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 19, 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."

Identification of Data and Coding Problems

The problem of data discrepancy refers to the practice that researchers consciously or unconsciously select to use only a part of data that are available. When hypothesis tests rely on statistical estimates, both frequentist and Bayesian methodologists have emphasized the importance of the sufficiency of qualified data for producing unbiased and efficient estimates [XXX]. Empirical studies have also well documented that the insufficient use of available data would cause lethal misunderstanding of data trends and unreliable conclusions (Solt et al. 2016, 2017).

In our illustrative case, the 2020 paper, the point of interest is the influence of the institutional democracy on people's support of democracy, i.e., "democracy mood." To measure the mood, the author uses existing survey questions about the "appropriateness or desirability of democracy, compare democracy to some undemocratic alternative, or evaluate one of these undemocratic forms of government" to draw a latent variable with a dynamic IRT method (Claassen 2020, 40). After the DGP, he collected 3,768 nationally aggregated opinions from 52 different survey questions of 14 survey projects.

For instance, the original study excludes observations from the third and fourth waves of Asian Barometers (the data were released in 2009 and 2017) on questions about to what extent people want their country to be democratic now, although it included them from the first and second waves. This accounts for 19 of 38 excluded country-year-items in 'available'.

The miscoding problem is caused by researchers' inconsistent coding. Here we are not talking about the coding manipulations for "p-hiking" or "p-fishing," but that scholars apply invalid or variant coding method on the data that should be coded consistently [XXXX]. The problems easily occur when scholars intend to use multiple measurements for the same variable. In Solt et al. (2016), the authors show that how three seemingly valid measurements produce considerably different outcome estimates (2016, 4). Howerver, the same problem can also occur even when only one measurement is used.

There are two potential miscoding types. One is operation-based miscoding. Researchers may mistakenly coded the values reversely, incompletely, or recognizing the value to mark missing response to a true variable value. For instance, we found in Claassen (2020) that

the author coded the option "Necesitamos un líder fuerte que no tenga que ser elegid" as "Strong_lapop_1" for the question in AmericasBarometer:

> AUT1:Hay gente que dice que necesitamos un líder fuerte que no tenga que ser elegido a través del voto. AUT1 Otros dicen que aunque las cosas no funcionen, la democracia electoral, o sea el voto popular, es siempre lo mejor. ¿Qué piensa usted? [Leer alternativas]

However, according to the publication's supplementary materials, "Strong_lapop_1" refers to the question in AmericasBarometer with wording, "On some occasions, democracy doesn't work. When that happens there are people that say we need a strong leader who doesn't have to be elected through voting. Others say that even if things don't function, democracy is always the best. What do you think?". "Strong_lapop_2" refers to the question with wording "There are people who say we need a strong leader who does not have to be elected. Others say that although things may not work, electoral democracy, or the popular vote,is always the best. What do you think?" Therefore, this question should be coded as "Strong_lapop_2" instead of "Strong_lapop_1".

The other type of problem is design-based mis(re)coding. Researchers conduct consistent method during the coding but the method per se under-represent or stretch the variance of the variables. For examples, when recoding the questions about how people evaluate democracy [2] in the Asia Barometer, Claassen (2020) counted the midpoint of its three-point scale as a positive, democracy-supporting response. The coding results in substantial overreports in 35 country-years. Similarly, in the *second* wave of the Asian Barometer, the study counts 5 along with 6 through 10 as positive responses for questions about the suitability of democracy[3] and desire for democracy[4]. This coding produces more modest overreports in 9 country-years. A third example in recoding Pew Global Attitudes surveys, the researcher counted only the highest value of the four-point scale of the question on how

[2]The question wording is "I'm going to describe various types of political systems. Please indicate for each system whether you think it would be very good, fairly good or bad for this country – A democratic political system"

[3]The question is "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?"

[4]The questions is "To what extent do you want our country to be democratic now?"

important to live in a courntry with regular elections[5] as a positive response and so leading to substantial underreports in 91 country-years.

A final example we found that the 2020 study counts nonresponses as negative responses (such as in Pakistan in 2005 in the South Asian Barometer, when 636 of 1324 respondents declined to answer whether they considered democracy suitable for their country, and these were all coded as unsupportive responses) yielding under-reports. Moreover, the study does not always employ survey weights, which can shift proportions somewhat in either direction.

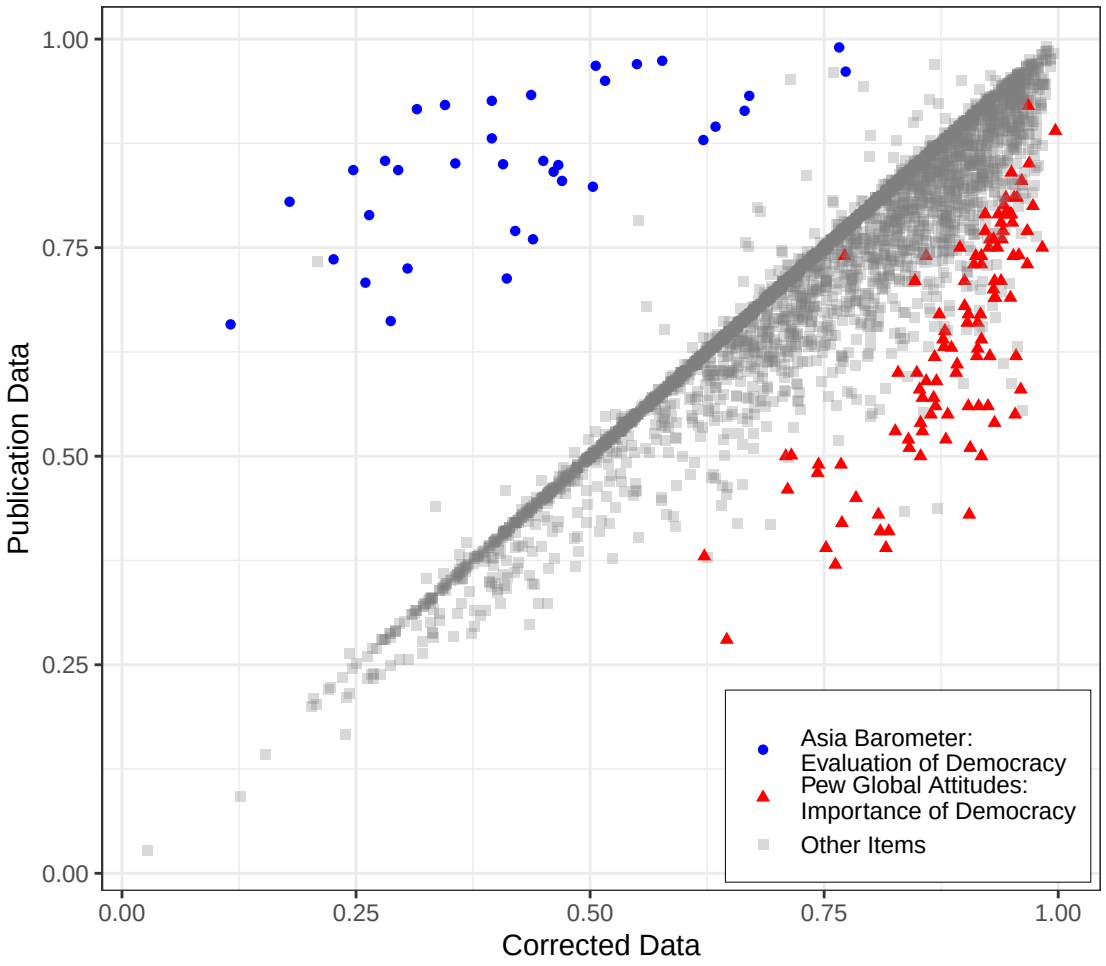
Both the data-discrepancy problem and design-based mis(re)coding are covert DGP problems that are often difficult to be detected merely through reproductions of the replication files. However, the discrepancy they cause could be substantial. We compare the distributions of the core explanatory variable, democracy support, from the replication data of the publication and a version that fixed all the problems discussed above at both country and survey levels.

In Figure 1, we compared democracy support from the publication and corrected data across surveys. The extent the points in the plot deviating from the 45° diagonal represents the discrepancy of the publication data from the full, corrected-coding data. The comparison witnesses a large amount of data from various surveys, especially from Pew Global Attitudes Survey, were underreported and a fair amount of data, mainly from the Asia barometers, were overreported.

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

[5]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, somewhate important, not too important or not important at all?"



Notes: Each point represents the proportion of respondents in a country−year to give a democracy−suppo
Publication data is as reported in Claassen (2020a); the corrected data was collected directly from the orig
on the evaluation of democracy accounts for most overreports, and the Pew Global Attitudes item on the in
large underreports. In both cases, these issues can be easily explained by errors in transcribing the data. I
inconsistent treatment of missing data and/or survey weights, reflecting in part differences in codebook rep

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

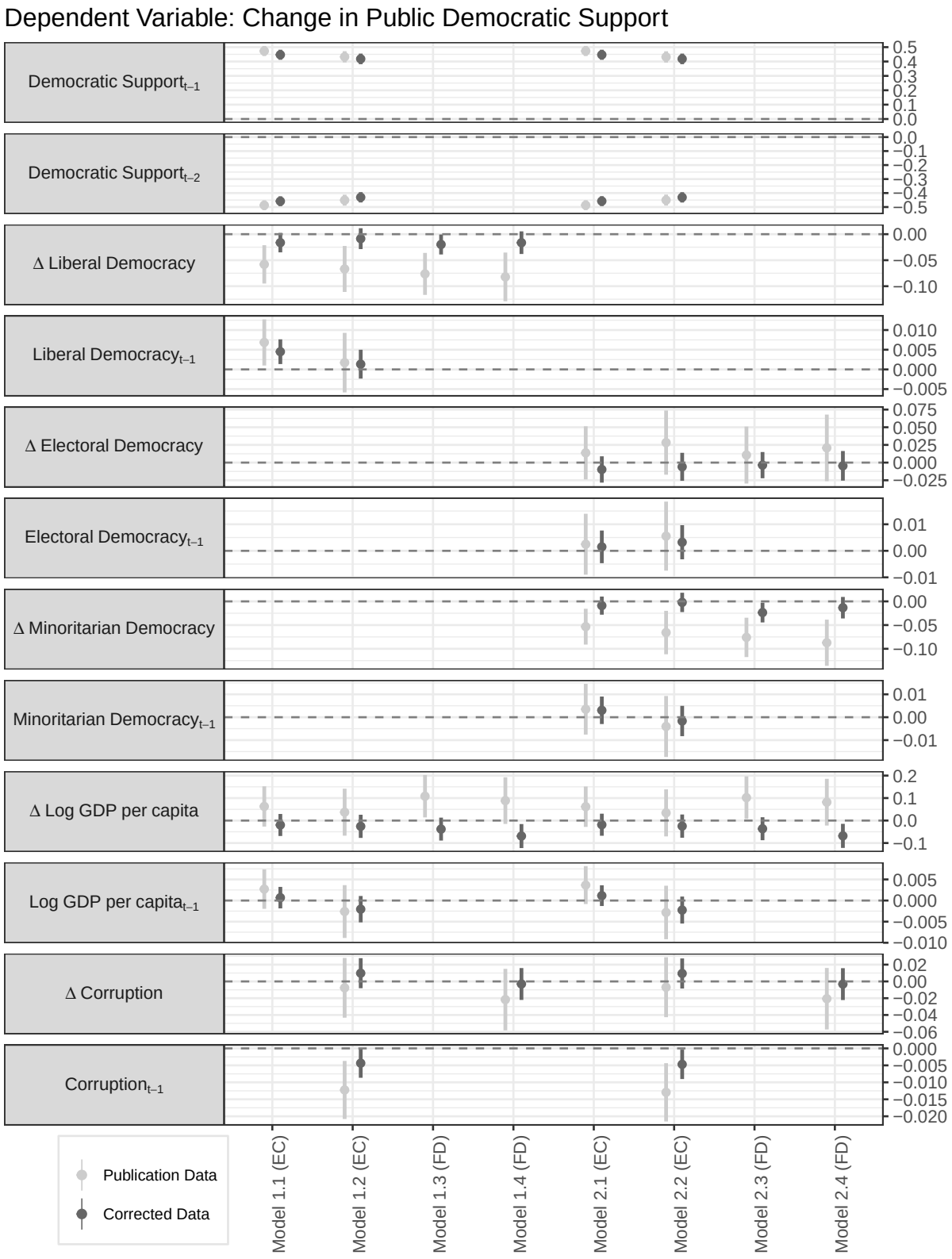
original dataset and then making the corrections to the errors in the data we described above. The results provide little support for either the classic or the thermostatic argument.

Figure 2 presents our results with 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 are point estimates and the whiskers represent 95% confidence intervals. Each row represents a variable's performance in its own scale across all of the models. The lighter entries 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 lagged levels of liberal democracy will be positive. When using the original data with 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 the models are better estimated in the corrected data, but so do the coefficients. The estimate for the change in liberal democracy remains negative and statistically significant only in the first-difference specification employed by Model 1.3. It fails to reach statistical significance in Claassen's (2020) preferred error-correction Models 1.1 and 1.2, and it likewise disappears when corruption is added to the first-difference model in Model 1.4.

Models 2.1 through 2.4, which break liberal democracy into its electoral democracy and minoritarian democracy components, follow the same pattern. The finding reached in analyses of the original dataset with its data-entry errors intact that public democratic support responds thermostatically to changes in minoritarian democracy evaporates when the 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

entry errors are corrected. Only in Model 2.3 is this coefficient negative and statistically significant, and even then it is much smaller than the published result. When the data-entry errors in the publication dataset are corrected, the evidence for the thermostatic theory is not robust.

Discussion

We draw two 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. When manual data entry cannot be avoided, double entry

Second,

With the previous sections, we identified two types of DGP problems, data discrepency and miscoding. The former refers to the problem that not all available data are included in the analysis. The latter relates to human or design errors to (re)code information from the original data source inconsistently or unvaildly. We use Claassen (2020) to illustrate these issues in practice and how correcting them would substantively change the conclusions of hypothesis tests.

The DGP problems are often too convert to be detected merely through reproduction of the results, but they can lead to severe misunderstanding of the empirical outcomes of scientific inquiries. A better way to deal with the DGP problems is more doing with the design than post-estimate phases. Here we give researchers four suggestions to minimize DGP problems in their research. First three are for the authors, and the last one is for the reviewers.

First, automatic downloading. We suggest researchers to maximumly utilize the data scraping functions of programming languages (such as R and Python) to collect data through with widecard searching and directly from the original sources. For instance, you can easily use the package `icpsrdata` to obtain all the latest data from ICPSR or use [`pewdata`] to access all types of data published by Pew. Instead of manually updates, the programming way collect data directly from the datasets, which reduces the missing probability.

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."

Second, automatic coding. We suggest researchers to process their data coding as automatically as possible. Instead of coding every variables with separate codes, building up packed-up functions or even packages for doing batch coding. For instance, if you work with R, you can find functions such as `stm::textProcessor` or packages like `DCPOtools`. The former prepares raw text data for more complex text-analysis tasks, and the latter transfer different survey questions into a consistent format ready for analysis and comparison. By this way, the authors can ensure every variable is treated in the same manner. Of course, the actual data cleaning process is complicated that maybe not all the steps can be packed into functions or batched as a process. In this scenario, we gives the following suggestion.

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

Third, replication from scratch. (Possibly roll this in with "first") Given the complication of the data management, we suggest researchers to provide fully replication codes on the raw data. Researchers ought to provide not only replication files that can reproduce the results in the publications but concrete coding programs or coding books that others can replicate the DGP process. This is an advanced requirement for data transparency and replicability. Yet as we showed, there could be serious drawbacks to ignore this step, especially for studies based on existing datasets.

Finally, we hope this replication can alert not only the researchers but also the reviewers of the importance of DGP. It is usually the very first empirical step for a research and a crucial step determining the validity of all the outcomes and conclusions. We appeal to both the academic journals and reviewers to pay more attention to this phase when assess a manuscript. The substantive difference elaborated in this paper demonstrate that the DGP problem is no less vital than reproduction or any other important methodological problems.

References

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 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, 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.

———. 2017. "Economic Inequality and Class Consciousness." *Journal of Politics* 79 (3): 1079–83.

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.