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

Yue Hu Yuehong 'Cassandra' Tai Frederick Solt

Abstract

Data 'janitor work', the task of getting data into a format appropriate for analysis, has grown increasingly important as political science research has come to depend on data drawn from hundreds and thousands of sources. One tempting solution is to simply enter the data by hand, but this approach raises serious risks of data-entry error, a difficult-to-catch problem with the potential to fatally undermine our conclusions. Underscoring these points, we identify data-entry errors in a prominent recent article, the 2020 study by Claassen that examines the effect of changes in democracy on public support for democracy. We then show that when these errors are corrected, the work's models provide no support for its conclusion that publics react thermostatically to changes in democracy. Researchers should refrain from hand-entering data as much as possible, and we offer suggestions for avoiding the practice.

A growing amount of political science projects tend to be characterized as data science: they employ 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 tiresome for researchers and their research assistants (RAs) but critically important for the scientific quality of the research(see Torres 2017). The problem a mistaken wrangling causes may be more lethal than researchers commonly thought.

While political methodologists have spent decades to develop methods to reduce measurement and specification errors, the equivalently fatal (if not more) problems in data wrangling are rarely systematically detected and researched. The issue is largely conceived in the lament of old-day works [A citation about *Civic culture*] and gossips of academic scandals [A citation of the fake data scandal]. In this letter, we go beyond stories of sloppy

*Corresponding author: yuehong-tai@uiowa.edu. Current version: June 02, 2022.

collectors or intentional p-fishing and uncover the consequences of mistaken data wrangling through a restrict replication.

In particular, we focus on a common wrangling problem that can affect every researchers: 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. As Barchard and Pace (2011) showed, "research assistants" carefully enter data manually, even those instructed to double-check their entries against the original, can cause error rates approaching 1% in just a single roughly half-hour session. Rates likely go up as the tedious task goes on.

In this piece, we illustrate how sneaky and easy data-entry errors can occur and what consequences they cause by scrutiny of the data janitor-work in Claassen (2020c). The research was published in a respectable journal with a strict replication policy. Thanks to that and the author's serious effort on research transparency, we were able to trace the study back to the data wrangling stage and identify the problems. In the following sections, we explain the frequency and magnitude of the data-entry errors and show how the empirics and their inferences change after correcting the errors. In this paper about mass democracy mood across countries, Claassen (2020c, 51) argued that when "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." After correcting the data-entry errors, no empirical evidence supports that public support responds thermostatically to changes in democracy. On this basis, we provide four practical suggestions to help researchers to reduce data-entry errors and their impact.

Data-Entry Errors in a Democratic-Support Context

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). Claassen (2020c) brought in an alternative view that democratic support behaves thermostatically. That is, the increases in democracy yield an authoritarian backlash in the

public, while democratic backsliding prompts the public to rally to democracy's cause.

The empirical test of the relevant inferences takes advantage of recent advances in modeling public opinion as a latent variable. The estimation of the latent variable of democratic support was conducted based upon the information from a variety of survey questions over one hundred countries for up to nearly three decades, constituting a much larger evidentiary base than any previous study (Claassen 2020c, 40). Two particular pieces of data were collected for each distinct survey item: the number of respondents to give a democracy-supporting response—defined, for ordinal responses, as those above the median value of the scale (Claassen 2020c, Appendix 1.3)—and the total number of respondents to whom the question was posed. Each of these 7,538 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. We then entry the data through a software-based automatic process to avoid human errors. By cross-checking our data output and the dataset from the replication file of Claassen (2020c) a following concrete comparison, we identify three types of data-entry errors in the original data wrangling process.

First, technical miscategorization. Researchers (and RAs) may erroneously categorize an answer option to a wrong category when aggregating the original scale to a new one. The error is especially easy to be made when multiple waves of surveys are used whereas the scales of the same question vary across the waves. For example, the Asian Barometer is an important source for Claassen (2020c)'s estimation. The four waves of this survey included the following question:

> 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

[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 2020b). 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 2020a).

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 percentage points in 9 country-years.

Second, theoretical miscategorization. Researchers and RAs may leave out or mistakenly put certain answer options in a category that is theoretically inappropriate during the aggregation. An example comes from the Pew Global Attitudes surveys' four-point question asking about the importance of living in a country with regular and fair contested elections:

> 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?

Rather than including respondents who gave both responses above the median—"very important" and "somewhat important"— Claassen (2020c) recorded only those respondents who answered "very important" as supporting democracy. This error caused substantial underreporting of the extent of democratic support in 91 country-years.

Another example relates, again, to the Asia Barometer, which 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 2020c, 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. 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.‘ The miscategorization led to overestimations of the percentage of democracy-supporting responses ranging from 19 to 63 percentage points and averaging 42 points.

Third, design misunderstanding. In many surveys, question items were not open to all the respondents. Neglecting such design can lead to For example, when the Americas

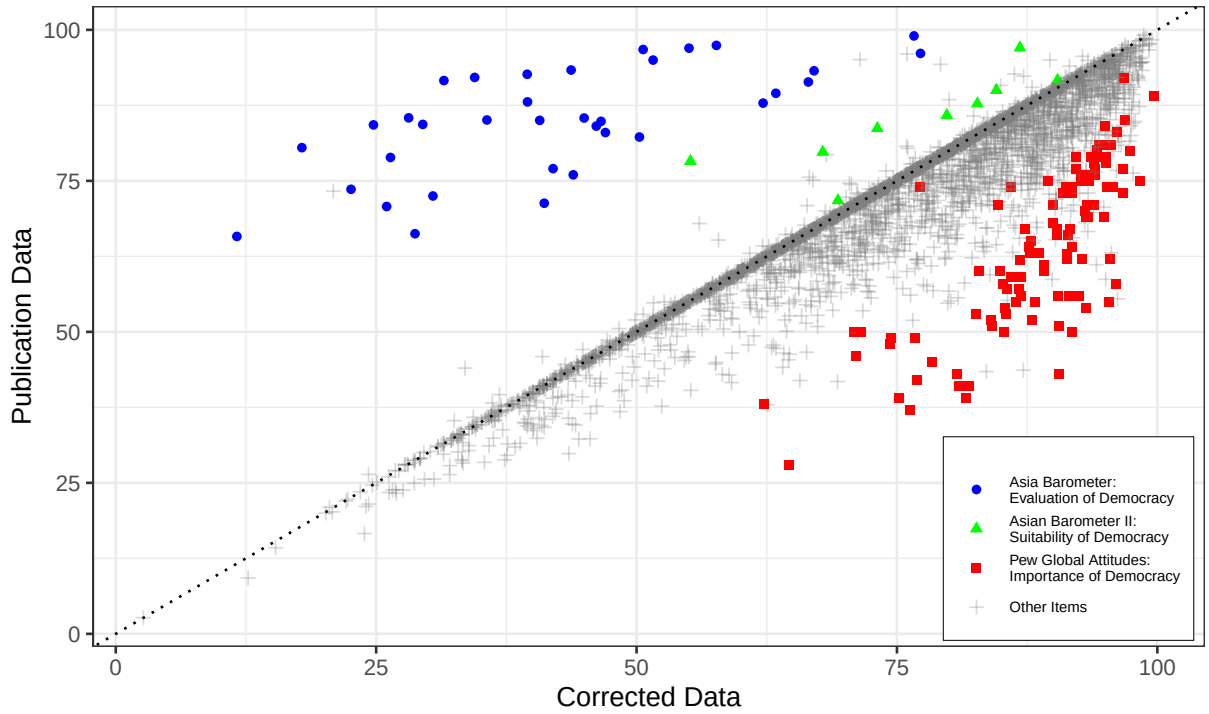
Barometer surveyed Canada in 2010, it included an item asking whether, when "democracy doesn't work," Canadians "need a strong leader who doesn't have to be elected through voting." It posed this question to only half of its sample. Those who were not asked the question, however, were included in the 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 2020c, Appendix 1.3). This rule may be a reasonable coding choice, but including in this category those who were never asked the question at all is clearly a data-entry error.

Fourth, weight misapplication. Weighting raw survey results to maximize the extent to which they are representative of the target population is important. Relying on toplines reported in 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.

Consequences for Inference

The accumulation of the above issues can lead to salient deviation of the output data from what researchers initially design. We illustrate such difference by comparing 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. The result is visualized in Figure 1. When points fall along the plot's dotted line, 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.

For 49% of the country-year-item observations, the difference between these percentages was negligible—less than half a percent—yielding points approximately along the plot's dotted line. But for the remaining observations, the difference was often substantial as a result of data-entry errors in the publication data.



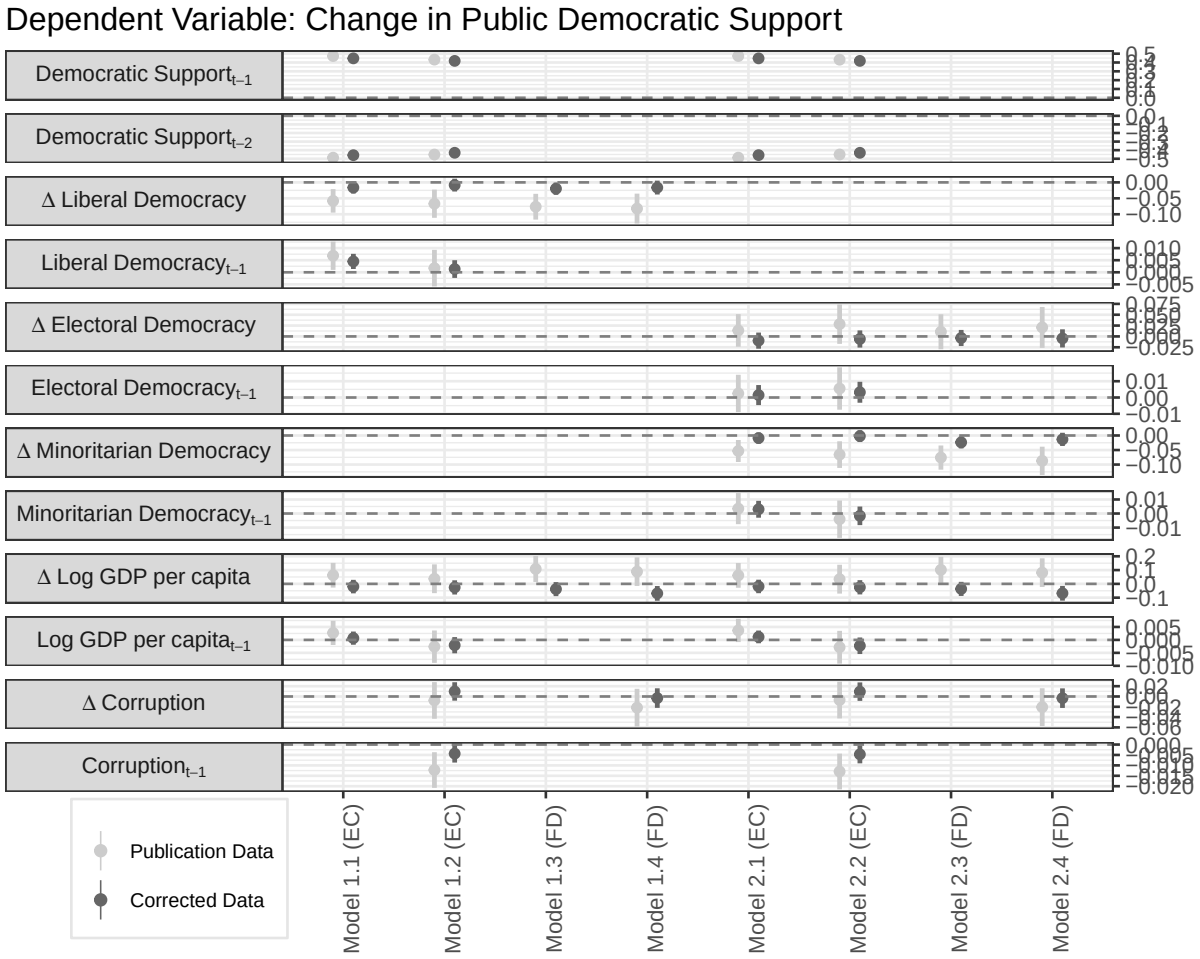
Notes: Each point represents the percentage of respondents in a country−year to give a democracy−supporting response to a particular survey item. Publication data is as reported in Claassen (2020b); the corrected data was collected directly from the original surveys. The Asia Barometer's item on the 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

If using estimates based on such data 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 (2020c) exactly using both of these 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 with the corrected data.



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

Models 1.1 through 1.4, which replicate those presented in Table 1 of Claassen (2020c, 47), examine the effects of overall liberal democracy using error-correction models (labeled EC in Figure 2) and first-difference models (labeled FD). As Claassen (2020c, 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 the thermostatic theory predicts. 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 2020c, 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 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 few parties actually engaged in eroding democracy put 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

The above analyses illustrate that data-entry errors are an especially pernicious threat to the credibility of our results. They are not only fatal but also sneaky—much sneaker than other errors. There is almost no way to find them unless through a scrutiny of the data collection process. Unfortunately, in practice, such scrutiny rarely happens. 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. These different courses put us in 'the garden of forking paths,' rendering our findings suspect even when we only ever perform a single analysis (Gelman and Loken 2014, 464).

To avoid staying at this position, we provide three practical suggestions for researchers to minimize the data-entry from the design stage **1. Automating the data entries.** Researchers should minimize reliance on manual entry and maximize the extent to which the "janitor work" of data science is performed computationally. Such work 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 2020) that we employed in our example. An additional benefit of programming the data-entry process is to make research much more reproducible (see, e.g., Benoit, Conway, et al. 2016) and hence more credible (see, e.g., Wuttke 2019). Christensen, Freese, and Miguel (2019, 197) admonish at a more specific level, "Write code instead of working by hand . . . don't use Microsoft Excel if it can be avoided."

2. Cross-checking, especially for large data-entry projects. When manual data entry *cannot* be avoided, each entry should be made twice. Double entry is labor intensive, but experiments have shown that 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. Even a cross-checking by the same author can be helpful. We suggest a certain period between the initial entry and the checking. On the other hand, if the researchers are resourceful, we highly suggest the cross-checking to be conducted by separate persons. This relates to our next suggestion:

3. Team work, if possible. For any project involving data entries, a team work is often preferred than a single-person work. Except for labors to conduct more reliable cross-checking, splitting the tasks can reduce the risk of human entry errors due to tiredness and ignorance. Each researchers (and RAs) are also more likely to read the codebook and instructions of the sources carefully and comprehensively than processing the entire process by oneself.

4. Being aware of the dangers of data-entry errors and common types. This is a suggestion especially 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.

The last but not least, some readers may reach a conclusion that, after reading this piece, researchers would be scared by the difficulty to conduct "errorless" data-entries or tend to provide a less transparent pile of replication files to prevent others to discover errors potential data janitor work. They are wrong. Looking back the history of political science, researchers gradually level up their requirement for an acceptable research, from single-variable analysis to multiple regression, from merely reporting the result to preregistration and *Data Access & Research Transparency*. Along the same line, five years ago, political scientists fought for data misconduct that the academia has reached a consensus now as an absolute "no-no" (Solt et al. 2016, 2017). Today, we propose a new movement to level up the requirement and self awareness at the data-entry stage. We hope the effort can promote scientific quality of the research in the discipline further.

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. https://doi.org/10.1017/S0003055416000058.

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. "Replication Data for: Does Public Support Help Democracy Survive?"

———. 2020b. "Replication Data for: In the Mood for Democracy? Democratic Support as Thermostatic Opinion."

———. 2020c. "Does Public Support Help Democracy Survive?" *American Journal of Political Science* 64 (1): 118–34. https://doi.org/10.1111/ajps.12452.

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 (01): 69–105.

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

Solt, Frederick. 2020. "Measuring Income Inequality Across Countries and Over Time:

The Standardized World Income Inequality Database." *Social Science Quarterly* 101 (3): 1183–99. https://doi.org/10.1111/ssqu.12795.

Solt, Frederick, and Yue Hu. 2015. "Dotwhisker: Dot-and-Whisker Plots of Regression Results." CRAN: Available at The Comprehensive R Archive Network (CRAN).

Solt, Frederick, Yue Hu, Kevan Hudson, Jungmin Song, and Dong "Erico" Yu. 2016. "Economic Inequality and Belief in Meritocracy in the United States." *Research & Politics* 3 (4): 1–7. https://doi.org/10.1177/2053168016672101.

———. 2017. "Economic Inequality and Class Consciousness." *The Journal of Politics* 79 (3): 1079–83. https://doi.org/10.1086/690971.

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.

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.