielniak & Aibar, 2016). It facilitates the spread of scientific discoveries by distilling the research findings, which are often oriented towards the research community and published in subscription-based journals, and amplifies their public values (Teplitskiy et al., 2017). Wikipedia thus acts as a filter and decides what scientific knowledge to display to its users. Therefore, Wikipedia editors, who usually choose what scholarly publications to cite in Wikipedia entries, become "janitors of knowledge" (Sundin, 2011), who control the dissemination of scientific discoveries and authoritative knowledge to various stakeholders.

Scientists have started to recognize the importance of having their research cited on Wikipedia. For scholarly publications, being cited by Wikipedia acknowledges the authors' work that adds to their symbolic capital (Shuai et al., 2013). Wikipedia citation is suggested as an alternative metric (*Altmetrics*) to complement academic citations, which is criticized (among others) for neglecting the social impact of research outside academia (Priem, 2014; Priem et al., 2010). There is also evidence that academic citations and Wikipedia citations to scholarly publications are positively correlated (Heydari et al., 2019; Shuai et al., 2013). Furthermore, under the increasing pressure to show the impact of funded research, scientists and funders are keen to know whether Wikipedia cites their research (Reich, 2011). Therefore, there is a demonstrated role of Wikipedia citations to scholarly publications in the science ecosystem. Ensuring Wikipedia cites scholarly publications based on quality and relevancy but not extrinsic factors such as social demographics is critical to the impartiality of knowledge dissemination and the fair recognition of contributions by scholars.

Nonetheless, concerns over Wikipedia's reliability and accuracy were raised due to its citation practices (Halavais & Lackaff, 2008; Mesgari et al., 2015; Thelwall & Kousha, 2016). Under Wikipedia's policies of "verifiability, not truth" and no original research presentation (Wikipedia, 2022b), Wikipedia editors may arbitrarily select sources with exemptions from the truth-checking responsibility. As Ford et al. (2015) described, Wikipedia editors are "active co-creators of knowledge" rather than "passive collectors of knowledge held in sources" because they support facts using "specially chosen sources." In reality, Wikipedia editors may have inadequate professional training to distinguish valuable research: two surveys in 2010 (Glott et al., 2010) and 2011 (Wikimedia Foundation, 2011) showed that only 4.43 and 8% of Wikipedia editors completed doctoral education, respectively. When faced with the uncertainty of distinguishing sound scholarly publications, editors may leverage other features to help make decisions (Murray et al., 2019). Studies suggested that certain groups of scholarly publications are more likely to be cited by Wikipedia: Wikipedia prefers studies published in highly prestigious journals (Arroyo-Machado et al., 2020; Benjakob et al., 2022; MacHado et al., 2020; Nielsen, 2007). Publications in specific fields, for example, biomedical fields, are more cited than those in other fields (Arroyo-Machado et al., 2020; Teplitskiy et al., 2017). A publication's open access (OA) status also increases its likelihood of being cited by Wikipedia (Benjakob et al., 2022; Teplitskiy et al., 2017). Wikipedia entries may have an inflated focus on recent events and tend to cite recent resources (Jemielniak et al., 2019; Sundin, 2011; Wikipedia, 2022a).

Despite the revelation from prior research, scarce research focused on the likely gender- or country-based biases of citation practices in Wikipedia. This topic is significant as the gender- and country-based disparities are stark in global academia. While there are slightly contradictory conclusions regarding women's disadvantage in academic citations (Larivière et al., 2013; Thelwall, 2020), recent research agrees that men appear more often as first authors in highly cited publications than women (Thelwall, 2020; Zhang et al., 2021). Publications from highly developed countries are also found to be cited more often (Sugimoto et al., 2015). If Wikipedia holds gender or country biases when citing scholarly publications, it may aggravate the existing disparities in global academia, marginalize minority researchers in knowledge dissemination, and distort the representation of knowledge to human beings. Unfortunately, gender and country biases in Wikipedia's coverage and content have been revealed. Demographic studies showed over 80% of Wikipedia editors are men (Glott et al., 2010; Hill & Shaw, 2013; Lam et al., 2011; Wikimedia Foundation, 2011). A significant number of editors across language versions speak English (Wikimedia Foundation, 2011). As the largest edition, English Wikipedia editors are more likely to be White and live in the United States and the United Kingdom (Wagner et al., 2016). A study shows that 56 and 13% of English Wikipedia's sources are from the United States and the United Kingdom (Ford et al., 2013). Consequently, Wikipedia is criticized for its "masculine culture" (Ford & Wajcman, 2017) and "geek culture" (Reagle, 2013), which

alienates women and expands gender disparities in Wikipedia. Wikipedia is suggested to reflect the Anglo cultures more than other cultures (Callahan & Herring, 2011). This can further exacerbate the disparity between countries with strong editing cultures and those on the peripheries that fail to reach critical masses (Graham et al., 2014).

This study aims to investigate gender and country biases in Wikipedia citation practices. The gender bias in this study is defined as the gap between the proportion of women-authored publications cited in Wikipedia (Wiki-cite group) and the proportion of available women-authored publications in the broader publication database (control group) when the two groups are homogeneous regarding other known characteristics. Likewise, country bias is defined as the gap between the proportion of publications by authors from certain countries in the Wiki-cite group and the proportion of available publications by authors from that country in the control group. Given English's standing as a global language and English Wikipedia's overwhelming popularity (Graham et al., 2014), this study uses English Wikipedia as an example. Accordingly, we dichotomized countries of affiliations into two groups by language and culture: Anglosphere and non-Anglosphere. We include the United States, the United Kingdom, Canada, Australia, and New Zealand in the Anglosphere countries. The concept of the Anglosphere historically has racial and imperialist roots but now emphasizes more the cultural homogeneity and collective identity shared by the community of English-speaking countries (Malcolm, 2021). These five countries are consistently admitted as the core of that community with strong historical, political, economic, and cultural ties (Vucetic, 2011) and share similar cultures, values, and ideologies (Legrand, 2016). Non-Anglosphere countries include other countries in the world. Using coarsened exact matching (CEM) (Iacus et al., 2012), this study investigates the biases by research field: Biomedical and Health Sciences (BHS), Life and Earth Sciences (LES), Physical Sciences and Engineering (PSE), Mathematics and Computer Science (MCS), and Social Sciences and Humanities (SSH). As intersectional inequalities are indicated to exist in science widely (Kozlowski et al., 2022), this study also explores the intersectional impact of gender and country to understand such disparities in Wikipedia citations.

2 | MATERIALS AND METHODS

2.1 | Data

This study mainly relies on two data sources: citations by English Wikipedia pages by Singh et al. (2020) and the Web of Science (WoS). The Wiki-cite dataset contains about 1.2 million distinct digital object identifiers (DOI) for scholarly publications cited by 6.1 million English Wikipedia entries as of May 2020. We then matched Wiki-cite with WoS and were able to find 746,046 records overlapped by the two sources. We further filtered these publications by document type, keeping publications labeled as articles, reviews, or proceeding papers by WoS. We also excluded non-English publications, given the scope of the study being limited to English Wikipedia only.

Additionally, we adopted the publication classification system by the 2021 Center for Science and Technology Studies at Leiden University (CWTS), which categorizes each publication into a fine-grained microlevel field and its associated broad main field (Centre for Science and Technology Studies, 2022). The classification system clusters publications in WoS into 4,139 microlevel fields based on publication-level citation relationships (Waltman & van Eck, 2012). The microlevel fields are further aggregated into the five broad main fields used in our analysis. It is noted that we limited our analyses to publications published between 2005 and 2020, during which the CWTS microlevel field classification has the most comprehensive coverage. The process kept 383,474 publications cited by English Wikipedia.

2.2 | Methods

2.2.1 | Control group construction using CEM

To understand the potential citation bias by Wikipedia, using CEM, we created a control group consisting of scholarly publications that were not cited by Wikipedia but shared a list of features with those cited, that is, the Wiki-cite group (see Appendix). Like the Wiki-cite group, the control group also derives from the pool of publications labeled as articles, reviews, and proceeding papers by WoS and using English as the document language. The shared features include publication year, publication venue, publication topic, team size, and open access (OA) status. Specifically, for each publication in the Wiki-cite group, we found their uncited pairs that are published during the same period in the same venue, within the same topical area, by an author team of similar size, and of the same OA status.

We selected uncited publications that were published in the same venue as the Wikipedia-cited ones for the control group, assuming publications in the same venue are of similar quality, readability, and disciplinary scope. Acknowledging publications in the same venue may vary

significantly in terms of topics of research, we further used the microlevel field classification by CWTS (Waltman & van Eck, 2012) as a fine-grained topic control for publications. With this strategy, only uncited publications under the same microlevel field as those cited by Wikipedia were included in the control group. Furthermore, as time accumulation and team sizes are shown to be related to a publication's impact, popularity, and novelty (Huang et al., 2019; Larivière et al., 2015; Lee et al., 2015), and the OA status of publications matters for their accessibility and visibility to the readers (Teplitskiy et al., 2017), we also controlled publication year, team size, and OA status in matching. Because we are interested in understanding Wikipedia citation bias based on authors' gender and country of affiliation, building a control group based on the criteria mentioned above allows us to rule out the potential effect of these criteria on Wikipedia's citation practices.

2.2.2 | Gender and country classification

The gender category of authors in the Wiki-cite group and control group was assigned using the method designed by Larivière et al. (2013), which was built based on the United States census data and some country-specific gender-name lists. The country category of publications was decided based on the country of author affiliation information provided by WoS. During the process, we excluded publications that only list initials for the first name of authors (16.9% of total) and lack information for author affiliations (1.1% of total), to which our classification approach cannot be applied. We successfully assigned gender and country classification to the key (first and last) authors of 89.0% of publications in the Wiki-cite group and control group, a similar percentage to Larivière et al. (2013). The above matching and classification processes create a collection of 198,344 publications in the Wiki-cite group and 1,900,708 publications in the control group with classifications. To examine potential skewness towards certain groups in our sample, we compared the distribution of publications in the Wiki-cite group and the group with gender and country classification (see Table S1, Supporting Information). The chi-square tests showed no significant differences in any group.

2.2.3 | Statistical analysis

We used the relative difference to measure the distance between the percentage of scholarly publications by a social group (e.g., gender) in the Wiki-cite group and the control group (see Appendix). We also performed weighted binary logistic regression (using CEM weights) to estimate the statistical significance of the role by gender or country (Blackwell et al., 2009) in the likelihood of a publication being cited in Wikipedia. Specifically, we used the Wikipedia citation status of each publication as the outcome variable and the gender (or country) classification as the independent variable while controlling for all the coarsened variables.

TABLE 1 Authorship composition of the analytical sample

	Wiki-cite group		Control group	
	#	%	#	%
Single-authored	21,798	11	174,819	9
Multi-authored	176,546	89	1,725,889	91
Total	198,344	100	1,900,708	100

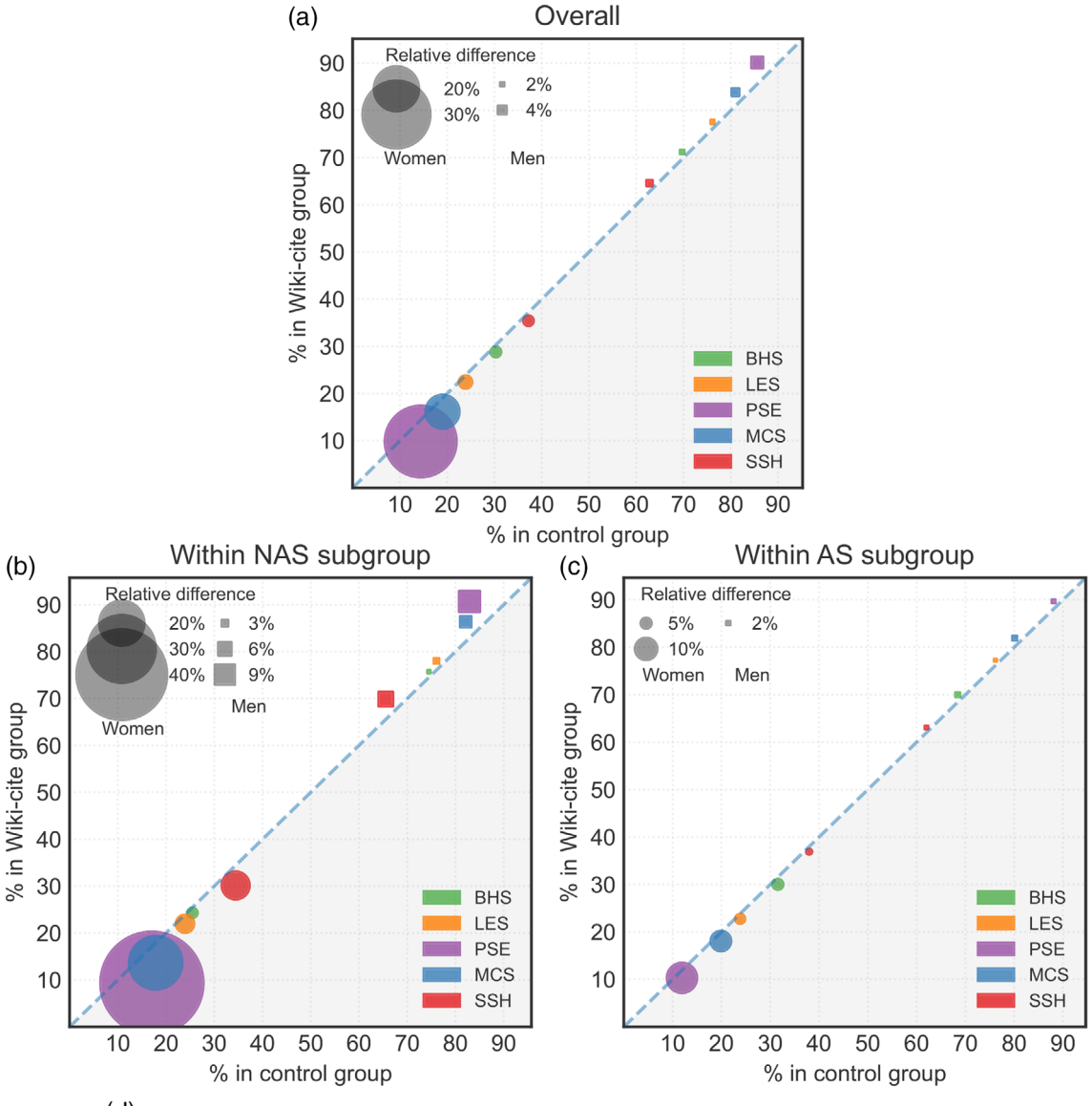
3 | RESULTS

3.1 | Gender-based biases

We split publications by the number of authors in their bylines into two subsets: single-authored publications and multi-authored publications (see Table 1). The gender classification of single-authored publications was decided based on the gender category of the author. For multi-authored publications, we considered both the first and last authors of each publication as key authors, who are generally the dominant contributors to scholarly publications (Ni et al., 2021). Therefore, the gender groups of multi-authored publications include women (both key authors are women), men (both key authors are men), and mix-gender (two key authors are of different gender categories). The country groups of multi-authored publications include Anglosphere (both key authors are affiliated with Anglosphere countries), non-Anglosphere (both key authors are affiliated with non-Anglosphere countries), and mix-sphere (two key authors are affiliated with different country groups).

3.1.1 | Single-authored publications

We first compared the difference between the percentage of publications by women and men in the Wiki-cite group and that percentage in the control group. The percentage of publications by women in the Wiki-cite group is lower than that in the control group in every field (falling into the gray area below the diagonal) (see Figure 1a



(d)

rd (%)	NAS subgroup		AS subgroup		All	
	Women	Men	Women	Men	Women	Men
BHS	-4.7	1.6	-4.9	2.3	-4.9	2.1
LES	-8.2	2.6	-4.6	1.4	-6.1	1.9
PSE	-45.5	9.4	-13.5	1.8	-31.7	5.3
MCS	-23.6	5.1	-9.2	2.3	-15.1	3.5
SSH	-12.5	6.6	-2.8	1.7	-4.8	2.8

FIGURE 1 Gender-level comparison of single-authored publications. Percentages in the Wiki-cite and control group and relative differences of women authors (circle) and men authors (square) by main field in (a) the overall sample, (b) non-Anglosphere subgroup, and (c) Anglosphere subgroup. (d) Relative differences for the women and men subset. AS, Anglosphere; NAS, non-Anglosphere; rd, relative difference

and Table S2); the relative difference ranges from −31.7% (PSE) to −4.8% (SSH) (see Figure 1d). In contrast, the percentage of publications by men in the Wiki-cite group is higher than that in the control group in every field (falling into the white area above the diagonal), with relative differences ranging from 1.9% (LES) to 5.3% (PSE). The logistic regression analysis (see Figure 2a and Table S3) shows that publications by women are significantly less likely to be cited by Wikipedia than those by men in PSE, MCS, and SSH.

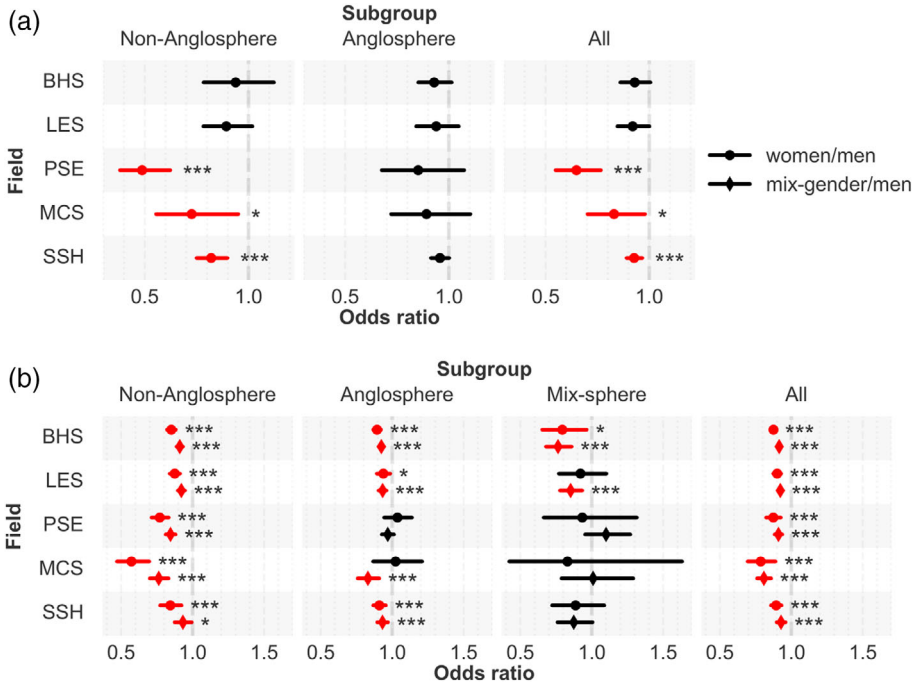


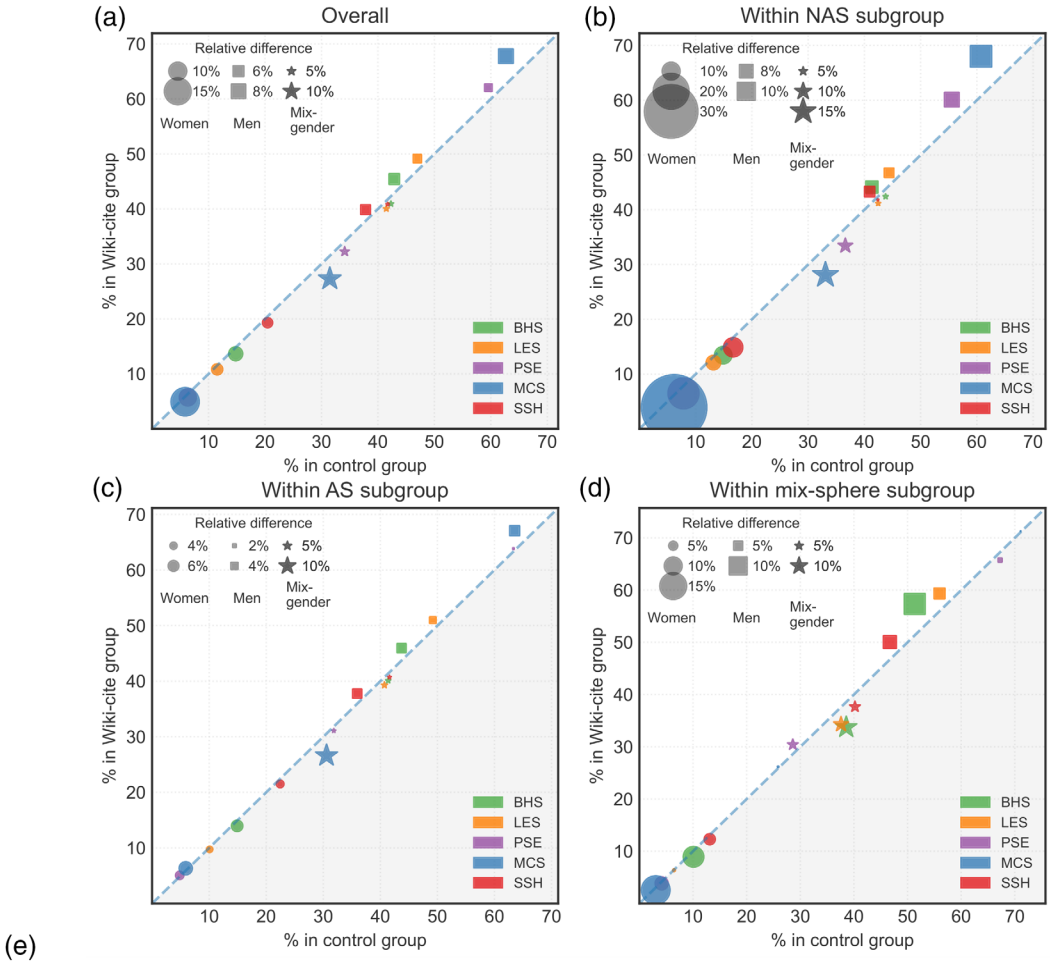
FIGURE 2 Weighted logistic regression analysis on getting cited and gender after CEM. The CEM was conducted by matching author's country of affiliation, publication venue, microlevel field, OA status, number of authors, and publication year. Men served as the reference group. Statistically significant negative odds ratios are in red color. ***$p < .001$, **$p < .01$, *$p < .05$. (a) Results of the single-author publications. Models control publication years. (b) Results of the multi-author publications. Models control publication years and number of authors

Aggregated by the country of author affiliation, Wikipedia citations to publications by women and men display different degrees of skewness. In both non-Anglosphere and Anglosphere subgroups, the percentages of publications by women are lower than expected (see Figure 1b,c). The relative differences for publications by women and men, however, are both larger in non-Anglosphere countries than in Anglosphere countries in most fields, especially PSE and MCS. The logistic regression analysis (see Figure 2a) suggests that in non-Anglosphere countries, publications by women are significantly less likely to be cited by Wikipedia than those by men in PSE, MCS, and SSH fields. In Anglosphere countries, however, we find no significant association between a publication's chance of being cited by Wikipedia and the gender of its author across all fields.

3.1.2 | Multi-author publications

For multi-author publications in the women subset (both key authors are women), the percentages of publications in the Wiki-cite group are lower than those in the control group in every field (see Figure 3a and Table S4), indicating women are cited less than expected in these fields. This trend also holds for those in the mix-gender subset (key authors are of opposite genders). Across fields, the relative differences in the women subgroup range from −15.8% (PSE) to −5.7% (SSH), and mix-gender subgroup from −13.3% (MCS) to −2.1% (SSH) (see Figure 3e). The percentages of publications in the men subgroup (both key authors are men) in the Wiki-cite group are higher than that in the control group in every field, with relative differences ranging from 4.1% (PSE) to 8.2% (MCS). The logistic regression analysis (see Figure 2b and Table S3) shows that publications by women are significantly less likely to be cited by Wikipedia than those by men across all fields. Publications in the mix-gender subgroup are also less likely to be cited than those by men across all fields.

Aggregated by the country of author affiliation (Anglosphere, non-Anglosphere, and mix-sphere), the publications by gender groups display different degrees of skewness. In the non-Anglosphere and mix-sphere subgroups, the percentages of publications by women in the Wiki-cite group are lower than that in the control group across all the five fields (see Figure 3b,d). Nonetheless, in the Anglosphere subgroup, the percentage of publications by women in the Wiki-cite group is higher than that in the control group in PSE and MCS. In the women subset, the relative differences are more prominent in non-Anglosphere and mix-sphere countries than in Anglosphere countries in most fields, particularly PSE and MCS. In the men subset, the relative differences are also more significant in non-Anglosphere and mix-sphere countries than in Anglosphere countries (see Figure 3e). The logistic regression analysis (see Figure 2b) suggests that in the non-Anglosphere subgroup, publications by women and mix-gender subgroups are significantly less likely to be cited by Wikipedia than those by men across all fields. In the Anglosphere subgroup, this is the case only in BHS, LES, and SSH: No evidence shows that in



(e)

rd (%)	NAS subgroup			AS subgroup			Mix-sphere subgroup			All		
	Women	Men	Mix-gender	Women	Men	Mix-gender	Women	Men	Mix-gender	Women	Men	Mix-gender
BHS	-9.7	6.8	-3.1	-6.3	5.1	-3.1	-11.5	11.7	-12.5	-7.8	5.9	-3.3
LES	-8.1	5.3	-3	-3.5	3.6	-3.5	-1.5	6.1	-8.8	-6.2	4.6	-3.5
PSE	-17.3	8.2	-8.7	4.6	0.8	-2.4	-7.2	-2.2	6.3	-9.5	4.1	-5.5
MCS	-36.6	12	-15.1	7.4	5.6	-13	-16	0.3	1.1	-15.8	8.2	-13.3
SSH	-10.7	5.8	-1.4	-4.2	5.1	-2.1	-6.2	7.2	-6.3	-5.7	5.4	-2.1

FIGURE 3 Gender-level comparison of multi-authored publications. Percentages in the Wiki-cite and control group and relative differences of women authors (circle), men authors (square), and mix-gender authors (star) by field in (a) the overall sample, (b) non-Anglosphere subgroup, (c) Anglosphere subgroup, and (d) mix-sphere subgroup. (e) Relative differences of women, men, and mix-gender subgroups. AS, Anglosphere; NAS, non-Anglosphere; rd, relative difference

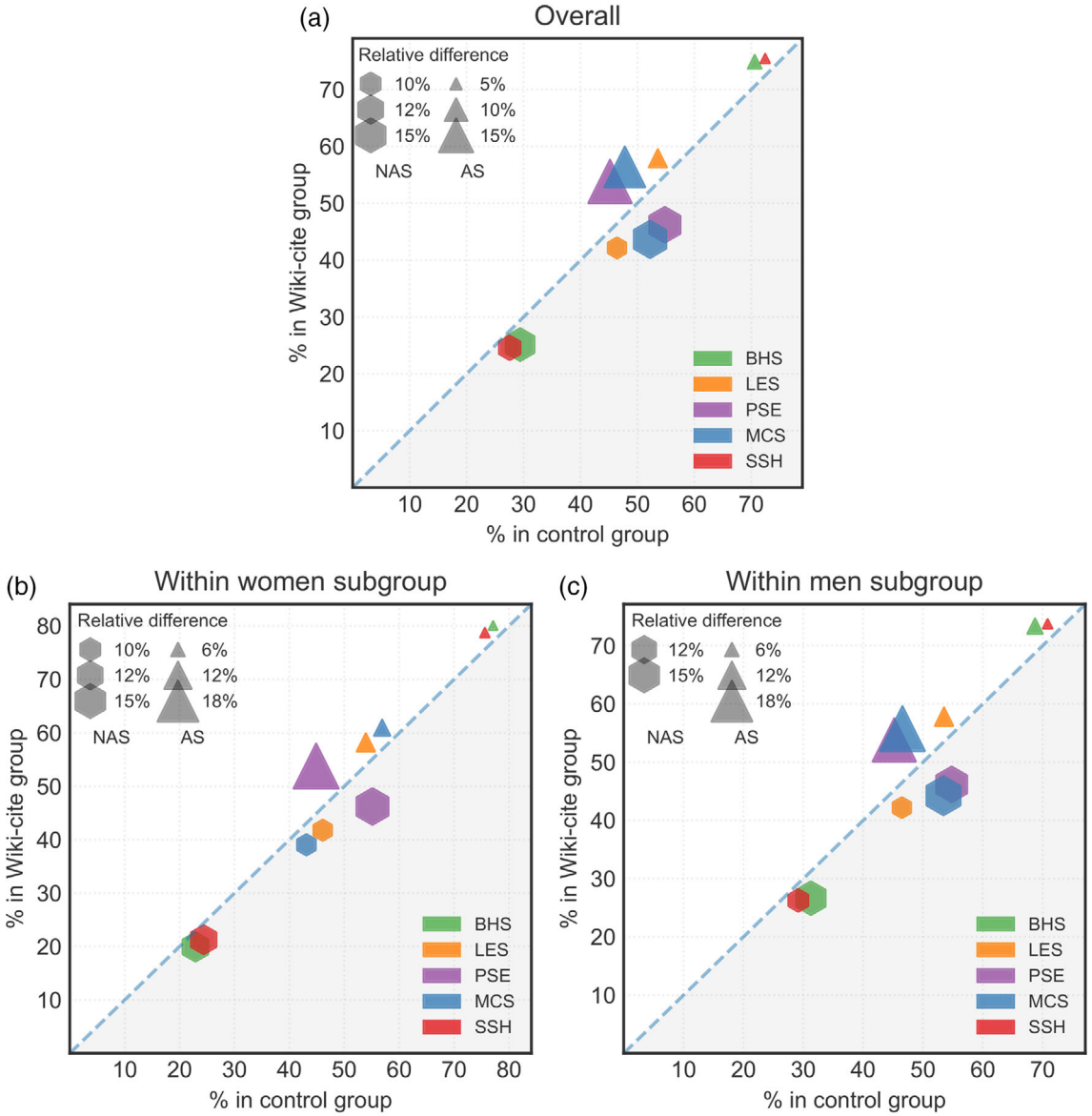
the Anglosphere subgroup, publications by women in PSE and MCS and publications by mix-gender authors in PSE are less likely to be cited by Wikipedia than those by men.

3.2 | Country-based biases

3.2.1 | Single-author publications

The percentage of publications by non-Anglosphere authors in the Wiki-cite group is lower than that in the control group in every field (see Figure 4a and Table S5); the relative differences range from −16.4% (MCS) to −9.3% (LES) (see Figure 4d). In contrast, the percentage of publications by Anglosphere authors in the Wiki-cite group is higher than that in the control group in every field, with relative differences ranging from 4.1% (SSH) to 19.1% (PSE). The logistic regression analysis (see Figure 5a and Table S6) shows that publications by non-Anglosphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors across all fields.

We further split publications based on author gender: women and men subgroups. The two subgroups are



(d)

rd (%)	Women subgroup		Men subgroup		All	
	NAS	AS	NAS	AS	NAS	AS
BHS	-13	3.9	-14.8	6.7	-14.5	6
LES	-9.3	8	-9.2	8	-9.3	8
PSE	-16.2	19.9	-15.7	19	-15.7	19.1
MCS	-9.4	7.1	-17.2	19.7	-16.4	18
SSH	-12.8	4.1	-10	4.1	-10.9	4.1

FIGURE 4 Country-level comparison of single-authored publications. Percentages in the Wiki-cite and control group and relative differences of non-Anglosphere authors (hexagon) and Anglosphere authors (triangle) by field in (a) the overall sample, (b) women subgroup, and (c) men subgroup. (e) Relative differences of non-Anglosphere and Anglosphere authors. AS, Anglosphere; NAS, non-Anglosphere; rd, relative difference

similar in terms of the degree of country-based skewness in getting cited by Wikipedia. In both subgroups, the percentages of publications by non-Anglosphere authors in the Wiki-cite group are lower than that in the control group in every field (see Figure 4b,c). The relative differences in the women subgroup are close to those in the men subgroup in all fields except MCS (see Figure 4d). The logistic regression analysis suggests that in the

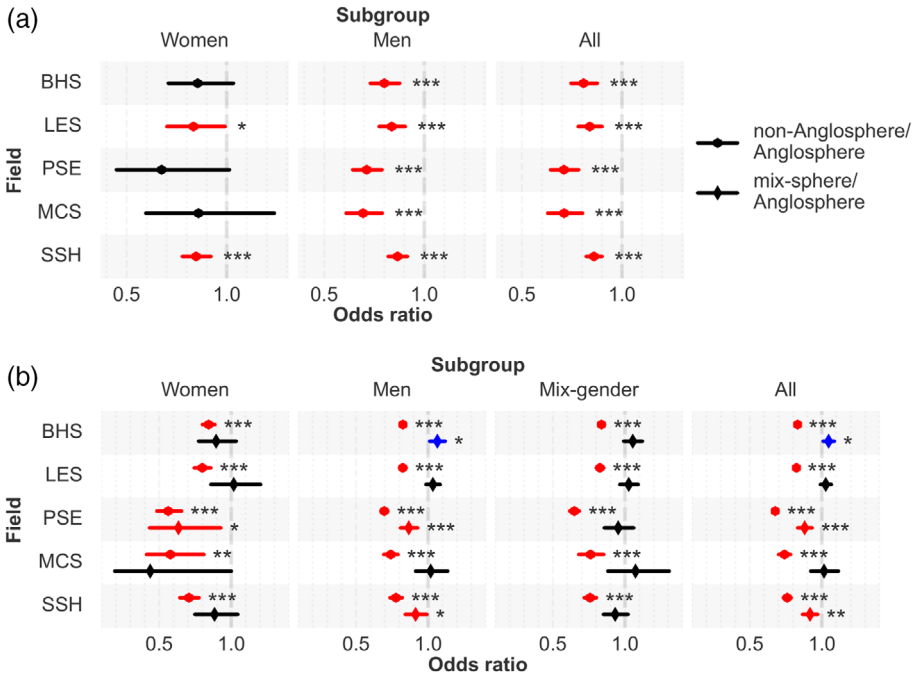


FIGURE 5 Weighted logistic regression on getting cited and country after CEM. The CEM was conducted by matching author's gender, publication venue, microlevel field, OA status, number of authors, and publication year. Anglosphere authors are the reference group. Statistically significant negative odds ratios are in red color; positive odds ratios are in blue color. ***$p < .001$, **$p < .01$, *$p < .05$. (a) Results of the single-author publications. Models control publication years. (b) Results of the multi-author publications. Models control publication years and number of authors

women subgroup, Wikipedia's preference for Anglosphere authors over non-Anglosphere authors is significant in LES and SSH. In the men subgroup, the publications by non-Anglosphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors across all fields (see Figure 5a).

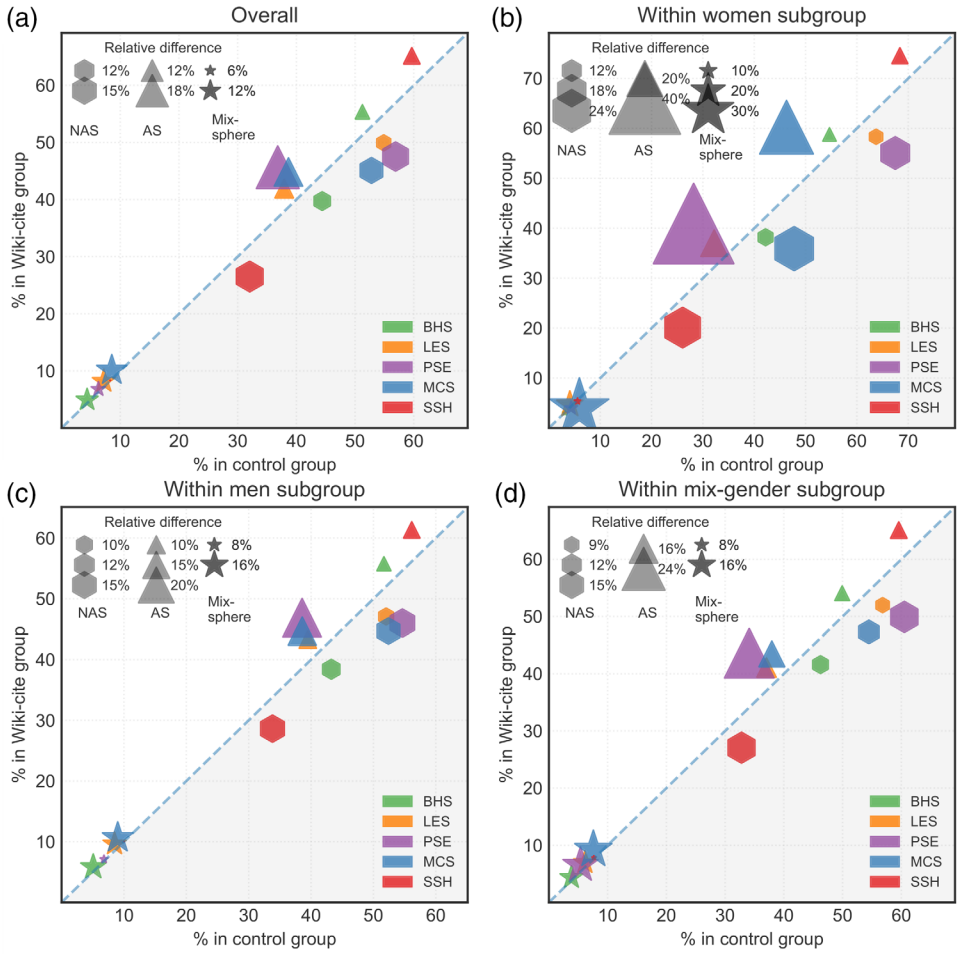
3.2.2 | Multi-author publications

The percentages of publications by non-Anglosphere authors in the Wiki-cite group are lower than those in the control group in every field (see Figure 6a and Table S7), with relative differences ranging from −17.3% (SSH) to −9% (LES) (see Figure 6e). In contrast, the percentages of publications by Anglosphere authors and by mix-sphere authors in the Wiki-cite group are both higher than those in the control group in every field. Anglosphere authors' relative differences range from 8% (BHS) to 24% (PSE), and mix-sphere authors' 0.5% (SSH) to 17.9% (MCS). The logistic regression analysis (see Figure 5b and Table S6) shows that publications by non-Anglosphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors across all fields. Publications by mix-sphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors in PSE and SSH and are more likely to be cited in BHS.

Aggregated by gender into women, men, and mix-gender subgroups, publications by country group display different degrees of skewness in getting cited by Wikipedia. The percentage of publications by non-Anglosphere authors in the Wiki-cite group is lower than that in the control group in every field. The percentages of publications by Anglosphere and by mix-sphere authors in the Wiki-cite group are both higher than those in the control group in most fields. Yet, among publications in the women subgroup, those by mix-sphere authors in the Wiki-cite group are larger than those in the control group only in LES (see Figure 6b,d). The relative differences for publications by non-Anglosphere, Anglosphere, and mix-sphere authors are larger in women subgroups than in other gender subgroups in most fields, especially PSE and MCS. The logistic regression analysis (see Figure 5b) suggests that across all subgroups and fields, the publications by non-Anglosphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors. For mix-sphere authors, however, their publications are significantly less likely to be cited by Wikipedia than those by Anglosphere authors only in PSE (for both the women and men subgroups) and SSH (for the men subgroup).

4 | DISCUSSION

Given Wikipedia's dominant role in disseminating knowledge in the public domain, it is critical to ensure that Wikipedia's representation of knowledge does not mirror its imbalanced editorial distribution (Glott et al., 2010; Hill & Shaw, 2013; Lam et al., 2011; Wagner et al., 2016; Wikimedia Foundation, 2011). Yet, our



(e)

rd (%)	Women subgroup			Men subgroup			Mix-gender subgroup			All		
	NAS	AS	Mix-sphere	NAS	AS	Mix-sphere	NAS	AS	Mix-sphere	NAS	AS	Mix-sphere
BHS	-9.5	7.5	-2.9	-11.2	7.9	14.8	-10	8.3	13.4	-10.5	8	13
LES	-8.5	14.9	15.4	-9.5	9.7	13.5	-8.5	10.9	13.4	-9	10.5	13.5
PSE	-18.4	45.2	-7.6	-15.8	21.6	4.7	-17.5	27.8	20.7	-16.5	24	8.5
MCS	-24.8	30.1	-34.8	-14.7	15.8	18.1	-13.2	14.5	22	-14.6	16	17.9
SSH	-22.8	9	-3.8	-15.3	9.2	0	-17.4	9.3	2.4	-17.3	9.2	0.5

FIGURE 6 Country-level comparison of multi-authored publications. Percentages in the Wiki-cite and control group and relative differences of non-Anglosphere authors (hexagon), Anglosphere authors (triangle), and mix-sphere authors (star) by main field in (a) the overall sample, (b) women subgroup, (c) men subgroup, and (d) mix-gender subgroup. (e) Relative differences of NAS, AS, and mix-sphere authors. AS, Anglosphere; NAS, non-Anglosphere; rd, relative difference

results showed that Wikipedia citations to scholarly publications reflect its known asymmetries for specific gender and country groups. Wikipedia's preference for citing scholarly publications by men is prominent. Publications with women in key authorship positions are significantly less likely to be cited by Wikipedia than those with men. This pattern is consistent across most fields, with the exceptions of BHS and LES when concerning single-authored publications, which might be related to a more balanced gender composition in these two fields (Leslie et al., 2015). While combined with countries of affiliation, we found that although Wikipedia less favors single-authored publications by women than those by men in non-Anglosphere countries, this bias does not necessarily hold for Anglosphere countries. This indicates an apparent disadvantage of women scholars in non-Anglosphere countries for having proper representations in the world's largest knowledge dissemination platform. Regarding multi-authored publications, women's disadvantage is significant in both non-Anglosphere and Anglosphere groups: Publications with women as key authors are less likely to get cited by Wikipedia in both Anglosphere and non-Anglosphere countries, except for PSE and MCS in the Anglosphere subgroup. Our study also

revealed country biases of Wikipedia citations: non-Anglosphere authors' publications are less cited by Wikipedia, while Anglosphere authors' publications are more cited. For multi-authored publications by mix-sphere authors, their percentage in the Wiki-cite group generally outstrips that in the control group for the men and mix-gender subgroups. Publications by mix-sphere authors are significantly less likely to be cited by Wikipedia than those by Anglosphere authors in PSE and SSH, suggesting a potential mitigation effect by having a key author affiliated with Anglosphere countries.

This study also confirmed the field differences in the Wikipedia citation biases: The gender gaps between the Wiki-cite group and the control group in PSE and MCS are generally larger than in other fields. As math-intensive STEM fields, PSE and MCS are known to suffer from gender disparities: Women in these fields are underrepresented, have less productivity and impact (Huang et al., 2020), and receive fewer awards (Meho, 2021); their contribution is more likely to be devalued (Ni et al., 2021). Research partially attributed these to the fact that women are stigmatized by a stereotyped ability belief that they are less likely to be gifted with the natural brilliance to succeed in these fields. This will likely also prevent women's publications from being acknowledged fairly by others, including Wikipedia. BHS, LSE, and SSH do not show equally large gaps, possibly because women participate and publish in these fields more (Thelwall & Nevill, 2019) and feel more sense of belonging and self-efficacy in a milder masculine culture (Cheryan et al., 2017). More importantly, our analysis of the intersectional impact of gender and country further reveals that such a gender gap in PSE and MCS mainly exists for non-Anglosphere authors. Even within the subgroup of women authors, compared with non-Anglosphere or mix-sphere women authors, women authors from Anglosphere countries still gain a remarkable advantage in receiving Wikipedia citations. Given Wikipedia's strong global impact on the public and the professionals, the gender-country intersectional biases in Wikipedia citations may further marginalize women authors in non-Anglosphere countries in the science ecosystem.

Future research may examine the causal mechanisms of the gender- and country-based biases discovered in this study. One possible mechanism is Wikipedia editors' homophily for information source selection. Homophily is the tendency that people prefer other people, works, or ideas similar to their own, which has been observed in other science gatekeeping processes, such as peer review (Murray et al., 2019). Homophily may explain the found biases in this study as English Wikipedia editors have been suggested to be men- and Anglosphere-dominated (see section 1). Given that the categorization of Anglosphere and non-Anglosphere countries is based more on cultural and language ties (Legrand, 2016), it is possible that due to homophily, publications with key authors being men and affiliated with Anglosphere countries are more likely to attract Wikipedia editors who belong to the same social demographic group. They may favor and prioritize publications by men and Anglosphere authors who are similar to them to protect their demographic group's benefits and social standing (Murray et al., 2019).

The gender and country disparities in the symbolic capital system may also contribute to the observed biases. Symbolic capital is the degree of accumulated prestige, celebrity, consecration, or honor founded on a dialectic of knowledge and recognition (Thompson, 1991). By this definition, an author's symbolic capital may involve multiple aspects, such as the prestige of affiliations, career rank, productivity, scholarly impact, awards, and personal reputation. There is credible evidence that women scientists are disadvantaged in authorship allocation, citation accumulation, honors and awards, and career development locally and globally (Kozlowski et al., 2022; Larivière et al., 2013; Murray et al., 2019; Ni et al., 2021). These disadvantages may restrict women's symbolic capital within the scientific communities, forming a vicious circle of glass ceilings and lower performance (Van Den Besselaar & Sandström, 2017) and downplaying women's contributions and achievements (Matilda effect) (Rossiter, 1993). Furthermore, Anglosphere countries, especially the United States and the United Kingdom, are top global scientific research powerhouses and are advantaged in their language, as English is the dominant language for modern scholarly communication (Ammon, 2011). Anglosphere countries' economic, political, and cultural hegemony and the Eurocentrism worldview still shape global knowledge production (Castro Torres & Alburez-Gutierrez, 2022). These factors may lead Wikipedia editors to pick publications consciously or unconsciously by men and Anglosphere authors, who possess more symbolic capital, over other similar publications by authors of other gender and country groups. The effect of different types of symbolic capitals remains to be tested further.

5 | CONCLUSION

Our study adds new evidence from the perspective of knowledge dissemination to the existing gender- and country-based, as well as the intersectional inequalities in science. Giving scholars of different social demographic groups across countries equal rights and opportunities to spread their knowledge and increase visibility should be vital to science equity. Strategies are needed

from Wikipedia and the scientific communities to ensure the fairness of knowledge dissemination and advance gender equality and decolonization in science.

Like many other studies using matching methods, our research is not immune to potential biases. First, CEM can only match observable variables when comparing two groups (Stuart, 2010). This indicates that the estimation may be biased by other unobservable variables that are not measured. In our study, although we balanced the publication quality, impact, and topical areas to some extent, we could not cover the unobservable variables, such as each publication's specific scientific contribution. This study is also likely limited by relying on the fine-grained microfield classification by CTWS, as a microfield may cover more than one topic. Although it was our goal to precisely match publications at an even lower granularity level (topic), we were limited by the resources needed for this task. Second, the Wikipedia dataset we used does not incorporate the editing history of the Wikipedia citations, restricting the possibility of observing and controlling variables that may vary over time, such as the number of academic citations attracted by publications. Further research will benefit from using more advanced datasets and causal inference methods that counter the confounding effects more effectively. Third, like in many other science of science studies, we have limited avenues to reflect the authors' symbolic capital and thus cannot decide the precise reason for the biases in Wikipedia citation to scholarly publications. Future research may consider implementing surveys to systematically collect relevant data under one universal framework to represent their symbolic capital.

AUTHOR CONTRIBUTIONS

Xiang Zheng and Chaoqun Ni designed research. Jiajing Chen and Chaoqun Ni collected and processed data. Xiang Zheng and Chaoqun Ni analyzed data. Xiang Zheng, Jiajing Chen, Erjia Yan, and Chaoqun Ni wrote the paper.

ACKNOWLEDGMENTS

We thank Dr. Giovanni Colavizza's response to our questions about the Wikipedia citation dataset. We thank CWTS at Leiden University for sharing the 2021 CWTS publication-level classification data and Observatoire des sciences et des technologies at the University of Quebec in Montreal for access to the Web of Science data. We also thank Dr. Ian Hutchins and Alison Tollas for constructive feedback on earlier drafts of the manuscript.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The Wikipedia citation dataset is available on Zenodo by Singh et al. (2020). This study's de-identified data and codes are available on GitHub (https://github.com/UWMadisonMetaScience/wikibias). For the complete datasets of CWTS publication-level classification and Web of Science, readers can contact CWTS at Leiden University (https://www.cwts.nl/) and Clarivate Analytics (https://clarivate.com/).

ORCID

Xiang Zheng https://orcid.org/0000-0002-6619-5504
Erjia Yan https://orcid.org/0000-0002-0365-9340
Chaoqun Ni https://orcid.org/0000-0002-4130-7602

REFERENCES

Ammon, U. (Ed.) (2011). The dominance of English as a language of science: Effects on other languages and language communities. In *The dominance of English as a language of science*. De Gruyter Mouton. https://doi.org/10.1515/9783110869484

Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., & Romero-Frías, E. (2020). Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLoS One*, *15*(2), e0228713. https://doi.org/10.1371/journal.pone.0228713

Benjakob, O., Aviram, R., & Sobel, J. A. (2022). Citation needed? Wikipedia bibliometrics during the first wave of the COVID-19 pandemic. *GigaScience*, *11*, giab095. https://doi.org/10.1093/gigascience/giab095

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: Coarsened exact matching in Stata. *Stata Journal*, *9*(4), 524–546. https://doi.org/10.1177/1536867X0900900402

Callahan, E., & Herring, S. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, *62*(10), 1899–1915. https://doi.org/10.1002/asi.21577

Castro Torres, A. F., & Alburez-Gutierrez, D. (2022). North and south: Naming practices and the hidden dimension of global disparities in knowledge production. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(10), e2119373119. https://doi.org/10.1073/pnas.2119373119

Centre for Science and Technology Studies (2022). Fields. In *CWTS Leiden ranking*. Author. Retrieved from http://www.leidenranking.com

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35. https://doi.org/10.1037/bul0000052

Eijkman, H. (2010). Academics and Wikipedia: Reframing Web 2.0+ as a disruptor of traditional academic power-knowledge arrangements. *Campus-Wide Information Systems*, *27*(3), 173–185. https://doi.org/10.1108/10650741011054474

Ford, H., Graham, M., & Meyer, E. (2015). *Fact factories: Wikipedia and the power to represent* (PhD thesis). University of Oxford. Retrieved from https://ora.ox.ac.uk/objects/uuid:b34fdd6c-ec15-4bcd-acba-66a777739b4d

Ford, H., Sen, S., Musicant, D. R., & Miller, N. (2013). Getting to the source: Where does Wikipedia get its information from? In *Proceedings of the 9th international symposium on open collaboration* (pp. 1–10). ACM. https://doi.org/10.1145/2491055.2491064

Ford, H., & Wajcman, J. (2017). "Anyone can edit," not everyone does: Wikipedia's infrastructure and the gender gap. *Social Studies of Science*, *47*(4), 511–527. https://doi.org/10.1177/0306312717692172

Glott, R., Schmidt, P., & Ghosh, R. (2010). *Wikipedia survey—Overview of results*. UNU-MERIT. Retrieved from https://www.merit.unu.edu/wp-content/uploads/2019/03/Wikipedia_Overview_15March2010-FINAL.pdf

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, *104*(4), 746–764. https://doi.org/10.1080/00045608.2014.910087

Halavais, A., & Lackaff, D. (2008). An analysis of topical coverage of Wikipedia. *Journal of Computer-Mediated Communication*, *13*(2), 429–440. https://doi.org/10.1111/j.1083-6101.2008.00403.x

Hew, K. F., & Cheung, W. S. (2009). Use of wikis in K-12 and higher education: A review of the research. *International Journal of Continuing Engineering Education and Life Long Learning*, *19*(2–3), 141–165. https://doi.org/10.1504/IJCEELL.2009.025024

Heydari, S., Shekofteh, M., & Kazerani, M. (2019). Relationship between altmetrics and citations a study on the highly cited research papers. *DESIDOC Journal of Library & Information Technology*, *39*(4), 169–174. https://doi.org/10.14429/djlit.39.4.14204

Hill, B. M., & Shaw, A. (2013). The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PLoS One*, *8*(6), e65782. https://doi.org/10.1371/journal.pone.0065782

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Huang, Y., Bu, Y., Ding, Y., & Lu, W. (2019). From zero to one: A perspective on citing. *Journal of the Association for Information Science and Technology*, *70*(10), 1098–1107. https://doi.org/10.1002/asi.24177

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, *20*(1), 1–24. https://doi.org/10.1093/pan/mpr013

Jemielniak, D., & Aibar, E. (2016). Bridging the gap between wikipedia and academia. *Journal of the Association for Information Science and Technology*, *67*(7), 1773–1776. https://doi.org/10.1002/asi.23691

Jemielniak, D., Masukume, G., & Wilamowski, M. (2019). The most influential medical journals according to Wikipedia: Quantitative analysis. *Journal of Medical Internet Research*, *21*(1), e11429. https://doi.org/10.2196/11429

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: Benefits, critiques, & a survival guide. *EJIFCC*, *25*(3), 227–243.

Kozlowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(2), e2113067119. https://doi.org/10.1073/pnas.2113067119

Lam, S. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., & Riedl, J. (2011). WP: clubhouse? An exploration of Wikipedia's gender imbalance. In *WikiSym 2011 conference proceedings: 7th annual international symposium on Wikis and open collaboration* (pp. 1–10). ACM. https://doi.org/10.1145/2038558.2038560

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, *66*(7), 1323–1332. https://doi.org/10.1002/asi.23266

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a

Laurent, M. R., & Vickers, T. J. (2009). Seeking health information online: Does Wikipedia matter? *Journal of the American Medical Informatics Association*, *16*(4), 471–479. https://doi.org/10.1197/jamia.M3059

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

Legrand, T. (2016). Elite, exclusive and elusive: Transgovernmental policy networks and iterative policy transfer in the Anglosphere. *Policy Studies*, *37*(5), 440–455. https://doi.org/10.1080/01442872.2016.1188912

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265. https://doi.org/10.1126/science.1261375

MacHado, D. B., Pescarini, J. M., Ramos, D., Teixeira, R., Lozano, R., Pereira, V. O. D. M., Azeredo, C., Paes-Sousa, R., Malta, D. C., & Barreto, M. L. (2020). Monitoring the progress of health-related sustainable development goals (SDGs) in Brazilian states using the Global Burden of Disease indicators. *Population Health Metrics*, *18*, 7. https://doi.org/10.1186/s12963-020-00207-2

Malcolm, D. (2021). Cricket, Brexit and the Anglosphere. *Sport in Society*, *24*(8), 1274–1290. https://doi.org/10.1080/17430437.2021.1876030

Meho, L. I. (2021). The gender gap in highly prestigious international research awards, 2001–2020. *Quantitative Science Studies*, *2*(3), 976–989. https://doi.org/10.1162/qss_a_00148

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F., & Lanamaki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, *66*(2), 219–245. https://doi.org/10.1002/asi.23172

Murray, D., Siler, K., Larivière, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2019). Author-reviewer homophily in peer review. BioRxiv: 10.1101/400515.

Ni, C., Smith, E., Yuan, H., Larivière, V., & Sugimoto, C. R. (2021). The gendered nature of authorship. *Science Advances*, *7*(36), eabe4639. https://doi.org/10.1126/sciadv.abe4639

Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday*, 12(8). https://doi.org/10.5210/fm.v12i8.1997

Okoli, C., Mehdi, M., Mesgari, M., Nielsen, F. Å., & Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. *Journal of the Association for Information Science and*

Technology, *65*(12), 2381–2403. https://doi.org/10.1002/asi.23162

Orduna-Malea, E., Thelwall, M., & Kousha, K. (2017). Web citations in patents: Evidence of technological impact? *Journal of the Association for Information Science and Technology*, *68*(8), 1967–1974. https://doi.org/10.1002/asi.23821

Priem, J. (2014). Altmetrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: Harnessing multidimensional indicators of performance* (pp. 340–374). MIT Press.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto*. Retrieved from http://altmetrics.org/manifesto/

Reagle, J. (2013). "Free as in sexist?" Free culture and the gender gap. *First Monday*, 18(1). https://doi.org/10.5210/fm.v18i1.4291

Reich, E. S. (2011). Online reputations: Best face forward. *Nature*, *473*(7346), 138–139. https://doi.org/10.1038/473138a

Rossiter, M. W. (1993). The Matthew Matilda effect in science. *Social Studies of Science*, *23*(2), 325–341. https://doi.org/10.1177/030631293023002004

Shuai, X., Jiang, Z., Liu, X., & Bollen, J. (2013). A comparative study of academic and Wikipedia ranking. In *Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries (JCDL '13)* (pp. 25–28). ACM. https://doi.org/10.1145/2467696.2467746

Singh, H., West, R., & Colavizza, G. (2020). Wikipedia Citations: A comprehensive dataset of citations with identifiers extracted from English Wikipedia [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3940692

Singh, H., West, R., & Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies*, *2*(1), 1–19. https://doi.org/10.1162/qss_a_00105

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Sugimoto, C. R., Ni, C., & Larivière, V. (2015). On the relationship between gender disparities in scholarly communication and country-level development indicators. *Science and Public Policy*, *42*(6), 789–810. https://doi.org/10.1093/scipol/scv007

Sundin, O. (2011). Janitors of knowledge: Constructing knowledge in the everyday life of Wikipedia editors. *Journal of Documentation*, *67*(5), 840–862. https://doi.org/10.1108/00220411111164709

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, *68*(9), 2116–2127. https://doi.org/10.1002/asi.23687

Thelwall, M. (2020). Female citation impact superiority 1996–2018 in six out of seven English-speaking nations. *Journal of the Association for Information Science and Technology*, *71*(8), 979–990. https://doi.org/10.1002/asi.24316

Thelwall, M., & Kousha, K. (2016). Are Wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology*, *68*(3), 762–779. https://doi.org/10.1002/asi.23694

Thelwall, M., & Nevill, T. (2019). No evidence of citation bias as a determinant of STEM gender disparities in US biochemistry, genetics and molecular biology research. *Scientometrics*, *121*(3), 1793–1801. https://doi.org/10.1007/s11192-019-03271-0

Thompson, J. B. (1991). Introduction. In P. Bourdieu (Ed.), *Language and symbolic power*. Harvard University Press.

Tomaszewski, R., & MacDonald, K. I. (2016). A study of citations to Wikipedia in scholarly publications. *Science & Technology Libraries*, *35*(3), 246–261. https://doi.org/10.1080/0194262X.2016.1206052

Van Den Besselaar, P., & Sandström, U. (2017). Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLoS One*, *12*(8), e0183301. https://doi.org/10.1371/journal.pone.0183301

Vucetic, S. (2011). *The Anglosphere: A genealogy of a racialized identity in international relations*. Stanford University Press.

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: Gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5. https://doi.org/10.1140/epjds/s13688-016-0066-4

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Wikimedia Foundation. (2011). *Wikipedia editors study: Results from the editor survey, April 2011*. Author.

Wikimedia Foundation. (2022). *Wikimedia Statistics—English Wikipedia*. Author. Retrieved from https://stats.wikimedia.org/#/en.wikipedia.org

Wikipedia. (2022a). *Wikipedia: Recentism*. Author. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Recentism

Wikipedia. (2022b). *Wikipedia: Verifiability*. Author. Retrieved from https://en.wikipedia.org/wiki/Wikipedia:Verifiability

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics*, *126*(11), 8861–8886. https://doi.org/10.1007/s11192-021-04171-y

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zheng, X., Chen, J., Yan, E., & Ni, C. (2023). Gender and country biases in Wikipedia citations to scholarly publications. *Journal of the Association for Information Science and Technology*, *74*(2), 219–233. https://doi.org/10.1002/asi.24723

APPENDIX A

A.1 | Coarsened exact matching

We used the coarsened exact matching (CEM) method to construct the control group of publications. CEM is a nonparametric matching method to reduce confounding effects by creating two new groups with balanced covariates, only different in whether they have been treated. Instead of exact matching, CEM groups ("coarsening") each of the n covariates into m_i nonoverlapping bins by predetermined cutoffs and then creating $\prod_{i=1}^{n} m_i$ strata by the Cartesian product of all sets of bins. Each unit is assigned to a stratum according to its coarsened covariates. Units in strata that do not contain at least one unit from the treatment group and one unit from the control group are removed. Units that share the same stratum will be matched. CEM also assigns weight to the units to ensure that the treated and control groups have the same proportioned distribution of units across strata.

To ensure that the Wiki-cite and control groups have the same proportioned distribution of publications across strata divided by the publication-level features, CEM assigns weight w_i to each matched unit i in stratum s by

$$w_i = \begin{cases} 1, \ i \in \text{Wiki}-\text{cite group} \\ \dfrac{m_{st}\sum m_c}{m_{sc}\sum m_t}, \ i \in \text{control group}, \end{cases}$$

where m_{st} and m_{sc} are the numbers of publications belonging to the Wiki-cite and control group, respectively, in stratum s; $\sum m_t$ and $\sum m_c$ are the numbers of units belonging to the Wiki-cite and control group, respectively, in all strata (Iacus et al., 2012).

A.2 | Relative difference

We calculated the relative difference between the Wiki-cite group and the control group using the following formula:

$$rd = \frac{P_{\text{wiki}-\text{cite}} - P_{\text{control}}}{P_{\text{control}}},$$

where rd denotes relative difference, $P_{\text{wiki}-\text{cite}}$ is the percentage of publications in a social group (gender or country) in the Wiki-cite group after CEM weighting, and P_{control} is the percentage of the group's publications in the control group after CEM weighting.