Updating the Standardized World Income Inequality Database*

Frederick Solt *University of Iowa*

Objective:*Methods*:*Results*:*Conclusion*:

Keywords: income inequality, measurement

The SWIID Source Data

The SWIID source data consists of observations of the Gini coefficient in various countries and years.[1] The Gini is most intuitively defined as the average difference in income between all pairs in a population, divided by twice the average income in the population. It is by far the most commonly encountered summary statistic for measuring income inequality. The Gini has drawbacks—it is most sensitive to changes in the middle of the income distribution, rather than among those with the highest or lowest incomes, and it is not easily decomposed–that other metrics, such as the Atkinson index and the Theil indices, overcome. However, in light of the SWIID's goal of providing data for the broadest possible sample of countries and years, the ubiquity of the Gini makes it the only plausible choice for this purpose.

To be included in the SWIID source data, Gini observations need to encompass the entire population of a country without regard to age, location, or employment status.[2] They need to have an identifiable welfare definition and equivalence scale (more on these below). Finally, to ensure that these original sources are easily available to SWIID users, observations need to be available online, although not necessarily without paywalls.[3]

Hand-entering data is tedious and error-prone work, so I automated as much of the process of data collection as practicable. Most international organizations and a few national statistical offices use application programming interfaces (APIs) that facilitate downloading their data, and often the R community has built packages using these APIs to make the task even easier (see **?????**). I took as much advantage of these resources as possible. Although the sources with APIs were relatively few, they contained the most data: 55% of the observations were collected this way. In the absence of an API, I scripted downloads of any available spreadsheets (see **?**). If there was no spreadsheet, but data were available in pdf files, I automated downloading these files and then used the `tabulizer` package (**?**) to read the tables into R. In the absence of a file to download, I

*The paper's revision history and the materials needed to reproduce its analyses can be found on Github here. Corresponding author: frederick-solt@uiowa.edu. Current version: 20 February 2019.

[1]Because the Gini index is simply the Gini coefficient multiplied by 100, the two are equivalent, and both can be referred to as 'the Gini' without much cause for confusion.

[2]The requirement for complete territorial coverage was relaxed for minor deviations such as data on Portugal that excludes Madeira and the Azores. It was relaxed somewhat further for early series that covered only the urban population of three highly urbanized countries: Uruguay, Argentina, and South Korea. The general rule, however, is that data is excluded if it measures the income distribution of only urban or rural populations, or of only selected cities, or some other such incomplete territory. This requirement that the observation must not be restricted to only the employed is new; it means nearly 600 observations on the distribution of wages across employed individuals that were included in the source data of earlier versions of the SWIID are now excluded. Between the lack of information on those out of the workforce and on how workers formed households, these data were not very strongly related to LIS data on income inequality in the entire population anyway.

[3]For scholarly articles, DOIs or JSTOR stable URLs were the preferred web addresses, but if those were unavailable the publisher's website or another repository was used. For books, the link is to the relevant page in Google Books.

scripted the process of scraping the data from the web.[4]

Still, for a variety of reasons, a source's data may have been consigned to being entered in a separate spreadsheet.[5] 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 reproducibility particularly challenging in any event. Some sources, like books, cannot be read straight into R. All of the entries in this spreadsheet were checked repeatedly for errors, and I excluded repeated reports of the exact same observation from different sources.[6]

In the end, I was able to automate the collection of approximately three quarters of the source data and a much higher percentage of the series that will be updated or are subject to revision, facilitating incorporating these changes in future versions.

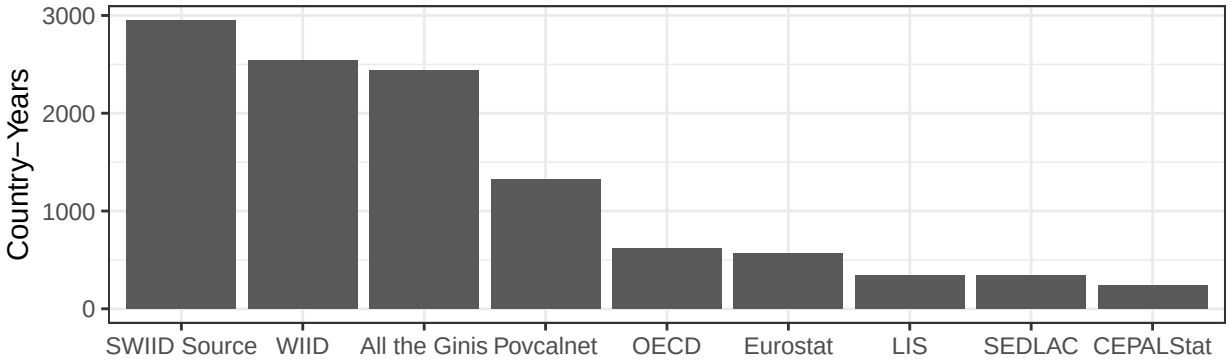

Figure 1: Income Inequality Datasets by Country-Years Covered

The resulting dataset comprises 15549 Gini coefficients from 2951 country-years in 196 countries or territories; as shown in Figure 1, this makes the coverage of the SWIID source data broader than that of any other income inequality dataset. This is not surprising given that, with the exceptions of the two other secondary collections–the World Income Inequality Database (**?**), which contains no original data and so is not drawn on at all, and the *All the Ginis* database (**?**), which contains very little original data and so is not drawn on much–the SWIID source data incorporates all of the data in these other datasets.

Turning from how the source data were collected to how they are composed reveals that there is much more data available about the income distribution in some countries than in others. Which countries are most data-rich? Figure 2 below shows the top dozen countries by the count of observations. Canada, by virtue of the excellent Statistics Canada as well as longstanding membership in the OECD and LIS, has 736 observations, many more than any other country. The United Kingdom and United States are next, followed by a group dominated by European countries but including Mexico and Taiwan. All of these data-rich countries are members of the LIS. On the other hand, eleven countries have only a single observation.

As discussed in the next section, observations for the same country in the same year, but with different welfare definitions and equivalence scales or from different sources, are important to generating the SWIID's cross-nationally comparable estimates. Still, we might be interested to know which countries have the most coverage of the years in the SWIID's current 57-year timeframe,

[4]Code for the entire process can be viewed here: https://github.com/fsolt/swiid/blob/master/R/data_setup.R.

[5]See https://github.com/fsolt/swiid/blob/master/data-raw/article_data/fs_added_data.csv.

[6]Which, of course, is not to say that these entries are error-free. If you spot any problems or know of sources I might have missed, *please* let me know at https://github.com/fsolt/swiid/issues/6.

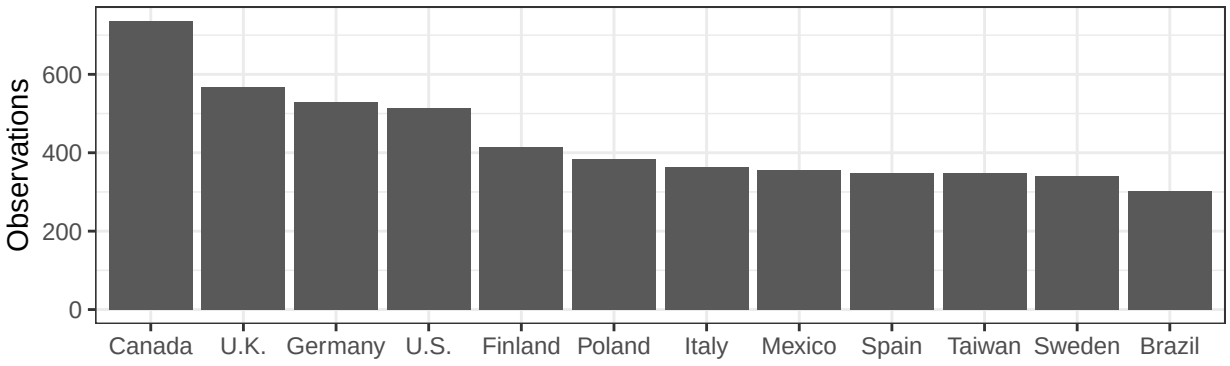

Figure 2: Countries with the Most Observations in the SWIID Source Data

from 1960 to 2017, because the SWIID's inequality estimates for countries with fewer country-year observations will include more interpolated values, which in turn will have more uncertainty. The countries with the most observed years are shown in the left panel of Figure 3. The source data includes observations in every covered year for the United Kingdom, in all but of these years for Sweden, and in all but four for the United States. Argentina—a country that is not a member of the LIS—makes the top ten, with 42 country-year observations. The median country has observations in just nine different country-years.

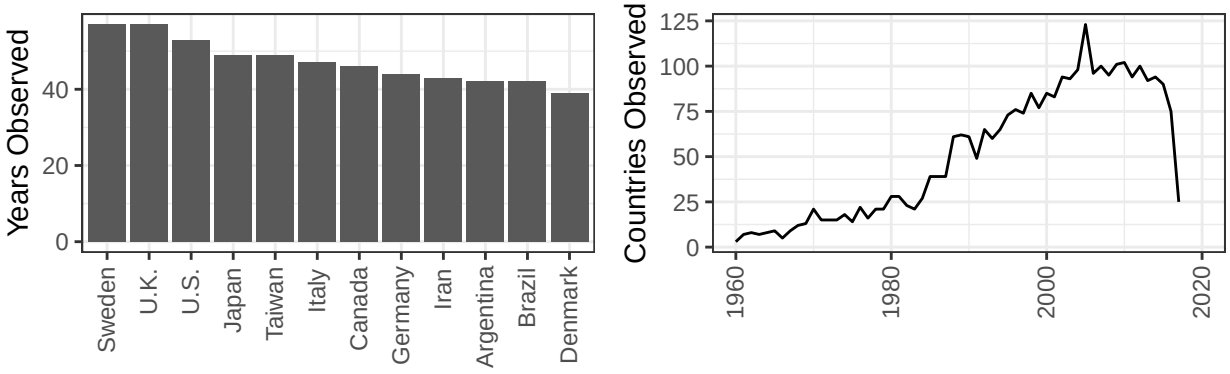

Figure 3: Country-Year Coverage in the SWIID Source Data

We can also get a sense of the available inequality data by turning the question around and asking about coverage of countries across time. The right panel of Figure 3 shows, for each year, the number of countries for which the SWIID source data includes at least one observation. There are observations for 123 countries in 2005, the year with the broadest coverage. Coverage is relatively good in all of the years from 2000 to 2015, with at least 80 countries observed per year, before dropping to 75 countries for 2016 and just 25 for 2017. Before then, country coverage is pretty thin each year through the 1960s and 1970s and still is not great until the late 1980s.[7]

Above I mentioned that to be included in the SWIID source data observations need to have an identifiable welfare definition and equivalence scale; let us now consider these two aspects of

[7]This is partly a result of the decision to insist on sources that are available online, but it's just as well: so little information is available about many of the so-excluded observations on that era that it is hard to have much confidence in them.

the data. A welfare definition is an answer to the question, this Gini measures the distribution of what? The four welfare definitions employed in the SWIID source data are market income, gross income, disposable income, and consumption. Market income is defined as the amount of money coming into the household, excluding any government cash or near-cash benefits, the so-called 'pre-tax, pre-transfer' income.[8] Gross income is the sum of market income and government transfer payments; it is 'pre-tax, post-transfer' income. Disposable income, in turn, is gross income minus direct taxes: 'post-tax, post-transfer' income.[9] Consumption does not refer to the money coming into the household at all but rather to the money going out.[10] As can be seen in the left panel of Figure 4, in the SWIID source data, Ginis of disposable income are much more common than those using other welfare definitions.

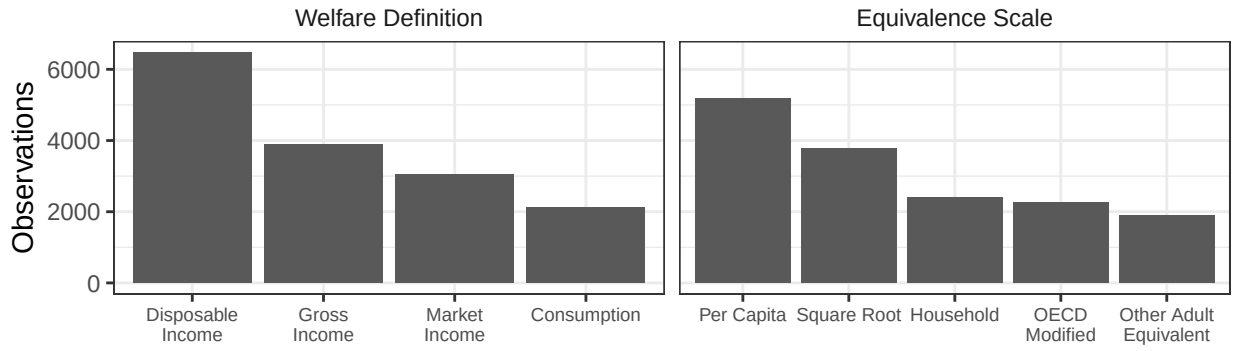

Figure 4: Welfare Definitions and Equivalence Scales in the SWIID Source Data

Equivalence scales are the ways in which the size and composition of a household is incorporated into the calculation of its members' welfare. On the one hand, these factors can simply be ignored, with all households with the same amount of income or consumption treated as if they enjoy the same level of welfare, regardless of their size. One can improve on this household 'scale'[11] by dividing the household's income by its number of members, that is, by using a per capita scale. Although undoubtedly superior to simply looking at households, the per capita scale is not ideal: a household of two members and an income of $100,000 is better off than one with a single member

[8]It's important, though, to not think of the distribution of market income as 'pre-government.' Beyond taxes and transfers, governments seeking to shape the distribution of income have a wide array of 'market-conditioning' or 'predistribution' policy options, with minimum wage regulation and labor policy two obvious examples (see, e.g., **?**). Moreover, even taxes and transfers can profoundly shape the distribution of market income through 'second-order effects.' Where robust public pension programs exist, for example, people save less for retirement, leaving many of the elderly without market income in old age and so raising the level of market-income inequality (see, e.g., **?**).

[9]Note that disposable income still does not take into account, on the one hand, indirect taxes like sales or value-added taxes, or, on the other, public services and indirect government transfers such as price subsidies. There is very little information available about the distribution of such 'final income,' pretty much only that generated by the Commitment to Equity Institute, so I exclude it from the SWIID source data at least for the time being.

[10]In previous versions of the SWIID, market and gross income were treated as a single welfare definition, and I am glad to finally be able to split them apart (cf. **?**, 1272). The consumption welfare definition might now be the most heterogeneous within the SWIID source data, varying considerably in whether and how observations treat expenditures on durable goods. Another source of differences within a single welfare definition is the extent to which nonmonetary income—such as the value of food grown for the household's own consumption or of housing that the owner occupies—is included. The SWIID source data include the variable `monetary` that indicates whether any nonmonetary income is taken into account, but at present this information is not incorporated into the classification of welfare definitions.

[11]Quoted because, strictly speaking, nothing is being scaled at all; it's simply treating the household as the unit of analysis.

and $50,000 due to economies of scale—that is the most important reason why people look for roommates. There are a variety of ways to try to account for these economies by calculating the number of 'equivalent adults' in the household. Of the most commonly used adult-equivalent scales, the square-root scale is the most straightforward: one simply divides the household income by the square root of the number of members. The 'OECD-modified' scale for the number of adult equivalents (which the OECD itself actually does not use, preferring the square-root scale) counts the first adult as 1, all other adults as .5, and each child as .3. And there are many other adult-equivalent scales, from the 'old OECD' scale (1 for the first adult, 0.7 for each additional adult, and 0.5 for each child) to caloric-requirement-based scales (which are in fact very nearly per capita, as it turns out) to a number of country-specific scales. In previous versions of the SWIID, all adult-equivalent scales were considered a single category. Now, the square-root scale and the OECD-modified scale have both been split out, leaving the remaining catch-all adult-equivalent category much smaller. The right panel of Figure 4 shows how many observations in the SWIID source data use each equivalence scale.

Differences in the welfare definition and the equivalence scale employed constitute the biggest source of incomparability across observations in the source data, and all twenty of the possible combinations are represented.

Standardizing the Data