On the Importance of 'Janitor Work' in Political Science: The Case of Thermostatic Support for Democracy*

Yue Hu   Yuehong 'Cassandra' Tai   Frederick Solt

Much recent political science research can be characterized as data science: it employs large quantities of data, often drawn from a large number of different sources. For such projects, data wrangling, the task of getting these data into the format required to perform analyses, is notoriously the bulk of the work (see, e.g., Lohr 2014). Such data 'janitor work' is often viewed as tiresome, as something to be delegated to research assistants, to someone—indeed anyone—else (see Torres 2017). Data wrangling is, however, critically important. Without wrangling our data, we cannot work with it, and without wrangling it *correctly*, we risk reaching mistaken conclusions. This is not, we think, particularly controversial.

We nevertheless seek to draw attention to a single, particularly insidious and perhaps common way in which data 'janitor work' can go wrong: data-entry errors. Faced with the task of getting data into the correct format, even some very sophisticated researchers will conclude that the most straightforward means to that end is to simply copy the needed data into a spreadsheet manually. Straightforward though this technique may be, it is very much prone to errors: Barchard and Pace (2011) found that 'research assistants' assigned in an experiment to carefully enter data manually, even those instructed to double-check their entries against the original, had error rates approaching 1% in just one roughly half-hour session. Rates likely go up as the tedious task goes on.

Here, we document data-entry errors in a prominent recent work, Claassen (2020a), that examines how changes in democracy affect democratic support among the public. After correcting these errors and re-running the article's models, we find that there is no empirical support for the paper's main conclusion that public support responds thermostatically to

*Corresponding author: yuehong-tai@uiowa.edu. Current version: May 28, 2022.

changes in democracy. There is no reason to think that, "should elected leaders start dismantling democratic institutions and rights, public mood is likely to swing rapidly toward democracy again, providing something of an obstacle to democratic backsliding," as Claassen (2020a, 51) concluded. Using these findings as a cautionary tale, then we offer suggestions to minimize data-entry errors and their impact.

Data-Entry Errors and Democratic Support

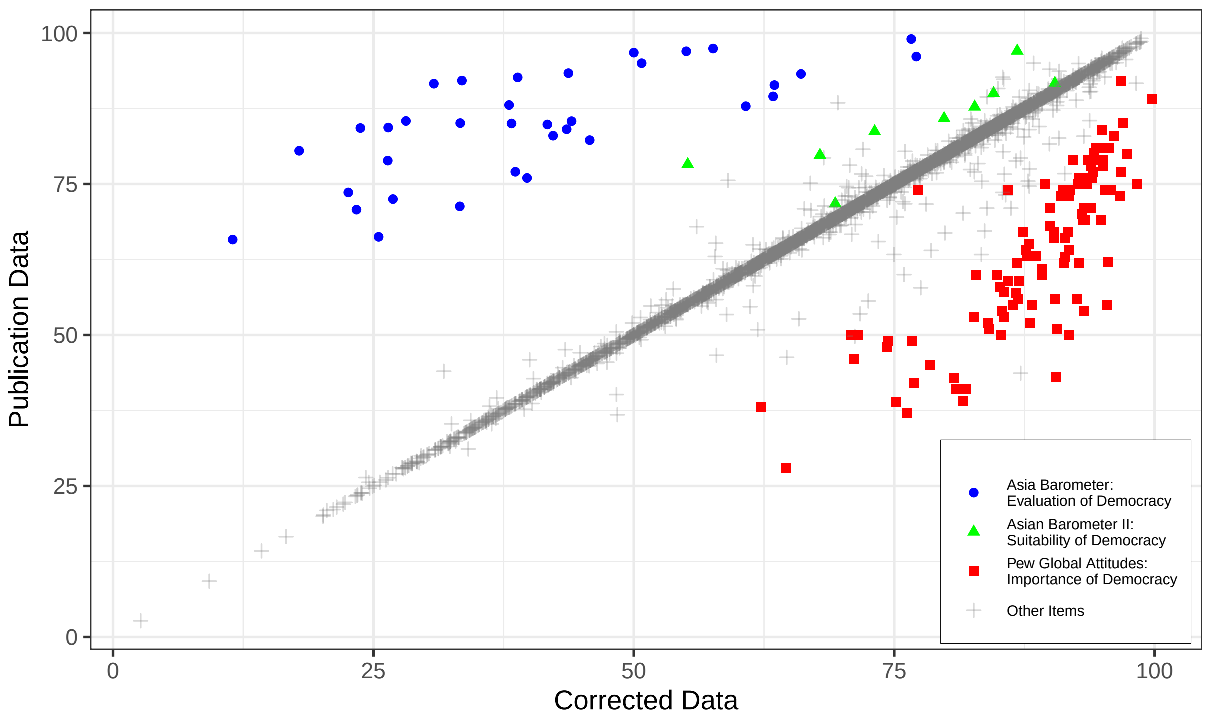
With democracy under increasing threat in countries around the world, how the public reacts is a crucial question. According to a now-classic literature, it is experience with democratic governance that generates robust public support for democracy (see, e.g., Lipset 1959). A prominent recent study, Claassen (2020a), argues instead that democratic support behaves thermostatically, that is, that increases in democracy yield an authoritarian backlash in the public, while democratic backsliding prompts the public to rally to democracy's cause. The evidence it offered in support of this argument takes advantage of recent advances in modeling public opinion as a latent variable to provide its dependent variable, estimates of democratic support for over one hundred countries for up to nearly three decades, constituting a much larger evidentiary base than any previous study.

These latent variable estimates of democratic support are based on source material consisting of thousands of nationally aggregated responses to dozens of different questions from cross-national survey projects (Claassen 2020a, 40). Two pieces of data were collected for each distinct survey item in each country and year it was asked: the number of respondents to give a democracy-supporting response—defined, for Likert scales and other ordinal responses, as those above the median value of the scale—and the total number of respondents to whom the question was posed (Claassen 2020a, Appendix 1.3). Each of these pieces of source data is recorded in a spreadsheet.[1]

We re-collected all of the source data for the publication from the original surveys using `DCPOtools`, an R package which automates the process of aggregating survey responses for

[1]The article's replication materials include only the latent variable estimates without the original survey aggregates that serve as their source data (see Claassen 2020c). Fortunately, however, the spreadsheet recording these original source data is included in the replication materials for a companion piece that employed the identical estimates (see Claassen 2020b).

use in estimating dynamic comparative public opinion as a latent variable (Solt, Hu, and Tai 2018). In Figure 1, we compare the percentage of respondents to give a democracy-supporting response in the publication spreadsheet with the percentage we found using our automated process of wrangling these same data. When points fall along the 45° diagonal, it indicates that the publication's source data and our own automated workflow reported the same percentages. Points above this diagonal represent observations for which the publication data overestimated the actual percentage of respondents who offered a democracy-supporting response, while points below this line are observations where the publication data underestimated this percentage.



Notes: Each point represents the percentage of respondents in a country−year to give a democracy−supporting response to a particular survey item. Publication data is as reported in Claassen (2020b); the corrected data was collected directly from the original surveys. The Asia Barometer's item on the evaluation of democracy accounts for most overreports, and the Pew Global Attitudes item on the importance of democracy accounts for most substantial underreports. In both cases, as well as the overreports of the suitability of democracy item in the second wave of the Asian Barometer, the issues can be easily explained by errors in transcribing the data. Deviations in other items result from inconsistent treatment of missing data and/or survey weights, reflecting in part differences in codebook reporting practices across surveys.

Figure 1: Comparing Democracy-Supporting Responses in the Publication Data and the Corrected Data

For 85% of the country-year-item observations, the difference between these percentages was negligible—less than half a percent—yielding points approximately along the 45° diagonal. But for the remaining observations, the difference was often substantial as a result of data-entry errors in the publication data. For example, the Asia Barometer asked respondents in 35 country-years to indicate whether they thought “a democratic political system” would be very good, fairly good, or bad for their country. According the study’s coding rules (see Claassen 2020a, Appendix 1.3), only answers above the median of the response categories should be considered as democracy supporting, yet in this case the lukewarm intermediate category was coded as supporting democracy as well.[2] This led to overestimations of the percentage of democracy-supporting responses ranging from 19 to 63 percentage points and averaging 44 points.

Similarly, the four waves of the Asian Barometer included the following item: “Here is a similar scale of 1 to 10 measuring the extent to which people think democracy is suitable for our country. If 1 means that democracy is completely unsuitable for [name of country] today and 10 means that it is completely suitable, where would you place our country today?” In accordance with the coding rules of the study, responses of 6 through 10 are considered democracy supporting, and that is how the first, third, and fourth waves of the survey are coded. For the second wave, however, 5 was erroneously also included among the democracy supporting-response. This data-entry error resulted in overestimates of as much as 23.1 percentage points in 9 country-years.

A third example comes from the Pew Global Attitudes surveys’ four-point item asking about the importance of living in a country with regular and fair contested elections: the question wording is “How important is it to you to live in a country where honest elections are held regularly with a choice of at least two political parties? Is it very important, somewhat important, not too important or not important at all?” In this case, rather than including respondents who gave both responses above the median—“very important” and “somewhat important”—only those respondents who answered “very important” were entered as supporting democracy. This error caused substantial underreporting of the extent of democratic support in 91 country-years.

[2]Although this might be interpreted as an exercise of researcher judgment as to what constitutes a democracy-supporting response rather than a data-entry error, examination of similar answers to similar questions shows that similarly lukewarm responses at and below the median response category (e.g., in the Arab Barometer, that democracy was “somewhat appropriate” for the country) were coded as not supporting democracy.

While these issues involve mistakes in recording the numerator of the percentage, the number of respondents who provided a democracy-supporting answer, entering the denominator, the total number of respondents asked a question, was also problematic on occasion. For example, when the Americas Barometer surveyed Canada in 2010, it included an item asking whether, when “democracy doesn’t work,” Canadians “need a strong leader who doesn’t have to be elected through voting.” It posed this question to only half of its sample. Those who were not asked the question, however, were included in the total number of respondents as if they had refused to answer. According to the study’s coding rules, refusing to answer is equivalent to answering in a fashion not supporting democracy, that is, in this case, agreeing that Canada needed a strong leader who need not bother with elections (see Claassen 2020a, Appendix 1.3). This rule might be a reasonable coding choice, but including in this category those who were never asked the question at all is clearly a mistake in data entry.

Another source of data-entry errors in this study involves survey weights. Weighting raw survey results to maximize the extent to which they are representative of the target population is important. Relying on the toplines reported in survey codebooks rather than the survey data itself evidently caused some mistakes in correctly entering the needed information here, as codebooks do not always take survey weights into account. These data-entry errors shifted the percentage of democracy-supporting responses in both directions, typically by relatively small amounts.

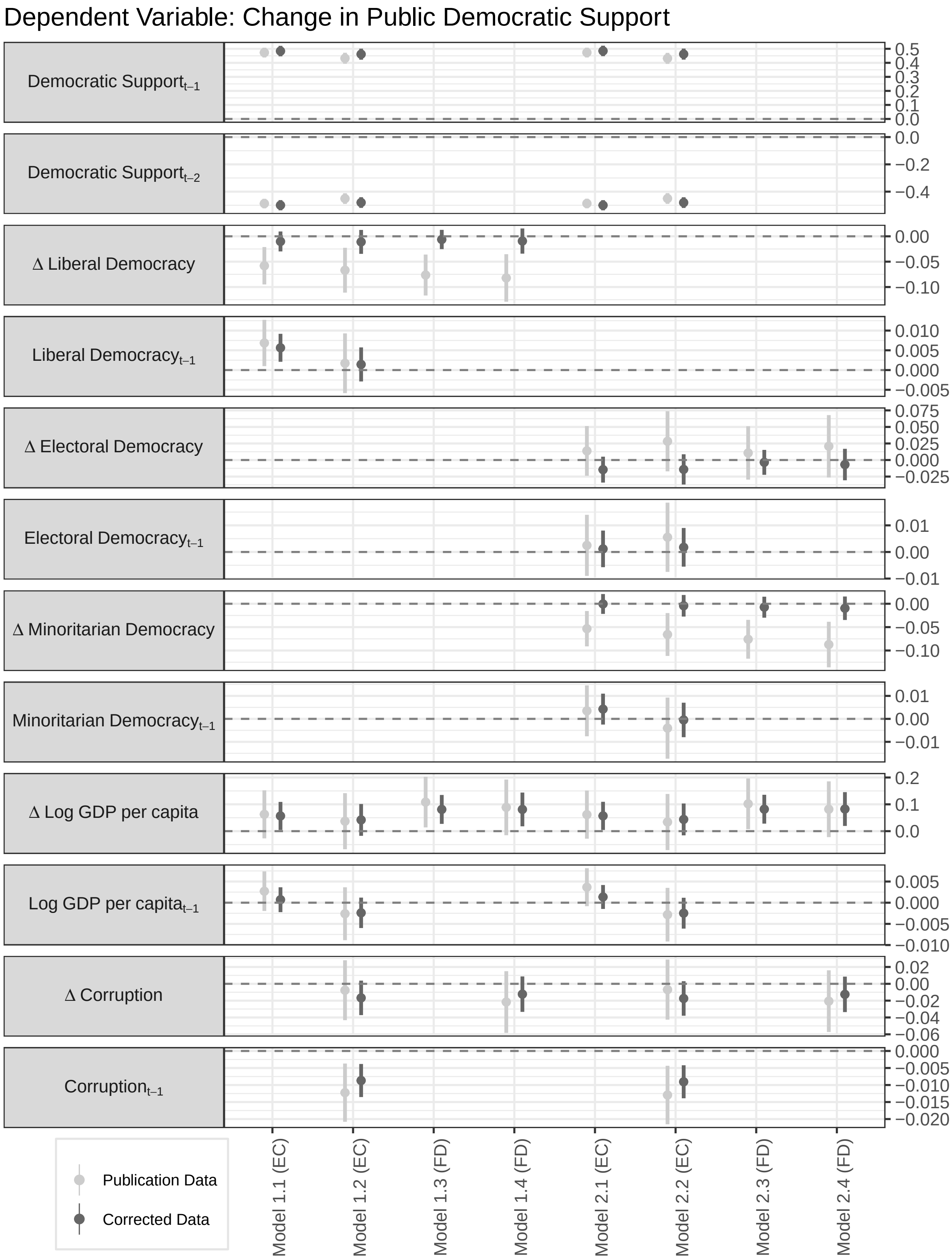
Consequences for Inference

Data-entry errors of this sort can yield erroneous conclusions. After replicating the latent variable of democratic support with first the article’s original data and then with the corrections to the errors we describe above, we replicated each of the models presented in Claassen (2020a) exactly using both versions of the latent variable. The results provide only limited support for the classic argument that democracy generates its own demand, at least in the short run, and none at all for a thermostatic relationship.

Figure 2 presents our results in a “small multiple” plot (Solt and Hu 2015) for a clear comparison of the coefficients of each variable in the article’s models. In the plot, the dots represent point estimates and the whiskers show the associated 95% confidence intervals. Each row depicts a variable's performance in its own scale across all of the models. The lighter dots and whiskers replicate those reported in the published article; the darker ones are the estimates obtained when using the corrected data.

Consider first Models 1.1 through 1.4, which replicate those presented in Table 1 of Claassen (2020a, 47). These models examine the effects of overall liberal democracy using error-correction models (labeled EC in Figure 2) and first-difference models (labeled FD). As Claassen (2020a, 46) notes, the thermostatic theory predicts that the estimated coefficient of the change in liberal democracy will be negative, while the classic theory suggests that lagged levels of liberal democracy will be positive. When using the original publication data with their data-entry errors, we replicate the results of the article exactly: the coefficients estimated for the change in liberal democracy are large, negative, and statistically significant across all four models, just as predicted by the thermostatic theory. The positive and statistically significant result for the lagged level of liberal democracy found in Model 1.1—supporting the classic theory—disappears when corruption is taken into account in Model 1.2, suggesting that "this effect is not particularly robust" (Claassen 2020a, 47).

When the data-entry errors are corrected, however, the results for these models suggest a very different set of conclusions. The standard errors shrink across the board—indicating that the models are better estimated in the corrected data—but so do the magnitudes of the coefficients. The positive and statistically significant result for the lagged level of liberal democracy remains in Model 1.1. The estimate is only slightly smaller than in the publication data, and as with the publication data, it disappears when corruption is added in Model 1.2: the evidence, such as it is, for the classic theory, operationalized as a short-run process, remains substantively unchanged. On the other hand, the estimates for the change in liberal democracy that provided support for the thermostatic theory are much smaller—very nearly exactly zero—and fail to reach statistical significance in any of these four models. Models 2.1 through 2.4, which break liberal democracy into its electoral democracy and minoritarian democracy components, similarly undermine claims for the thermostatic theory. The strong and statistically significant negative coefficients for the change in minoritarian democracy on public democratic support that are found using the publication dataset evaporate when



Notes: Replications of Claassen (2020), Table 1, 47, and Table 2, 49. Models denoted 'EC' are error−correction models; those marked 'FD' are first−difference models.

Figure 2: The Effect of Democracy on Change in Public Support

the data-entry errors are corrected. There is no support for the thermostatic theory.

This is not, we contend, a particularly surprising finding. As much as those who favor democracy might wish it were so, and as well as the thermostatic theory performs with regard to many other topics in public opinion, it is not a particularly likely candidate for explaining trends in democratic support—the mechanism required for it to operate is not present. In its original formulation, the theory requires citizens to possess a level of knowledge of politics that a long line of public opinion research shows is unrealistic; and as recently re-elaborated it requires the issue in question to be debated by political parties so as to provide cues to the broader public as to what is going on (Atkinson et al. 2021, 5–6). But virtually no party engaged in eroding democracy actually puts their actions in such terms: instead they claim to be defending democracy, or saving democracy, or putting forth a different model of democracy that better suits the nation's needs. And to the extent they succeed, their opponents are increasingly unable to make their case to the public at all. Absent its mechanism, the thermostat cannot operate on the public's democratic support.

Discussion

We draw several conclusions from this case. First, data-entry errors are an especially pernicious threat to the credibility of our results. Although failure to find support for a research hypothesis may prompt us to undertake a close review of the dataset to confirm that it is free of data-entry errors, an analysis that yields statistical significance is unlikely to trigger what may be, as in the above example, a time-consuming and difficult effort. This difference in the course taken depending on our data places us within 'the garden of forking paths,' rendering our findings suspect even when we only ever perform a single analysis (Gelman and Loken 2014, 464).

This leads to our second conclusion: to reduce the possibility of data-entry errors, researchers should minimize reliance on manual data entry and maximize the extent to which data wrangling—the 'janitor work' of data science—is performed computationally. Automating 'janitor work' will sometimes require considerable programming effort, but often software is available that makes the task straightforward, such as the `readtext` R package (Benoit, Obeng, et al. 2016) for formatting the contents of text files for text analysis or the `DCPOtools`

R package (Solt, Hu, and Tai 2018) that we employed in our example. In addition to minimizing data-entry errors, writing computer code that starts from the raw source material and works forward has the added benefit of making research much more reproducible (see, e.g., Benoit, Conway, et al. 2016) and hence more credible (see, e.g., Wuttke 2019). As Christensen, Freese, and Miguel (2019, 197) admonish, "Write code instead of working by hand . . . don't use Microsoft Excel if it can be avoided."

Third, when manual data entry *cannot* be avoided, each entry should be made twice, either by different people working independently or by the same person working at a different time, to allow for cross-checking. Double entry is labor intensive, but experiments have shown that while visually inspecting entered data is no better at catching mistakes than simply entering the data once and making no checks, the double-entry approach reduces error rates by thirty-fold (Barchard and Pace 2011, 1837). Given that data-entry errors can completely undermine the validity of our conclusions, as in the example above, double entry is worth the extra effort.

Our final point is for manuscript reviewers. If data-entry errors are invisible to the authors themselves, they are doubly so to reviewers (though if editors provided reviewers with replication materials at the time of the review it may help them to better assess the work's credibility). But the case described above nevertheless suggests a valuable heuristic: when a work's conclusions suggest that a difficult problem will be easily solved—that democratic erosion will reflexively trigger a backlash and a renewed public support for democracy, in the present instance—it warrants especially careful scrutiny.

References

Atkinson, Mary Layton, K. Elizabeth Coggins, James A. Stimson, and Frank R. Baumgartner. 2021. *The Dynamics of Public Opinion.* Cambridge: Cambridge University Press.

Barchard, Kimberly A., and Larry A. Pace. 2011. "Preventing Human Error: The Impact of Data Entry Methods on Data Accuracy and Statistical Results." *Computers in Human Behavior* 27 (5): 1834–39.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95.

Benoit, Kenneth, Adam Obeng, Paul Nulty, Aki Matsuo, Kohei Watanabe, and Stefan Müller. 2016. "readtext: Import and Handling for Plain and Formatted Text Files." Available at the Comprehensive R Archive Network (CRAN).

Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Berkeley: University of California Press.

Claassen, Christopher. 2020a. "In the Mood for Democracy? Democratic Support as Thermostatic Opinion." *American Political Science Review* 114 (1): 36–53.

———. 2020b. "Replication Data for: Does Public Support Help Democracy Survive?" https://doi.org/10.7910/DVN/HWLW0J, American Journal of Political Science Dataverse.

———. 2020c. "Replication Data for: In the Mood for Democracy? Democratic Support as Thermostatic Opinion." https://doi.org/10.7910/DVN/FECIO3, American Political Science Review Dataverse.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–65.

Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53: 69–105.

Lohr, Steve. 2014. "For Data Scientists, 'Janitor Work' Is Hurdle to Insights." *New York Times*, B4.

Solt, Frederick, and Yue Hu. 2015. "dotwhisker: Dot-and-Whisker Plots of Regression Results." Available at the Comprehensive R Archive Network (CRAN). http://CRAN.R-project.org/package=dotwhisker.

Solt, Frederick, Yue Hu, and Yuehong 'Cassandra'Tai. 2018. "DCPOtools: Tools for Dynamic Comparative Public Opinion." https://github.com/fsolt/DCPOtools.

Torres, Rachel. 2017. "Me: Shouldn't There Be Someone in a Basement That We Just Pay to Do All This Awful Data Cleaning? Advisor: That's Who You Are." Twitter. https://twitter.com/torrespolisci/status/886993701855268865.

Wuttke, Alexander. 2019. "Why Too Many Political Science Findings Cannot Be Trusted and What We Can Do about It: A Review of Meta-Scientific Research and a Call for Academic Reform." *Politische Vierteljahresschrift* 60 (1): 1–19.