Cite the Good Cite: Making Citations in Political Science More Transparent: Online Appendix

Jonathan Grossman, Hebrew University of Jerusalem

Jonathan.grossman@mail.huji.ac.il

Contents

List of replication files

Each of the following files was assembled for this research. They are provided for replication purposes. All the files can be accessed at https://github.com/jonathan-grossman/Cite-the-Good-Cite.

2019_Articles_Data.csv – The database of political science articles from 2019 that was used in the section "The State of the Discipline: General Citations as Political Science Convention" in the main article.

RegEx_Formulas.csv – The regular expression patterns used to extract citations, citations of several works by the same author(s), detailed citations, and superfluous words in the list of references in every article.

Replication_Data.dta – The 2019 articles database in a STATA format (STATA 16, Windows 10).

Replication_Code.do – The syntax of statistical analysis in the article, using STATA 16 (Windows 10).

AJPS_Citations.xlsx – An Excel spreadsheet containing the raw data of all the articles published in the *American Journal of Political Science*, which were coded into the database.

APSR_Citations.xlsx – An Excel spreadsheet containing the raw data of all the articles published in the *American Political Science Review* that were coded into the database.

BJPS_Citations.xlsx – An Excel spreadsheet containing the raw data of all the articles published in the *British Journal of Political Science* that were coded into the database.

IO_Citations.xlsx – An Excel spreadsheet containing the raw data of all the articles published in *International Organization* that were coded into the database.

WP_Citations.xlsx – An Excel spreadsheet containing the raw data of all the articles published in *World Politics* that were coded into the database.

How the journals in the database were selected

Work on this research started in January 2020 (although it was soon halted for several months because of the COVID-19 pandemic outbreak). My goal was to obtain the most recent dataset that represented a full calendar year. For this reason, I reviewed articles published in 2019.

In selecting a sample of journals, I considered the top eight journals in four rankings: Clarivate's Journal Citation Reports for 2018 (Political Science and International Relations); Scimago Journal & Country Rank for 2018 (Political Science and International Relations); Google Scholar's Political Science ranking; and the Teaching, Research, and International Policy Project (TRIP) 2017 survey. Seven journals appeared more than once in these rankings: *American Journal of Political Science*, *American Political Science Review*, *International Organization*, and *International Security* occurred three times, while *British Journal of Political Science*, *Foreign Affairs*, and *World Politics* occurred twice.

I excluded *International Security*, which does not have author-date format citations, and *Foreign Affairs*, which does not have citations. This wide variety of citation practices represents one of the challenges of working toward greater precision in political science. The sample thus consists of five journals: *American Journal of Political Science*, *American Political Science Review*, *British Journal of Political Science*, *International Organization*, and *World Politics*.

The database was created using *Web of Science* bibliometric data. On 21 January 2020, I searched the *Web of Science Social Sciences Citation Index* for articles published in the five journals during 2019 (including "early access" articles). Out of the 267 results, I excluded articles labeled "Editorial Material" and "Correction," and was left with a database of 256 articles. I exported these articles' bibliographic details from the *Web of Science* into a Google Sheets spreadsheet. I then individually reviewed all the articles in the database and coded their attributes, as I explain in the next section and in the codebook.

How general and detailed citations were extracted

To aggregate both general and detailed citations, I employed a semi-automated approach: I used regular expressions to capture all the possible citations in every individual article, and then I manually browsed through the captured citations and removed those that were not actually citations. To make sure that the automated process did not fail to capture actual citations, I also manually skimmed through the content of each article and closely reviewed all the endnotes.

Regular expressions (often referred to as RegEx) are common patterns in a text as well as the language that allows us to find such patterns.[1] For example, author-date citations usually include at least one word (the author's name) followed by a four-digit number (the year of publication). A basic RegEx to capture all such citations would be **\w+ \d{4}**. This pattern, which refers to a combination of any number of letters or digits

[1] Friedl, Jeffrey E. F. 2006. *Mastering Regular Expressions*. 3rd ed. Sebastapol, CA: O'Reilly, 4.

followed by four consecutive digits, would capture citations such as "**Tarrow 1996**" or "Gerring and **Christenson 2017**" (in the latter case, the captured string would be "Christenson 2017"). However, it will also capture other combinations of words and digits, such as "**Trump 2020**" or "**November 2020**" in a description of the 2020 US election. What is more, it will fail to capture citations such as "King et al. 1994" (because of the dot after the word "al," which is neither a letter nor a digit) or "Tarrow (1996)" (because of the brackets that separate the author's name from the date). Therefore, more complex patterns, and in a few cases, a totally manual process, were needed to ensure accuracy in the extraction and classification of citations.

I used a slightly different pattern for each journal based on the peculiar characteristics of its citations and other features. Some of these formulas are not very elegant, and sometimes I had to use "brute force" – including all possible patterns to extract all possible citations. What mattered to me was not the elegance of my formulas but their efficiency in extracting

all the required citations. As my manual verification checks indicated, they have achieved this goal.

To capture citations and other patterns, I used the HTML version of the articles (an early attempt to use RegEx in PDF documents proved to be more complicated and less accurate). I used the *{find+}* browser extension for Google Chrome for this purpose. Having experimented with a number of RegEx tools, *{find+}* proved to be the most efficient and reliable, as it was able to capture all of the in-text citations in all of the journals and all of the citations in the notes, except for references that appeared in the notes of articles published in the *American Journal of Political Science*, which I had to extract manually.

Before applying each pattern to all the articles in a journal, I did a pilot extraction on three or four articles (depending on the modifications that I had to introduce into the RegEx during this process). At this pilot stage, I carefully combed through the entire article for each citation and made sure that the RegEx captured all of them. Only after I was certain that the

pattern worked did I apply it onto all the articles in the journal. In cases where I had to modify the pattern after the pilot (that is, when my manual revision detected detailed citations that the pattern had failed to capture previously), I retroactively applied the new RegEx onto all the previous articles. Nonetheless, as this was, to a large extent, a manual process, human errors could have occurred despite these control measures. Needless to say, any errors are my own and are unrelated to the extraction tool that I used.

The extraction and coding procedure

The following procedure was applied individually to the HTML version of every article in the database.

1. All words in the article's main text and notes were counted by copying and pasting them into a blank Microsoft Word document and using the word processor's Word Count function (following the

methodology of Gerring and Cojocaru[2]). In accordance with the journals' word limit rules, I counted the words in the epigraph (if there was any) and all parts of the article and notes, including headings, tables, and appendices that were printed with the article but excluding abstracts, online appendices, text in figures, acknowledgments, author biographies and other supporting information. I also excluded the number of words in the list of references, which I counted separately.

2. The footnotes or endnotes were counted so that their number, doubled, could be extracted from the calculation. For example, if there were 26 notes, I discounted 52 words from the final calculation of words in the article. That is because, in the HTML version, both the note's number (e.g., the figure 1 before Note Number 1) and the reference to the note in the main text (e.g. the

[2] Gerring, John, and Lee Cojocaru. 2016. "Arbitrary Limits To Scholarly Speech: Why (Short) Word Limits Should Be Abolished." *Qualitative and Multi-Method Research*, 5. https://zenodo.org/record/823308.

superscripted figure 1 in the sentence “All is good.[1]”) are counted as distinct words.

3. The words in the article’s list of references were counted by copying and pasting them into a blank Word document.
4. A RegEx pattern was used to capture all the superfluous words in the list of references – links to different academic indexes in which the cited articles could be accessed, such as *Web of Science* or *Google Scholar*. These words do not appear in the printed version of the article and are not included in its word count. The extracted text was pasted into a Word document and the number of words in this document was subtracted from the total number of words in the references.
5. All of the citations in the article and the notes, both general and detailed, were extracted using a RegEx pattern (with the exception of citations that appeared in the notes section of articles published in the *American Journal of Political Science*, which I extracted manually). The captured citations were then pasted into two

different columns in a designated spreadsheet. The first column was left intact for replication.

6. Each article was manually reviewed to ensure that no citation was overlooked. If additional citations were found, they were copied and pasted into the spreadsheet.
7. All of the results were reviewed manually, such that any non-citations accidentally included were deleted. If needed, I went back to the original article to verify whether a match was a citation or not (for instance, references to the author Luke March could have been mistaken for dates, and hence created false positives). In this stage, I retained only citations of secondary sources (that is, sources that appeared in the article's list of references). Other citations, such as those of primary sources not in the list of references, newspaper articles, internet sources (unless they appeared in the list of references), and court decisions, were excluded. Citations of "forthcoming" papers, which did not have page numbers, were equally removed.

8. I used a different pattern to capture all of the references to several works by the same author (for example, “Adamson 2006/2018” or “Heimann 2010, 2015”), which the first pattern was unable to find. The citations were pasted into a designated spreadsheet column, from which false positives were removed. The number of values left in this column was added to the number of values in the previous column to determine the total number of references in the article (the **CIT_NO** variable in the database).
9. All of the detailed citations in the article and notes were captured using a RegEx pattern (with the exception of the notes in the *American Journal of Political Science*, from which citations were extracted manually). I also made sure that no detailed citation was left out by manually reviewing the notes section in every article. The captured citations were pasted into two different columns in the spreadsheet – once again, the first column was left intact for replication.

10. All potential detailed citations were manually reviewed and false positives were deleted. The number of values left in this column was coded as the number of detailed citations in the article (the **DET_CIT_NO** variable in the database).

11. If, during the phase of capturing detailed citations, I noticed a type of detailed citation that the existing RegEx pattern could not capture, I changed the formula to include this type of citation and then reapplied the RegEx search to all articles that had been coded previously to ensure that I did not miss any detailed citation. There were, however, some exceptions, mostly because of citation style irregularities in some articles, as I detail in the next section of this appendix.

Journal-specific notes

American Journal of Political Science

- AJPS uses a slash (“/”) to separate works by the same author, instead of a comma. The RegEx pattern for capturing such works was modified accordingly.
- In Richard A. Nielsen’s article “Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers,” the RegEx pattern that I used for this journal did not work, for a reason that I was not able to ascertain. Therefore, I extracted all the citations from this article manually.
- Jeffrey Church’s article “Exemplary Lives and the Normative Theory of Culture” had an appended list of abbreviations, which I did not include in the article’s word count.

British Journal of Political Science

- Several articles in this journal contained references to secondary sources that did not appear in the list of references. For example, in Niheer Dasandi and Lior Erez's article "The Donor's Dilemma: International Aid and Human Rights Violations," there were references to "Rai 1980" and "Wang 1999," which were omitted from the list of references. I decided to count such citations nonetheless, as their exclusion was most likely made by mistake.
- In Alan Lucardi's article "The Effect of District Magnitude on Electoral Outcomes: Evidence from Two Natural Experiments in Argentina," the RegEx for capturing citations of several works by the same author(s) yielded too many false positives to be useful. Therefore, I extracted such citations manually.
- In Jennifer Hadden and Lorien Jasny's article "The Power of Peers: How Transnational Advocacy Networks Shape NGO Strategies on Climate Change," there is a reference to "DiMaggio and Powell 1983, 1952." This does not appear to be a citation of two works by

the same authors (their 1983 article is the only one in the references) nor a page number (DiMaggio and Powell's 1983 article has the page range of 147–160). Therefore, I counted this reference as a single general citation.

- In Joseph Asunka, Sarah Brierley, Miriam Golden, Eric Kramon, and George Ofosu's article "Electoral Fraud or Violence: The Effect of Observers on Party Manipulation Strategies," the citation style is different from the regular *BJPS* style – page numbers are indicated by the abbreviations "p." or "pp." As a result, the RegEx pattern could not capture them. For this reason, I extracted detailed citations manually.
- In Simon Bornschier's article "Historical Polarization and Representation in South American Party Systems, 1900–1990," there were some inconsistencies regarding detailed citations of chapter numbers, which were cited in different styles. Therefore, I extracted these citations manually.

- In Mahvish Shami's article "Connectivity, Clientelism and Public Provision," I counted the words in the "List of Interviews" as part of the list of references.

International Organization

- In Phillip Y. Lipscy and Haillie Na-Kyung Lee's article "The IMF As a Biased Global Insurance Mechanism: Asymmetrical Moral Hazard, Reserve Accumulation, and Financial Crises," there was a reference to "Reinhart and Rogoff 2009 (chapter 1)," which the RegEx pattern could not capture. Therefore, I extracted it manually (as a detailed citation).
- In Dan Honig's article "When Reporting Undermines Performance: The Costs of Politically Constrained Organizational Autonomy in Foreign Aid Implementation," there was an "Ibid., abstract" citation that the RegEx could not count. I extracted it manually as a detailed citation.

- In Jack Paine's article "Ethnic Violence in Africa: Destructive Legacies of Pre-Colonial States," I manually added the reference "Coded by the author using Harkness's 2018 appendix" as a detailed citation.
- Benjamin O. Fordham's article "The Domestic Politics of World Power: Explaining Debates over the United States Battleship Fleet, 1890–91" included detailed citations with square brackets (e.g. "Sprout and Sprout 1966 [1939], 205–17"), which the Regex pattern did not capture. Therefore, I added them manually.

World Politics

- Jack Paine's article "Democratic Contradictions in European Settler Colonies" included the citation of tables in another work ("Source: Land data from Lützelschwab 2013, Tables 5.1 and 5.2"). I extracted these detailed citations manually.
- Margarita H. Petrova's article "Naming and Praising in Humanitarian Norm Development" had references to footnotes

without page numbers (e.g. " Finnemore and Sikkink 1998, fn. 57"). It also used "pp." when referring to a specific page range. I added these citations manually, and I retroactively checked all the other *World Politics* entries manually (by searching the articles for the strings "p." and "pp.") to make sure that I did not fail to capture such citations, which did not conform to the journal's style. Apparently, this style was unique to Petrova's article, except for one reference in David Hope and Angelo Martelli's article "The Transition to the Knowledge Economy, Labor Market Institutions, and Income Inequality in Advanced Democracies" ("As Arestis and Paliginis 1995, p. 267, neatly summarize"). I retroactively added this citation manually.

- Darin Christensen, Mai Nguyen, and Renard Sexton's article "Strategic Violence during Democratization: Evidence from Myanmar" included references to sections (e.g. "See Christensen, Nguyen, Sexton 2019b, sec. A.3"), figures (e.g. "Christensen, Nguyen, and Sexton 2019b, Figure A.1."), and tables (e.g.

"Christensen, Nguyen, Sexton 2019b, Table A1"), all of them in previous works by the article's authors. The RegEx pattern failed to capture these detailed citations, which I added manually.

- Magnus Feldmann's article "Global Varieties of Capitalism" included many detailed citations in the main article rather than in the footnotes (as in other *World Politics* articles). For the sake of precision, I added all of these citations manually.

The Regular Expressions used to capture the citations

For the RegEx formulas that I used to capture each type of citation in each journal, please see the **RegEx_Formulas.csv** file in this article's replication materials.

The “pilot” articles used for testing the regular expressions patterns

American Journal of Political Science

1. Hun Chung, “The Well-Ordered Society under Crisis: A Formal Analysis of Public Reason vs. Convergence Discourse”

2. Christian Gläßel and Katrin Paula, “Sometimes Less Is More: Censorship, News Falsification, and Disapproval in 1989 East Germany”

3. Lilla V. Orr and Gregory A. Huber, “The Policy Basis of Measured Partisan Animosity in the United States”

4. Patrick J. Egan, “Identity as Dependent Variable: How Americans Shift Their Identities to Align with Their Politics”

American Political Science Review

1. Ramya Parthasarathy, Vijayendra Rao, and Nethra Palaniswamy, "Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies"

2. Robert A. Blair, Sabrina M. Karim, and Benjamin S. Morse, "Establishing the Rule of Law in Weak and War-torn States: Evidence from a Field Experiment with the Liberian National Police"

3. Andrew B. Hall, Connor Huff, and Shiro Kuriwaki, "Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War"

British Journal of Political Science

1. Nicholas A. Valentino, Stuart N. Soroka, Shanto Iyengar, Toril Aalberg, Raymond Duch, Marta Fraile, Kyu S. Hahn, Kasper M. Hansen, Allison

Harell, Marc Helbling, Simon D. Jackman, and Tetsuro Kobayashi, "Economic and Cultural Drivers of Immigrant Support Worldwide"

2. Mahvish Shami, "Connectivity, Clientelism and Public Provision"

3. James R. Hollyer, B. Peter Rosendorff and James Raymond Vreeland, "Transparency, Protest, and Democratic Stability"

International Organization

1. Nikhil Kalyanpur and Abraham L. Newman, "Mobilizing Market Power: Jurisdictional Expansion as Economic Statecraft"

2. Phillip Y. Lipscy and Haillie Na-Kyung Lee, "The IMF As a Biased Global Insurance Mechanism: Asymmetrical Moral Hazard, Reserve Accumulation, and Financial Crises"

3. Darin Christensen, "Concession Stands: How Mining Investments Incite Protest in Africa"

World Politics

1. Milena Ang and Monika Nalepa, "Can Transitional Justice Improve the Quality of Representation in New Democracies?"

2. Egor Lazarev, "Laws in Conflict: Legacies of War, Gender, and Legal Pluralism in Chechnya"

3. Timothy Frye, Ora John Reuter, and David Szakonyi, "Vote Brokers, Clientelist Appeals, and Voter Turnout: Evidence from Russia and Venezuela"

4. Sarah Zukerman Daly, "Voting for Victors: Why Violent Actors Win Postwar Elections"

Codebook: The Variables in the Database

Note: *for the database itself please see the* ***2019_Articles_Data.csv*** *file in this article's replication materials*.

AUTHORS: The name(s) of the author(s) of the article.

TITLE: The title of the article.

JOURNAL: The title of the journal in which the article was published.

JOURNAL_CODE: The journal in which the article was published.

1 = *American Journal of Political Science*

2 = *American Political Science Review*

3 = *British Journal of Political Science*

4 = *International Organization*

5 = *World Politics*

DOI: The article's digital object identifier.

URL: The article's web address.

A_1_COUNTRY: The country in which the first author's institution is located according to the information in the article.

Notes: If the first author was affiliated with several institutions, the first institution on the list was coded. The journal *International Organization* does not contain information on authors' affiliation. Therefore, I used Google to look up this information. I did the same for Matto Mildenberger and Dustin Tingley's article "Beliefs about Climate Beliefs: The Importance of Second-Order Opinions for Climate Politics" in the *British Journal of Political Science*, in which I could not find this information.

A_1_US: This variable indicates whether the article's first author was affiliated with an institution within or outside the United States at the time of publication.

0 = The first author is affiliated with a non-US institution

1 = The first author is affiliated with a US institution (including New York University Abu Dhabi)

A_2_COUNTRY: The country where the institution of the second author (if there is any) is located.

A_3_COUNTRY: The country or countries where the institution(s) of any additional authors, if there are any, are located.

SHORT_ARTICLE: This variable indicates the type of short articles that usually have a stricter word limit than full-length articles. Different journals may use different terms for this type of article, such as “note,” “letter,” or “reflection.”

1 = The article is a short article.

2 = *British Journal of Political Science* articles entitled “Notes and Comments.” While articles in this category are shorter than most full-length articles, I was unable to find any word limit specifications regarding this category in the journal’s instructions for contributors. Therefore, I applied the journal’s regular word limit to these articles.

WORD_NO: The number of words in the body of the article, beginning in the epigraph or introduction and ending in the notes or the printed

appendix (if there is any). The list of references and online materials are excluded from this count.

NOTES_NO: The number of footnotes or endnotes in the article.

REF_WORD_NO: The number of words in the list of references, excluding such strings as “Google Scholar” and “Crossref” that refer readers to a digital version of the cited work and do not appear in the article’s print version.

WORDS_CALC: The total number of words in the article, including the body of the article, the notes, the list of references, and any offline appendix.

K_WORDS_TOTAL: The total number of words in the article, divided by 1,000. This count includes words in the body of the article, the notes, the list of references, and any offline appendix.

REF_WEIGHT: The number of words in the list of references (“**REF_WORD_NO**”) divided by the number of words in the entire article

(“**WORDS_CALC**”). In other words, this variable indicates how many references per word there are in the article.

CIT_NO: The number of citations in the article.

DET_CIT_NO: The number of detailed citations in the article, that is, citations that refer the reader to a specific location within the cited secondary source (e.g. pages, chapters, sections, appendices, tables, figures, notes, etc.).

DET_CIT_WEIGHT: The number of detailed citations (“**DET_CIT_NO”**) divided by the number of citations in the article (“**CIT_NO**”). This variable indicates how many citations in the article are detailed out of all the citations. It is the dependent variable in the linear regression model presented in the article.

WORD_LIMIT: The official word limit for the article as stipulated in the journal's instructions for contributors.

LENGTH_RATIO: The total number of words in the article (“**WORDS_CALC**”) divided by the journal’s word limit for this type of

article (“**WORD_LIMIT**”). This variable shows whether an article exceeds the official word limits imposed by the journal and to what extent it goes beyond these limits.

CIT_PER_WORD: The number of citations in the article (“**CIT_NO**”) divided by the number of words (“**WORDS_CALC**”). This variable indicates how many citations-per-word there are in the article.