Research Articles

Self-citation and corruption: cross-sectional, cross-country study

Alexander C. Tsai[1]

[1] Center for Global Health and Mongan Institute, Massachusetts General Hospital, Boston, Massachusetts, USA; Harvard Medical School, Boston, Massachusetts, USA; Mbarara University of Science and Technology, Mbarara, Uganda

Keywords: citation manipulation, self-citation, corruption

https://doi.org/10.29392/001c.24588

Journal of Global Health Reports

Vol. 5, 2021

Background

Self-citation appears to be widely prevalent. However, the structural drivers of self-citation are poorly understood.

Methods

Data for this study were obtained from a recently published study of Scopus data aggregated across all authors with >5 publications, across all scientific fields, which yielded aggregate, country-level data on the mean co-author self-citation rate for the period 1960-2018. These data were merged with 2018 data from Transparency International on corruption, and additional data extracted from the World Development Indicators. The country-level association between the self-citation rate and the corruption index was estimated using multivariable linear regression.

Results

Across 178 countries, the correlation between the mean self-citation rate and the corruption index was -0.52, 95% confidence interval, CI=-0.62 to -0.41. Among the 49 countries in the lowest quartile of the corruption index, the mean self-citation rate was 0.24 (standard deviation, SD=0.06). Among the 44 countries in the highest quartile of the corruption index, the mean self-citation rate was 0.21 (SD=0.05). In a weighted linear regression model with robust estimates of variance, the corruption index had a statistically significant association with the mean self-citation rate (2nd quartile compared with 1st quartile: *b*=-0.08 (95% CI=-0.17 to -0.01); 3rd quartile: *b*=-0.11 (95% CI=-0.19 to -0.02); 4th quartile: *b*=-0.10 (95% CI=-0.19 to -0.01; N=165). The implied effect size was large in magnitude and robust to potential confounding by unmeasured covariates.

Conclusions

In this cross-sectional, cross-country analysis, there was a strong correlation between a country's overall level of corruption and the mean self-citation rate. The estimated association was statistically significant, large in magnitude, and unlikely to be explained away by unmeasured confounding. Better understanding of how corruption norms evolve is likely to be critical in addressing the problem of extreme self-citation and other forms of citation manipulation.

Self-citation, which occurs when the authors of a published journal article cite a previously published journal article in which any of their names appear as authors,[1] is widely employed by researchers to disseminate their findings and influence the trajectory of the literature.[2,3] This behavior is not inappropriate by definition, given that it can also reflect genuine acknowledgment of scientific influence and priority.[4,5] Further, while this behavior is most often discussed in reference to individual authors,[6] it can also be a characteristic of journals and journals editors (i.e., to inflate journal impact factors)[7,8] and institutions.[9]

One type of self-citation behavior, extreme self-citation[10] – which occurs when an inordinately large proportion of an author's total citation count is derived from self-citation behavior – has also been described in the literature. No specific threshold proportion (i.e., of self-citations relative to total citations) has been identified, but Ioannidis et al.[10] identified more than 250 researchers for whom at least half of their total citations were derived from self-citation. This practice is a problematic behavior in academic research, as it is a form of citation manipulation that distort decisions about hiring, promotions, and research funding. There is debate in the literature about the extent to which men are more likely to engage in self-citation compared with women,[11–13] but other structural drivers of self-citation are poorly understood.

METHODS

DATA SOURCES

Data for this study were obtained from Baas et al.[14] In brief, Scopus data were aggregated across all authors with >5 publications, across all scientific fields, to calculate the mean co-author self-citation rate for 1960-2018.[10,14] These data were merged with: 2018 data from Transparency International, which calculates an annual country-level Corruption Perceptions Index, a composite measure based on data from a variety of sources intended to measure "the overall extent of corruption (frequency and/or size of bribes) in the public or political sectors"[15]; and 2018 data on per capita gross domestic product and the total number of scientific and engineering articles, both extracted from the World Development Indicators.[16]

STATISTICAL ANALYSIS

First, I estimated the correlation, at the country-level, between the mean self-citation rate and the Corruption Perceptions Index. I estimated the mean self-citation rate in each quartile of the Corruption Perceptions Index. Then I used linear regression with robust estimates of variance to estimate the association between the two variables, specifying the mean self-citation rate as the dependent variable and quartiles of the Corruptions Perception Index as the primary explanatory variables of interest, adjusting for per capita gross domestic product and the total scientific publication output. Observations were weighted by the total number of authors, computed by Baas et al.[14]

I conducted several sensitivity analyses. First, I specified the median rather than the mean self-citation rate as the dependent variable. Second, I used the e-value to estimate the degree of unmeasured confounding that would be needed to completely explain the observed association.[17] Third, I used the method proposed by Oster[18] to estimate the extent to which selection on unmeasured variables would be, relative to selection on measured variables (per capita gross domestic product and total scientific publication output), to completely explain the observed association.

RESULTS

Across countries, the mean self-citation rate was 0.23 (standard deviation, SD=0.06) and the median self-citation rate was 0.18 (interquartile range, 0.15-0.22) (N=228). The countries with the 10 highest median self-citation rates (ranging from 0.29 to 0.42) were: Guinea Bissau, Timor-Leste, Ukraine, Kazakhstan, Russia, Armenia, Afghanistan, Indonesia, North Korea, and Moldova. The mean number of authors represented per country was 31,553 (SD=138314; median=563; interquartile range=54-6735). The Corruption Perceptions Index (mean=43.1; SD=19.1) was available for 180 countries in the dataset.

Across countries, the correlation between the mean self-citation rate and the Corruption Perceptions Index was -0.52 (95% confidence interval, CI=-0.62 to -0.41; N=178) (Figure 1). Among the 49 countries in the lowest quartile of the Corruption Perceptions Index, the mean self-citation rate was 0.24 (SD=0.06). Among the 44 countries in the highest quartile of the Corruption Perceptions Index, the mean self-citation rate was 0.21 (SD=0.05).

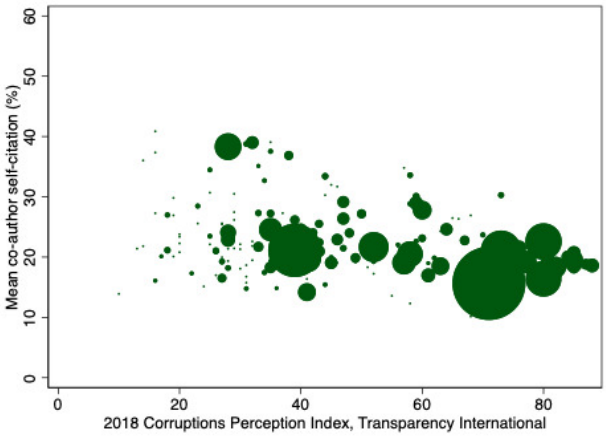

Figure 1. Country-level correlation between corruption and self-citation rate (N=178)

In a weighted linear regression model with robust estimates of variance, the Corruption Perceptions Index had a statistically significant association with the mean self-citation rate (2nd quartile compared with 1st quartile: *b*=-0.06 (95% CI=-0.15 to 0.04); 3rd quartile: *b*=-0.10 (95% CI=-0.19 to -0.02); 4th quartile: *b*=-0.13 (95% CI=-0.22 to -0.04); N=178). After adjustment for per capita gross domestic product and total scientific publication output, these estimates remained statistically significant (2nd quartile compared with 1st quartile: *b*=-0.08 (95% CI=-0.17 to 0.01); 3rd quartile: *b*=-0.11 (95% CI=-0.19 to -0.02); 4th quartile: b=-0.10 (95% CI=-0.19 to -0.01); N=165). The implied effect size was large in magnitude (3rd quartile compared with 1st quartile, Cohen's *d*=-1.60; 4th quartile, Cohen's *d*=-1.54).

Specifying the median self-citation rate as the dependent variable did not substantively shift the estimated associations (2nd quartile compared with 1st quartile: *b*=-0.08 (95% CI=-0.18 to 0.01); 3rd quartile: *b*=-0.12 (95% CI=-0.21 to -0.03); 4th quartile: *b*=-0.11 (95% CI=-0.20 to -0.01); N=165). The e-value associated with the highest quartile of the Corruption Perceptions Index was 7.58 for the point estimate and 5.13 for the confidence interval, indicating that an unmeasured confounding variable would need to have a very strong association with both corruption and self-citation (greater than 5 on the risk ratio scale) in order to shift the confidence interval of the estimated association to include zero. The R-squared from the regression model with all covariates was 0.60, so I assumed a maximum R-squared value of 0.60 × 1.3 = 0.78 in applying the procedures described by Oster.[18] I calculated a delta of 7.23, indicating that the model for the association between corruption and self-citation is fairly robust: selection on unmeasured variables would need to be more than 7 times as important as selection on the measured variables to generate an estimated regression coefficient equal to zero.

DISCUSSION

In this cross-sectional, cross-country analysis of data from 178 countries, I estimated a strong correlation between the country's overall level of corruption and the mean self-citation rate. The estimated association was statistically significant, large in magnitude, and unlikely to be explained away by unmeasured confounding. My findings are consistent with prior work from Italy showing that self-citation behavior can be shifted dramatically in response to incentives.[19,20] What my analysis adds is an assessment of country-level norms in explaining a behavior that is widely understood to be a form of citation manipulation. In this regard, my findings are consistent with those of Fisman and Miguel,[21] who found that New York City-based foreign diplomats from high-corruption countries were more likely than diplomats from low-corruption countries to accumulate unpaid parking tickets.

Interpretation of my findings is subject to several important limitations. First, unlike the study by Fisman and Miguel,[21] the data are ecological in nature and therefore potentially subject to the ecological fallacy: it would be an overreach to conclude that individuals from more corrupt countries are more likely to engage in self-citation behavior, or to conclude that individuals who are more corrupt are more likely to engage in self-citation behavior. Second, the ecological variables used in this analysis were aggregate (derived) measures, and no covariate adjustment for individual-level variables was used.[22] Third, the estimated association between corruption and self-citation could potentially be confounded by unmeasured variables. Fourth, and relatedly, the self-citation measures were based on data from Scopus,[10,14] which is known to have significantly broader coverage of journals and research output compared with other leading scholarly databases such as Web of Science.[23] If the coverage is over-inclusive of lower-quality research output characterized by a greater rate of self-citation, and is also over-inclusive of research output from countries with a higher degree of corruption, the estimated association between corruption and self-citation could be biased away from the null. However, the sensitivity analyses indicate that only *very* strong confounding could completely explain the estimated associations.

CONCLUSIONS

The limitations notwithstanding, my analysis shows that, at the country level, corruption is strongly associated with self-citation. Better understanding of how corruption norms evolve is likely to be critical in addressing the problem of extreme self-citation and other forms of citation manipulation.

ACKNOWLEDGMENTS

I thank Jeroen Baas (Director, Funding Content & Analytics, Elsevier Research Intelligence) for providing the data on self-citation.

FUNDING

None.

COMPETING INTERESTS

The author completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available upon request from the corresponding author) and declares no conflicts of interest.

CORRESPONDENCE TO:

Alexander C. Tsai
Center for Global Health and Mongan Institute, Massachusetts General Hospital, Boston, Massachusetts, USA.
actsai@mgh.harvard.edu

Submitted: January 24, 2021 GMT, Accepted: June 03, 2021 GMT

REFERENCES

1. Carley S, Porter AL, Youtie J. Toward a more precise definition of self-citation. *Scientometrics*. 2013;94(2):777-780. doi:10.1007/s11192-012-0745-2

2. Bartneck C, Kokkelmans S. Detecting h-index manipulation through self-citation analysis. *Scientometrics*. 2011;87(1):85-98. doi:10.1007/s11192-010-0306-5

3. James F, Dag A. Does self-citation pay? *Scientometrics*. 2007;72(3):427-437.

4. Szomszor M, Pendlebury DA, Adams J. How much is too much? The difference between research influence and self-citation excess. *Scientometrics*. 2020;123(2):1119-1147. doi:10.1007/s11192-020-03417-5

5. Ioannidis JPA. A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *J Psychosom Res*. 2015;78(1):7-11. doi:10.1016/j.jpsychores.2014.11.008

6. Crandall C. Letter to APS on PoPS. *PsyArXiv*. Published online April 2, 2018. doi:10.31234/osf.io/w2exa

7. Wilhite AW, Fong EA. Scientific publications. Coercive citation in academic publishing. *Science*. 2012;335(6068):542-543. doi:10.1126/science.1212540

8. The PLoS Medicine Editors. The impact factor game. It is time to find a better way to assess the scientific literature. *PLoS Med*. 2006;3(6):e291. doi:10.1371/journal.pmed.0030291

9. Hendrix D. Institutional self-citation rates: A three year study of universities in the United States. *Scientometrics*. 2009;81(2):321-331. doi:10.1007/s11192-008-2160-2

10. Ioannidis JPA, Baas J, Klavans R, Boyack KW. A standardized citation metrics author database annotated for scientific field. *PLoS Biol*. 2019;17(8):e3000384. doi:10.1371/journal.pbio.3000384

11. King MM, Bergstrom CT, Correll SJ, Jacquet J, West JD. Men set their own cites high: Gender and self-citation across fields and over time. *Socius*. 2017;3:1-22. doi:10.1177/2378023117738903

12. Azoulay P, Lynn FB. Self-citation, cumulative advantage, and gender inequality in science. *Sociol Sci*. 2020;7:152-186. doi:10.15195/v7.a7

13. Mishra S, Fegley BD, Diesner J, Torvik VI. Self-citation is the hallmark of productive authors, of any gender. Schiller NO, ed. *PLoS ONE*. 2018;13(9):e0195773. doi:10.1371/journal.pone.0195773

14. Baas J, Ioannidis J, Klavans R, Boyack K. Data for aggregate statistics in "Hundreds of extreme self-citing scientists revealed in new database." *Mendeley Data*. Published online August 21, 2019. doi:10.17632/gw684hwcyb.1

15. Transparency International. *Transparency International Corruption Perceptions Index 2007: A Short Methodological Note*. Transparency International; 2007. Accessed June 2, 2021. https://images.transparencycdn.org/images/2007_CPI_ShortMethodology_EN.pdf

16. The World Bank. *World Development Indicators*. The World Bank; 2021. Accessed June 2, 2021. https://datacatalog.worldbank.org/dataset/world-development-indicators

17. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: Introducing the E-value. *Ann Intern Med*. 2017;167(4):268. doi:10.7326/m16-2607

18. Oster E. Unobservable selection and coefficient stability: theory and evidence. *J Bus Econ Stat*. 2019;37(2):187-204.

19. Seeber M, Cattaneo M, Meoli M, Malighetti P. Self-citations as strategic response to the use of metrics for career decisions. *Res Policy*. 2019;48(2):478-491. doi:10.1016/j.respol.2017.12.004

20. Baccini A, De Nicolao G, Petrovich E. Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. Bornmann L, ed. *PLoS ONE*. 2019;14(9):e0221212. doi:10.1371/journal.pone.0221212

21. Fisman R, Miguel E. Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *J Polit Econ*. 2007;115(6):1020-1048. doi:10.1086/527495

22. Morgenstern H. Ecologic studies in epidemiology: Concepts, principles, and methods. *Annu Rev Public Health*. 1995;16(1):61-81. doi:10.1146/annurev.pu.16.050195.000425

23. Singh VK, Singh P, Karmakar M, Leta J, Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*. 2021;126(6):5113-5142. doi:10.1007/s11192-021-03948-5