On 'Janitor Work' in Political Science: Best Practices for Wrangling Data Before Harmonization*

Yue Hu[1] Yuehong Cassandra Tai[2] Frederick Solt[3]

This article focuses on a preliminary step in any ex-post data harmonization project—wrangling the pre-harmonized data—and suggests best practices for helping scholars avoid errors in this often-tedious work. To provide illustrations of these best practices, the article uses the examples of pre-harmonizing procedures used to produce the Standardized World Income Inequality Database (SWIID), a widely used database that uses Gini indices from multiple sources to create comparable estimates, and the Dynamic Comparative Public Opinion (DCPO) project, which creates a workflow for harmonizing aggregate public opinion data.

[1] Department of Political Science, Tsinghua University, Beijing, China

[2] Center for Social Data Analytics, Pennsylvania State University, University Park, USA

[3] Department of Political Science, University of Iowa, Iowa City, USA

*Corresponding author: yuehong-tai@uiowa.edu. Current version: February 09, 2025. Replication materials and complete revision history may be found at https://github.com/fsolt/wrangling_data. The authors contributed equally to this work. Yue Hu appreciates the funding support from the National Natural Science Foundation of China (72374116) and Tsinghua University Initiative Scientific Research Program (2024THZWJC01).

1 The Problem with 'Janitor Work'

Most data harmonization projects—and a growing volume of other political science research—can 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 harmonization or analysis, is notoriously the bulk of the work (see, e.g., Lohr 2014). Such '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 to scientific inquiry, and errors that arise during this process can undermine our data-harmonization goals.

One kind of data-wrangling error presents a particularly insidious problem: errors that occur during manual data entry. Faced with the task of getting data into the correct format before harmonizing, 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. This technique may be straightforward, but it is very much prone to error. 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 a single roughly half-hour session. Rates likely go up as the tedious task goes on. Although the pernicious consequences of data-entry errors are easily grasped in everyday contexts—Haegemans, Snoeck, and Lemahieu (2019, 1) collects examples of misrouted financial transactions and airline flights—they have thus far gained little attention in political science, even among those working to harmonize large quantities of data.

We suggest three best practices for reducing the rate of data-entry errors. First, *automate data entry* to the greatest extent possible. Second, *use the double-entry method*: when manual data entry cannot be avoided, each entry should be made twice, either by separate researchers or sequentially. Third, *embrace teamwork* for any project involving entering data by hand, splitting the task up among team members will reduce the risk of errors going undetected. We next demonstrate the application of these practices within two ongoing harmonization efforts, the Standardized World Income Inequality Database (SWIID) and the Dynamic Comparative Public Opinion (DCPO) project.

2 Wrangling Income Inequality Data for Harmonization

The Standardized World Income Inequality Database (SWIID) is a long-running project that seeks to provide harmonized income inequality statistics for the broadest possible coverage of countries and years (Solt 2009, 2015, 2016, 2020a). As of its most recent update at the time of this writing, its source data consists of some 27,000 observations of the Gini coefficient of income distribution in nearly 200 countries over as many as 65 years, collected from over 400 separate sources including international organizations, national statistics bureaus, and academic studies.[1]

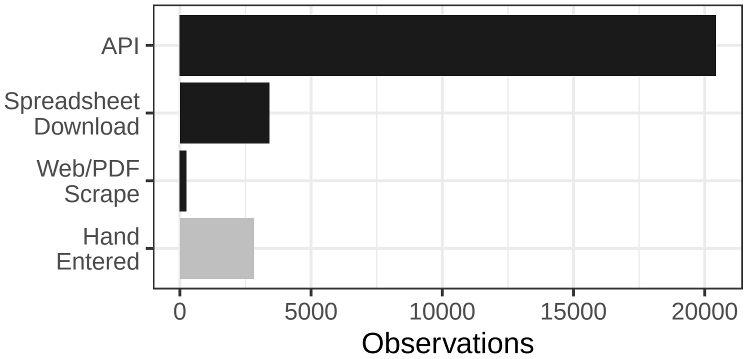


Figure 1: Income Inequality Observations by Method of Collection

In early versions of the SWIID, all of these source data were entered by hand, and checks of newly entered observations revealed that error rates were high. Moreover, many of the sources consulted frequently update or revise their figures. To avoid data-entry errors and ensure that updates and revisions are automatically incorporated as they become available, for the past decade the process of collecting the source data has been automated to the greatest extent practicable (Solt 2020a, 1184–85).[2]

Most international organizations and a few national statistical bureaus use application programming interfaces (APIs) that facilitate automating the inclusion of their data; the R community has often built packages using these APIs to make the task even easier (see Blondel 2018; Lahti et al. 2017; Lugo 2017; Magnusson, Lahti, and Hansson 2014; Wickham, Hester, and Ooms 2018). The SWIID takes as much advantage of these resources as possible, as shown in Figure 1. Although the sources with APIs are relatively few, they

[1] Those who are interested can access and explore these data on the web at https://fsolt.org/swiid/swiid_source.html.

[2] The R code for this automated data collection can be viewed here: https://github.com/fsolt/swiid/blob/master/R/data_setup.R.

contain by far the most data: 76% of the observations are collected in this way. When no API is available, the automation script downloads and reads any available spreadsheets (see Wickham 2016). In the absence of a spreadsheet, the process of scraping the data either directly from the web or, preferably, from a pdf file (see Sepulveda 2024) is automated. Together the collection of 90% of the source data is scripted. This means not only that the possibility of errors introduced by hand entry for a vast majority of observations is eliminated but also that the updates and revisions that are frequent in these data are automatically incorporated as they become available.

However, it also means that some 10% of the observations are entered by hand.[3] Many sources contain just a handful or fewer observations, making the payoff to the often laborious process of data cleaning too small to justify the effort. Some sources—including most academic articles—are behind paywalls, making automation particularly challenging. When these sources contain more than a handful of observations, these are still collected using Sepulveda's (2024) `tabulapdf` R package to avoid data-entry errors. Other sources, such as many books, cannot be read directly into R. And finally, one source contains crucial information encoded in the typeface of its tables (see Mitra and Yemtsiv 2006, 6); this information would be lost if the tables were read directly into R. All such new observations are entered twice into separate spreadsheets. Most often this has been done by two different investigators, but sometimes sequentially by a single researcher. Either way, using this double-entry method allows for automated cross-checks of the newly entered data that increase the chances that errors are identified and corrected (see Barchard and Pace 2011).

To summarize, the process of collecting the source data for the SWIID is 90% automated, and the dual-entry method is employed for the remaining 10%. The upshot of this process is that the SWIID's harmonized estimates of income inequality are reliable, frequently updated, and employed around the world by international organizations, central banks, and researchers in academia and beyond.

[3] The resulting spreadsheet can be found at https://github.com/fsolt/swiid/blob/master/data-raw/fs_added_data.csv.

3 Wrangling Public Opinion Data for Harmonization

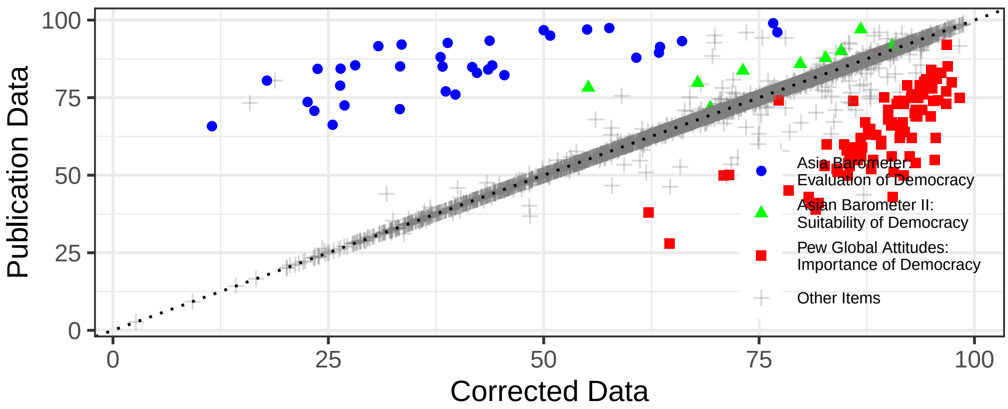
Scholarship on comparative public opinion only rarely benefits from relevant items asked annually by the same survey in many countries (examples of such fortunate works include Norris (2011, 70–77) on trust in government and Hagemann, Hobolt, and Wratil (2017) on attitudes toward European integration). To address the lack of cross-national and longitudinal data on many topics, a number of works have presented latent variable models that harmonize available but incomparable survey items (Caughey, O'Grady, and Warshaw 2019; see Claassen 2019; Kołczyńska et al. 2024; McGann, Dellepiane-Avellaneda, and Bartle 2019; Solt 2020b). This approach has been used to generate cross-national time-series measures of public opinion on a range of topics, from economic, social, and immigration conservatism (Caughey, O'Grady, and Warshaw 2019) to trust in government (Kołczyńska et al. 2024). The Dynamic Comparative Public Opinion (DCPO) model presented in Solt (2020b) in particular has been employed to measure gender egalitarianism (Woo, Allemang, and Solt 2023), political interest (Hu and Solt Forthcoming), and support for gay rights (**?**), among other aspects of public opinion (see https://dcpo.org/).

This work is by nature extremely data intensive, drawing on information extracted from dozens if not hundreds of survey datasets. In DCPO projects, the collection of the source data proceeds in two steps: first, gathering the relevant survey items, and second, accumulating the survey responses.

The first step, gathering the relevant survey items, involves identifying surveys that contain questions on the topic of interest and then recording in a spreadsheet the survey, the variable representing the questions of interest in the dataset, the question text, the response values ordered from least to most of the concept being investigated, and the original textual response categories. At present, this step is done entirely by hand. To minimize data-entry errors, the dual-entry method is used here too. Multiple collaborators go through the entire process separately, and the resulting spreadsheets are compared to catch the omnipresent data-entry errors.

The second step, accumulating the survey responses, is fully automated through the use of the `DCPOtools` R package (Solt, Hu, and Tai 2018). When passed the spreadsheet created in the first step, this package reads in each survey dataset recorded, extracts the variable of interest, reorders the response values for this variable from least to most of the concept investigated, and aggregates the number of respondents in each of the reordered response categories in each country and in each year the survey was fielded. The package also automatically ensures that country names are standardized using the excellent `countrycode` package (Arel-Bundock, Enevoldsen, and Yetman 2018) and that the years accurately reflect actual fieldwork dates using internal crosswalk tables. The aggregated number of respondents for each observed response-item-country-year then serve as the source data for the latent variable model.

We re-collected all of the source data for the publication from the original surveys. We identified the variables of the survey items used by the article within each survey dataset, and then we used an automated process to collect the needed data from the survey datasets while avoiding data-entry errors (see Solt, Hu, and Tai 2018). In Figure 2, we compare the percentage of respondents to give a democracy-supporting response in the hand-entered publication spreadsheet with the percentage we found using our automated process of wrangling these same data. 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.



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

Figure 2: Comparing Democracy-Supporting Responses in Hand-Entered and Machine-Collected 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 plot's dotted line. But for the remaining observations, the difference was often substantial due to 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 to the study's coding rules (see Claassen 2020, 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.[4] 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 responses. This data-entry error resulted in overestimates of as much as 23 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.

[4] 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 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 supportive.

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, asked half its sample, when "democracy doesn't work," Canadians "need a strong leader who doesn't have to be elected through voting." Those who were not asked the question were included in the total number of respondents. According to the study's coding rules, refusing to answer is equivalent to answering in a fashion not supporting democracy (see Claassen 2020, Appendix 1.3). This rule may or may not 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.

Another source of data-entry errors here involves survey weights. 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. These errors shifted the percentage of democracy-supporting responses in both directions, typically by relatively small amounts.

Finally, although not depicted on this plot, data-entry errors were also evident in the variable recording the year in which a survey was conducted: these typically reflected differences between the nominal year of a survey wave and when the survey was actually in the field in a particular country. This was an issue for some 9% of the country-year observations.

4 Discussion

The analysis above reveals that data-entry errors are an especially pernicious threat to the credibility of our results. The threat is a subtle one that is not easily detected. To discern it requires close scrutiny of every manual entry; merely examining the data and their distribution will uncover few errors (Barchard and Pace 2011, 1837–38). Although failure to find support for a research hypothesis may prompt us to undertake a such a close review, an analysis that yields statistical significance is unlikely to trigger what will likely be, as in the above example, a time-consuming and difficult effort (see Gelman and Loken 2014, 464). 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).

In making this recommendation, we are aware that being open and transparent in this way takes effort (?). But as researchers automate more of their data entry, the chances that they can reuse their code in subsequent projects improve. In fact, many common janitor-work chores already have been packaged as open-source software that to make researchers' task more straightforward and easier.[5] "Write code instead of working by hand," as Christensen, Freese, and Miguel (2019, 197) admonish, "don't use Microsoft Excel if it can be avoided."

Second, **use the double-entry method**: when manual data entry cannot be avoided, each entry should be made twice. Double entry is labor intensive, but experiments have shown that it reduces error rates by thirty-fold even when done immediately after the initial collection and by the same person (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.

Third, **embrace teamwork**: for any project involving entering data by hand, splitting the task up among team members will reduce the risk of errors going undetected. When double entries are performed by different people, discrepancies will be noted, discussed, and resolved correctly; having two sets of eyes on complex materials like survey codebooks also increases the chances that nuances of the presentation like survey weights will be uncovered. Further, by dividing the load, teamwork also lessens the probability of errors due to fatigue arising in the first place.

Data-entry errors are inevitable, and even following these recommendations is unlikely to eliminate them entirely. Further, the above suggestions follow closely from a specific case and, although they successfully help us identify and fix its data-entry issues, they do not constitute a panacea to cure all data-processing problems in all types of research.

Nonetheless, we also hope the readers to see the shared logic of these suggestions and the growing literature to guide political scientists to conduct more reliable and credible research. For instance, in the same vein as our first suggestion, (?) provides a book-

[5] For example, see `readtext` (Benoit et al. 2016) for formatting text files and `DCPOtools` (Solt, Hu, and Tai 2018) for aggregating cross-sectional time-series public-opinion surveys.

length set of illustrations on how to reduce "manual point and click" tasks found in a variety of studies with the `tidy`-data framework in the R language. (?) even suggests that qualitative researchers should consider using "open-exchange format" of qualitative data analysis software to be more "transparent about the generation and analysis of data." Furthermore, we regard our efforts and recommendations as a contribution to the open science movement to produce more robust and credible research in the social sciences (see, e.g., ?) and beyond (Lohr 2014; see, e.g., ?). With careful attention, not only can the threat of data-entry errors to our 'janitor work', our research, and our understanding of the world be minimized, but the transparency, openness, and credibility of our research can continuously grow.

Reference

Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman. 2018. "Countrycode: Convert Country Names and Country Codes." *Journal of Open Source Software* 3(28): 848–49.

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, Adam Obeng, Paul Nulty, Aki Matsuo, Kohei Watanabe, and Stefan Müller. 2016. "readtext: Import and Handling for Plain and Formatted Text Files."

Blondel, Emmanuel. 2018. "Rsdmx: Tools for Reading SDMX Data and Metadata."

Caughey, Devin, Tom O'Grady, and Christopher Warshaw. 2019. "Policy Ideology in European Mass Publics, 1981–2016." *American Political Science Review* 113(3): 674–93. doi:10.1017/S0003055419000157.

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. 2019. “Estimating Smooth Country–Year Panels of Public Opinion.” *Political Analysis* 27(1): 1–20. doi:10.1017/pan.2018.32.

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

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

Haegemans, Tom, Monique Snoeck, and Wilfried Lemahieu. 2019. “A Theoretical Framework to Improve the Quality of Manually Acquired Data.” *Information & Management* 56(1): 1–14.

Hagemann, Sara, Sara B. Hobolt, and Christopher Wratil. 2017. “Government Responsiveness in the European Union: Evidence from Council Voting.” *Comparative Political Studies* 50(6): 850–76.

Hu, Yue, and Frederick Solt. Forthcoming. “Macrointerest Across Countries.” *British Journal of Political Science.*

Kołczyńska, Marta, Paul-Christian Bürkner, Lauren Kennedy, and Aki Vehtari. 2024. “Modeling Public Opinion over Time and Space: Trust in State Institutions in Europe, 1989-2019.” *Survey Research Methods* 18(1): 1–19.

Lahti, Leo, Janne Huovari, Markus Kainu, and Przemysław Biecek. 2017. “Retrieval and Analysis of Eurostat Open Data with the eurostat Package.” *The R Journal* 9(1): 385–92.

Lohr, Steve. 2014. “For Data Scientists, ‘Janitor Work’ Is Hurdle to Insights.” *New York Times*: B4.

Lugo, Marco. 2017. “CANSIM2R: Directly Extracts Complete CANSIM Data Tables.”

Magnusson, Mans, Leo Lahti, and Love Hansson. 2014. "Pxweb: R Tools for PX-WEB API."

McGann, Anthony, Sabastian Dellepiane-Avellaneda, and John Bartle. 2019. "Parallel Lines? Policy Mood in a Plurinational Democracy." *Electoral Studies* 58: 48–57.

Mitra, Pradeep, and Ruslan Yemtsiv. 2006. "Increasing Inequality in Transition Economies: Is There More to Come?"

Norris, Pippa. 2011. *Democratic Deficit: Critical Citizens Revisited.* New York: Cambridge University Press.

Sepulveda, Mauricio Vargas. 2024. *Tabulapdf: Extract Tables from PDF Documents.*

Solt, Frederick. 2009. "Standardizing the World Income Inequality Database." *Social Science Quarterly* 90(2): 231–42.

Solt, Frederick. 2015. "On the Assessment and Use of Cross-National Income Inequality Datasets." *Journal of Economic Inequality* 13(4): 683–91.

Solt, Frederick. 2016. "The Standardized World Income Inequality Database." *Social Science Quarterly* 97(5): 1267–81.

Solt, Frederick. 2020a. "Measuring Income Inequality Across Countries and over Time: The Standardized World Income Inequality Database." *Social Science Quarterly* 101(3): 1183–99.

Solt, Frederick. 2020b. "Modeling Dynamic Comparative Public Opinion." doi:10.31235/osf.io/d5n9p.

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."

Wickham, Hadley. 2016. "Rvest: Easily Harvest (Scrape) Web Pages."

Wickham, Hadley, James Hester, and Jeroen Ooms. 2018. "Xml2: Parse XML."

Woo, Byung-Deuk, Lindsey Allemang, and Frederick Solt. 2023. "Public Gender Egalitarianism: A Dataset of Dynamic Comparative Public Opinion Toward Egalitarian Gender Roles in the Public Sphere." *British Journal of Political Science* 53(2): 766–75. doi:https://doi.org/10.1017/S0007123422000436.