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

Yue Hu Yuehong 'Cassandra' Tai Frederick Solt

Many facets of Computational social science Data wrangling, the task of getting the needed data into the format required to perform analyses.

(Wickham and Grolemund 2017, xi)

Such data 'janitor work' is often viewed as tiresome and as better performed by someone, anyone, else.[1]

DGP Problem and Consequences

We illustrate above DGP problems and their potential consequences with Claassen (2020). The study published in a very prestige journal of political science with clear replication requirements. Similar or relative data were also used in publications in other top journals of the field. We appreciate the author's replication materials to enable the this scientific scrutiny. Based on them and the author's description in the paper, we can largely infer how the measurements of variables are constructed. We apply the same methods on the full available data with consistent coding. By comparing the results with the original paper, we identify two primary problems of DGP, data discrepancy and coding inconsistency, which lead to results implying substantively different conclusions from the original publications. We tend to use this case to show that the current consensus of publication replication for sure progress the scientificness of political science, whereas it does not prevent research from DGP problems and that may lead to severe consequences.

*Corresponding author: yuehong-tai@uiowa.edu. Current version: May 23, 2022.

[1]As Torres (2017) wrote, recounting her experience as a research assistant, "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."

Maybe also something about the documentation by Herndon, Ash, and Pollin (2014) of the problematic data selection in Reinhart and Rogoff (2010)

Related: Troeger (2019, 285) says, "Another potential measure is to make all data publicly available. Again many journals require data and code to be made available to the public before publication. But often there are no requirements whether source data has to be included. When source data is original, confidential, or personalized, publication might not be possible or undesirable. However, new avenues to make this kind of data available for replication need to be explored."

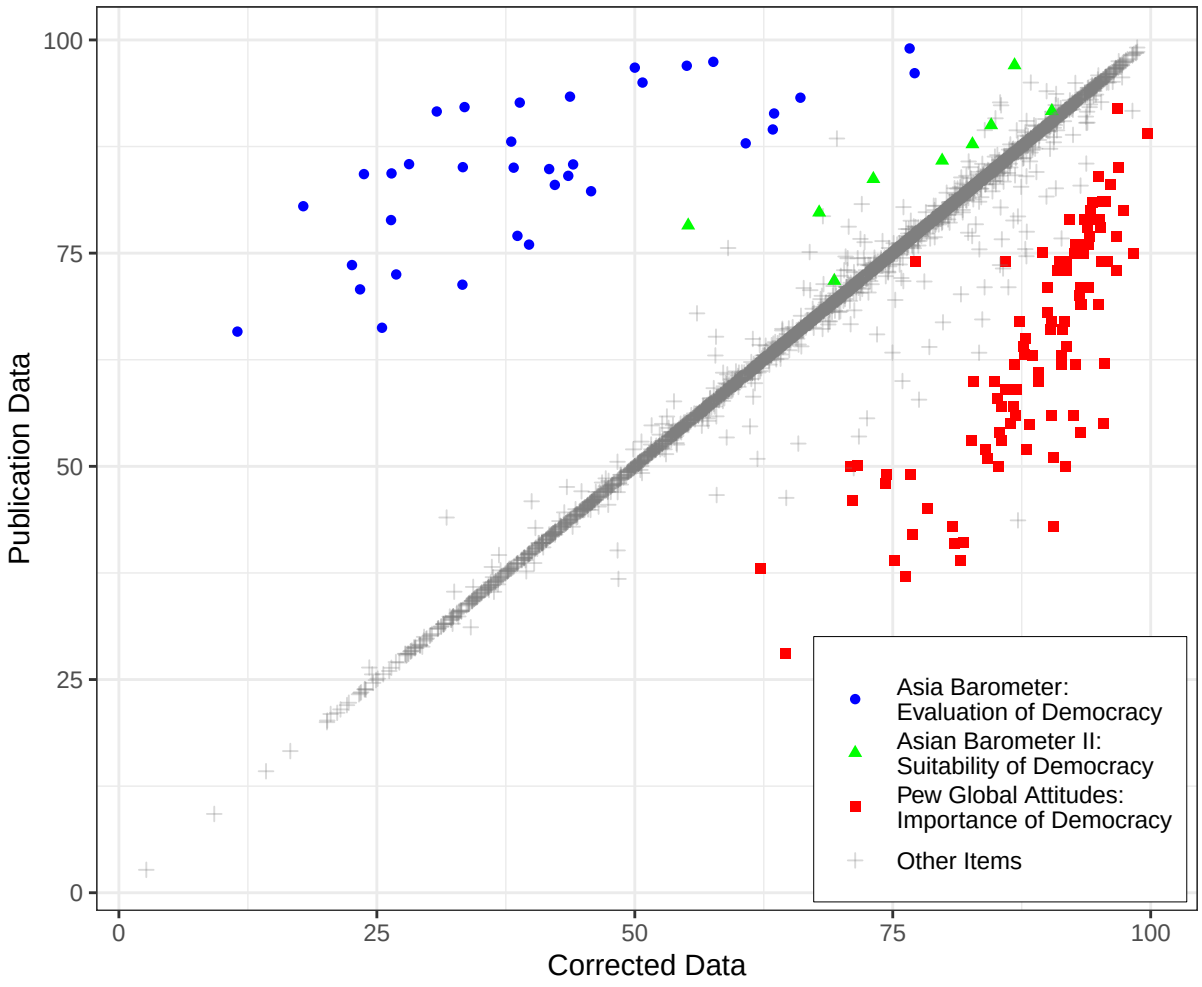
Data-Entry Errors: The Case of Thermostatic Democratic Support

Claassen's coding rule

In Figure 1, we compare the percentage of respondents to give a democracy-supporting response in the publication's replication data with the percentage we found when we automated the process of wrangling these same data from their original surveys. When points fall along the 45° diagonal, it indicates that the two data-wrangling processes yielded 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 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, as a result of data-entry errors, the difference was often substantial. 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 2020, Appendix 1.3), only answers above the median response category 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-

[2]Although this may 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



Notes: Each point represents the proportion of respondents in a country−year to give a democracy−supporting response to a particular item. Publication data is as reported in Claassen (2020a); 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

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 included. This data-entry error resulted in overestimates of as much as 23.1% 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.

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 only posed this question to 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. This

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.

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.

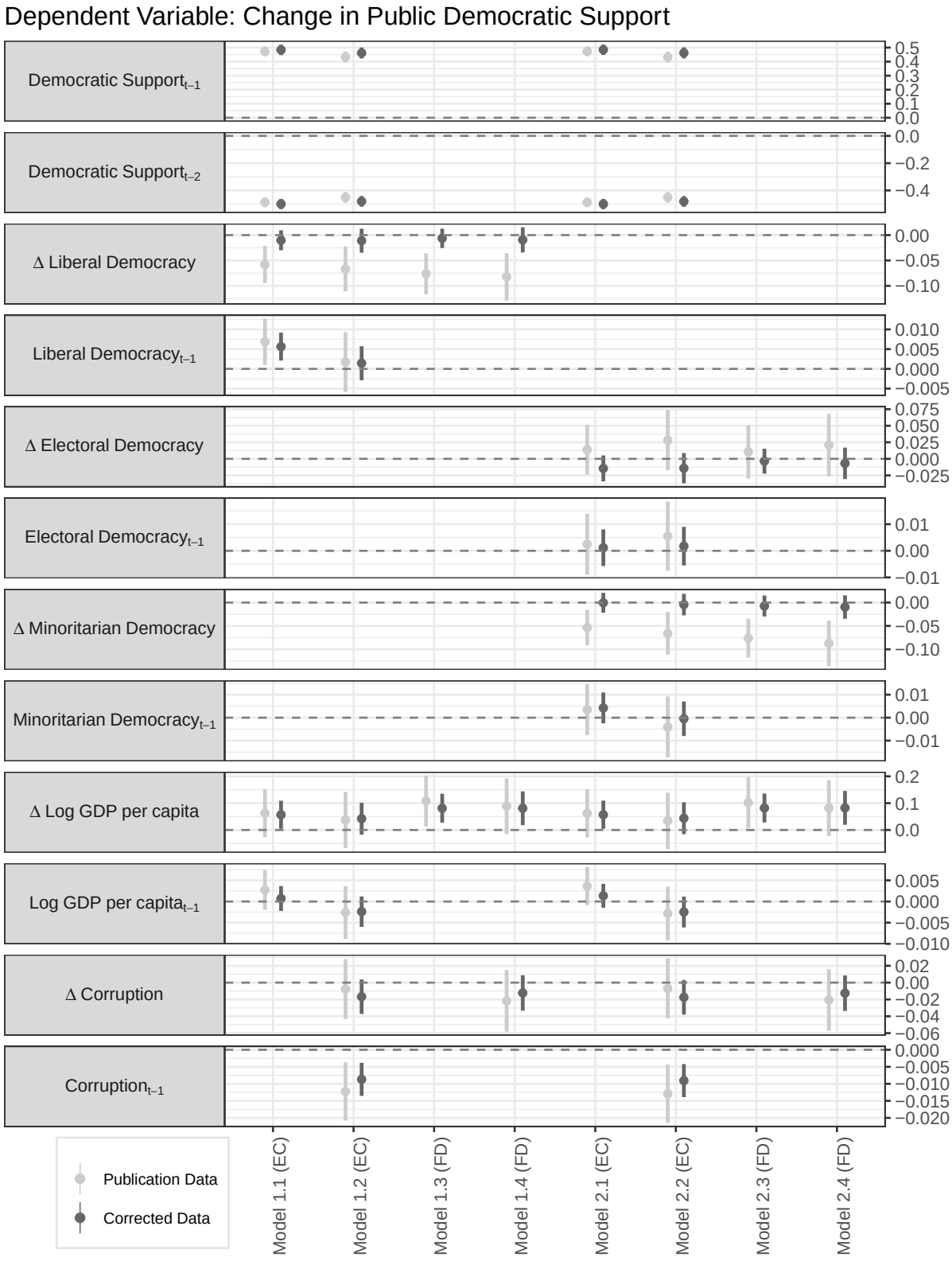
Another source of data-entry errors in this study involves survey weights. Weighting raw survey results to maximize the extent to which they are representative of the target population is important. Relying on the toplines reported in survey codebooks rather than the survey data itself evidently caused 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

Data-entry errors of this sort can yield erroneous conclusions. A now-classic literature maintains that experience with democratic governance generates robust public support for democracy (see, e.g., Lipset 1959). Claassen (2020) argued 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. We replicated each of the models presented in Claassen (2020) exactly, first with the article's original dataset and then making the corrections to the errors in the data we described above. The results provide little support for either the classic or the thermostatic argument.

Figure 2 presents our results in a "small multiple" plot (Solt and Hu 2015) for a clear comparison of the coefficients of each variable across the 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 (2020, 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 (2020, 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



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

lagged levels of liberal democracy will be positive. When using the original publication data with their data-entry errors, the coefficients estimated for the change in liberal democracy are large, negative, and statistically significant across all four models, 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, indicating that "this effect is not particularly robust" (Claassen 2020, 47).

When the data-entry errors are corrected, however, the results for these models suggest a 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 for the classic theory, such as it is, 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.

Discussion

We draw three conclusions from the foregoing. First, researchers should minimize reliance on manual data entry and maximize the extent to which data collection and wrangling—the 'janitor work' of data analysis—is performed computationally. While automating 'janitor work' will sometimes require considerable programming effort, often software is available that makes the task straightforward, such as the `readtext` R package for formatting the

contents of text files for text analysis (Benoit, Obeng, et al. 2016) or the `DCPOtools` R package for wrangling survey data to be used in latent variable analyses (Solt, Hu, and Tai 2018).[3]

Second, 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, as shown above, data-entry errors can completely undermine the validity of our conclusions, double entry is worth the extra effort.

Finally…

References

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, Benjamin E. Conway Drew And 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).

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

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal*

[3] 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 replicable (see, e.g., Benoit, Conway, et al. 2016).

of Economics 38 (2): 257–79.

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

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100 (2): 573–78.

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.

Troeger, Vera. 2019. "To p or Not to p? The Usefulness of p-Values in Quantitative Political Science Research." *Swiss Political Science Review* 25 (3): 281–87.

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science*. O'Reilly. http://r4ds.had.co.nz.