Codebook and supporting information for

Can We Do Better? Replication and Online Appendices in Political Science

by Jonathan Grossman and Ami Pedahzur

Forthcoming in *Perspectives on Politics*

Table of Contents

The Article Database

Our full database can be accessed and downloaded in its raw form here. We created it using *Airtable*, an online spreadsheet and database tool that is particularly intuitive and user-friendly. In this part of the appendix, we detail our data collection process and present some relevant descriptive statistics. As an *Airtable* file, users may utilize the software's other options for sorting, grouping, and filtering the data in the spreadsheet. However, in case users do not have access to Airtable or wish to use another software, the database can also be downloaded as a CSV file here.

How We Created the Article Database

In this section, we will briefly describe how we examined the state of transparency and data sharing in current political science articles.

First, we read Ellen M. Key's article "How Are We Doing? Data Access and Replication in Political Science" (2016).This article maps the transparency and replicability standards in six leading political science journals – *American Political Science Review* (APSR), *American Journal of Political Science* (AJPS), *British Journal of Political Science* (BJPS), *International Organization* (IO), *Journal of Politics* (JOP), and *Political Analysis* (PA). Assuming that the norms prevalent in such prominent journals represent the discipline's highest standards, and that these norms will eventually permeate other publications, we decided to use Key's six journals as the baseline for our examination of data sharing in political science articles.

We began the data collection in November 2019. To ensure systematic categorization, we reviewed the most recent issues of each one of the six journals (that is, not "first view" articles, but ones that have already been assigned specific page numbers in a particular issue). This means that many of these articles were submitted in 2018 and even 2017.

In each issue, we surveyed all the articles, research notes, letters, and comments. We did not include editorials, addenda, corrigenda, review essays, and responses (for example, we examined a "Comment" in the BJPS that included an analysis and a dataset, but not the original article authors' response to this comment, as the data used in this response are the same data used in the original article).

We opened each article's web page individually and manually located its supplementary materials or any links to a data archive. In most journals, there were either direct hyperlinks from the article's page to the replication data archive or the data was stored on the article's page itself. In *The Journal of Politics*, however, the main text of the article only contains a general link to the journal's *Dataverse* rather than a direct hyperlink to the article's data archive. Thus, we had to manually search for the relevant article in the JOP's *Dataverse*. In some instances, searching for an article's appendix by typing the first author's last name did not yield the article's data repository. Only further searches (for example, by the second author's name or by words from the article's title) led to the archive. In three cases – Sellars 2019, Meijers and van der Veer 2019, and Gueorguiev and Malesky 2019 – such a search did not yield the purportedly available repository of replication materials. In *Political Analysis*, the text appendices on the

article's page as well as the data and code files on the journal's *Dataverse* could only be downloaded as compressed ZIP files. In such cases, we downloaded the compressed files and opened them to view the text documents and data repositories.

We coded the variables with the spreadsheet/database online tool *Airtable*, which has advanced and intuitive sorting and filtering tools and offers a good free version, which make it an excellent tool for descriptive statistics. We present our codebook in the next section.

Codebook

“Authors”: The names of all the article’s authors as they appear in the article’s page.

“Article Title”: The title of the article.

“Journal”: The name of the journal in which the article is published.

“Volume”: The number of the volume in which the article is published.

“Issue”: The number of the issue in which the article is published.

“Year”: The year in which the issue containing the article was published (*not* the year in which the article was published online).

“Pages”: The article’s page numbers in the relevant volume.

“Link”: A clickable URL of the article’s page on the journal's website.

“Has Appendix?”: Does the article have an online or printed appendix?

“Yes” = the article includes supplementary information of any kind, including appendices that are attached to the text of the main (printed) article.

“No” = we could not locate any supplementary materials for the article.

“Appendix Includes a Text Document?”: Does the article’s online version offer the option to download at least one text file?

“Yes” = the appendix contains at least one text document.

“No” = no such document is available for download.

“Unclear” = we could neither verify or rule out the existence of such a document (this value does not occur in our data, but we included it for future expansions of the dataset).

“In the Article” = the main text of the article includes an appendix.

“Number of Pages of Main Document?”: The number of pages in the appendix’s text document, if the appendix includes such a document. If there are more than one text files, this value represents the number of pages in the longest document.

“**Appendix Includes Table of Contents?**”: Does the appendix contain a table of contents or a detailed list that allows users to navigate it and find the exact location of the information they seek?

“Yes” = the appendix contains a detailed table of contents.

“No” = the appendix does not contain a table of contents.

Note: We coded this variable “No” if the text document had a list describing the different parts or sections of the appendix but this list did not point to specific page numbers or contained hyperlinks to these parts.

“Text Document Includes Page Numbers?”

“Yes” = the pages of the main text document(s) of the appendix are numbered, regardless of whether there is a table of contents.

“No” = the pages of the appendix are unnumbered.

“Appendix Includes Replication Data?”: Is there a replication dataset associated with the article? This dataset can be hosted on the journal’s website, the journal’s data archive on a separate website (such as *Dataverse*), or the author’s personal website.

“Yes” = the article offers raw or structured data for replication.

“No” = no such data are available.

“Broken Link” = The article contains a link or reference to a data archive but this link does not lead to the article’s actual data repository.

Note: we coded this variable “Broken Link” only if a manual search in the journal’s Dataverse (by first author’s surname, second author’s surname, or words from the article’s title) did not lead to any repository associated with the article.

“Replication Data Available on Journal's Website?”:

“Yes” = the journal stores the datasets used for analysis on its website or in a file storage system (e.g. *Dataverse*).

“No” = the article’s dataset is available on the author’s personal website or any other site not officially associated with the journal.

“Linked” = there is a hyperlink to the data in the journal’s website but the data themselves are stored elsewhere (including in a personal *Dataverse* that is not the journal’s official *Dataverse*).

"Broken Link" = the link provided in the article does not lead to the article's data repository, and a search of the journal's *Dataverse* (by first author's surname, second author's surname, or words from the article's title) did not lead to the article's repository.

"Link to Replication Materials Page": A clickable URL of the article's data archive.

"Replication Dataset Includes Instructions?": Does the replication materials repository include a text document of any sort (usually a README file or a codebook) that explains which files are included in the repository and how to use them for replication?

"Yes" = at least one text document is included in the replication data repository.

"No" = no text document is included in the data archive.

Note: When coding this variable, we were agnostic to the question of whether the main text document of the appendix contained such instructions or not. We only took into account the files offered in the data repositories themselves. As R. Michael Alvarez, Ellen M. Key, and Lucas Núñez expound in their 2018 article, "although codebooks may be developed internally (e.g., as part of a Stata .dta file), they will be lost when files are saved in a flat-file format. For this reason, we also recommend the inclusion of a separate codebook file." (page 426, note 11). The same caveat applies to codebooks or README instructions that were embedded in statistical software files (and could thus be opened only by users who owned the software in question) rather than published as independent text files.

"Replication Dataset Includes Data?": Does the article's data repository contain any file(s) that the authors define as "data"?

"Yes" = there is at least one data file in the repository.

"No" = no such file could be located.

"Unclear" = we could not determine whether the files in the repository were data files or not because the authors did not describe the nature of these files in their metadata and we did not have the software required for opening the files.

Note: When determining whether a replication data repository contained data and/or code, we considered the description of each file in the article's Dataverse or in the repository's README file. If we could not find such a description, we attempted to open the files in question. If there was no description and we were not able to open the files, we coded this variable as "Unclear." Whether these data were raw or structured did not affect our coding.

"Replication Dataset Includes Code?": Does the article's data repository contain any file(s) that the authors define as "code" or "syntax"?

"Yes" = there is at least one code or syntax file in the repository.

"No" = no such file could be located.

"Unclear" = we could not determine whether the files in the repository were code files or not because the authors did not describe the nature of these files in their metadata and we did not have the software required for opening the files.

Note: When determining whether a replication data repository contained data and/or code, we considered the description of each file in the article's Dataverse or in the repository's README file. If we could not find such a description, we attempted to open the files in question. If there was no description and we were not able to open the files, we coded this variable as "Unclear."

"Text Document Includes Link to Replication Data?": Does the article's main text document include a link to the article's replication data repository?

"Yes" = the main text document includes a link to the article's replication data repository.

"No" = we could not find such a link.

Note: To code the "Text Document Includes Link to Replication Data?" variable, we opened the main documents of every article whose appendix contained both a text document and a replication data repository (using the software Adobe Acrobat Pro 2017 for PDF files and Microsoft Word 2016 for DOC files). To locate such links, we searched the documents for the keywords "Dataverse" and "http" (all the replication data archives that we surveyed were stored in Dataverse). In one case where the text of the appendix was not machine-searchable (Hassell and Visalvanich 2019), we used Adobe Acrobat Pro's Optical Character Recognition (OCR) tool to make it digitally searchable. If no such hyperlink could be found, we coded the variable as "No" (regardless of links to any other databases that the text document might have included).

"Notes" = brief explanatory notes that we have written to ourselves while coding the database.

Some Descriptive Statistics of the Article Database

In the next pages are three tables that we created by downloading our dataset from *Airtable* as a CSV file, uploading it to *Google Sheets*, and converting text values to numbers (for example, “yes” to 1 and “no” to 0). We included them here because we refer to data from them in the main article.

Table 1: The Accessibility of Text Documents in Appendices

Journal	Number of Articles Reviewed	Number of Articles with an Appendix	Number of Appendices with a Main Text Document (Excluding Appendices Attached to the Main Article)	Number of Text Documents that Include a Table of Contents	Number of Text Documents that Include Page Numbers
American Journal of Political Science	15	14	14	10	12
American Political Science Review	14	13	12	4	12
British Journal of Political Science	17	16	15	1	13
International Organization	7	7	7	3	7
Political Analysis	13	13	10	4	10
The Journal of Politics	31	29	28	11	24
Grand Total	**97**	**92**	**86**	**33**	**78**

Note: In "Number of Appendices with a Main Text Document" we excluded appendices that were attached to the main article (N=5).

Table 2: The Accessibility of Data Archives in Appendices

Journal	Number of Articles Reviewed	Number of Articles with an Appendix	Number of Appendices that Include a Replication Data Archive	Number of Replication Data Archives that Include an Instructions File	Number of Replication Data Archives that Include Data Files (Excluding Unclear Cases)	Number of Replication Data Archives that Include Code/Syntax Files	Number of Appendix Text Documents that Include a Link to the Article's Replication Data Archive
American Journal of Political Science	15	14	12	12	12	12	0
American Political Science Review	14	13	10	5	10	10	1
British Journal of Political Science	17	16	16	5	16	14	2
International Organization	7	7	2	1	1	1	0
Political Analysis	13	13	13	13	13	13	3
The Journal of Politics	31	29	26	11	26	26	0
Grand Total	**97**	**92**	**79**	**47**	**78**	**76**	**6**

Note: When counting the "Number of Appendices that Include a Replication Data Archive," we excluded cases where there was a reference to a replication data archive but we could not retrieve this archive; when counting the "Number of Replication Data Archives that Include Code/Syntax Files," we had to exclude one article (Shami 2019) whose replication file we were not able to open and thus unable to determine whether it included any code in addition to the data.

Table 3: Other Descriptive Statistics

Query	Number of Articles	Percentage
Total number of articles in the database	97	
Articles that include an appendix	92	
Out of which, appendices that include a text document	86	93.48%
Out of which, documents that include a table of contents	33	38.37%
Appendices that include a replication dataset	79	
Out of which, datasets that include an instructions file	47	59.49%
datasets that include replication data	78	98.73%
datasets that include code/syntax	76	96.20%
Appendices that include both a text document and a replication dataset	74	
Out of which, the number of text documents that include a link or reference to the data archive	6	8.11%