The Standardized World Income Inequality Database

Frederick Solt
frederick-solt@uiowa.edu

October 2, 2014

Abstract

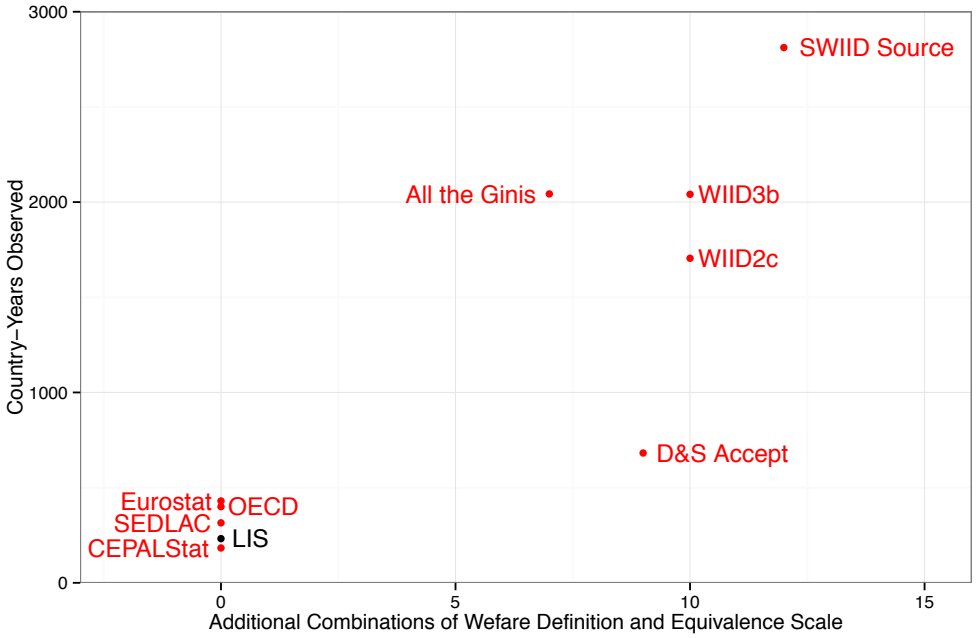
Objective. Investigating the causes and consequences of income inequality requires comparable data, but greater cross-national and temporal coverage is generally available only with sharply reduced comparability. The Standardized World Income Inequality Database (SWIID) provide researchers with data that maximize comparability for the broadest possible sample of countries and years. *Methods.* The SWIID employs a custom missing-data algorithm that minimizes reliance on problematic assumptions by using as much information as possible from proximate years within the same country to estimate inequality statistics for the missing country-years in the Luxembourg Income Study using data drawn from regional collections, national statistical offices, and academic studies. *Results.* The SWIID provides comparable estimates of the Gini index of net- and market-income inequality for 174 countries for as many years as possible from 1960 to the present, as well as measures of absolute and relative redistribution. *Conclusions.* As its coverage and comparability far exceed those of alternate datasets, the SWIID is better suited for broadly cross-national research on income inequality than other sources.

Valid comparisons of levels and trends in income inequality across countries and over time depend on comparable data. There is a great deal of data on inequality available for cross-national and over-time analyses, but unfortunately most of these data are simply not comparable due to differences in the population covered, in terms of geography, age, and employment status; the welfare definition employed, such as market income or consumption; the equivalence scale applied, such as household per capita or household adult equivalent; and the treatment of various other items, such as non-monetary income and imputed rents. The Standardized World Income Inequality Database (SWIID) was introduced in 2008 to provide researchers with income inequality data that maximize comparability for the broadest possible sample of countries and years (see Solt 2009). While it retains that same goal, the SWIID has evolved and expanded considerably since that time. This article reviews the problem of comparability in cross-national income inequality data, explains how the SWIID addresses the issue, assesses the SWIID's performance in comparison to the available alternatives, and explains how to use the SWIID data in cross-national analyses.

The Problem of Comparability

The tradeoff between coverage and comparability is readily evident in Figure 1, which graphs the number of country-year observations of the Gini index provided by different cross-national datasets against the number of different combinations of welfare definition and income scale used in calculating these statistics. The Luxembourg Income Study is the only source that provides inequality statistics calculated using a uniform set of assumptions and definitions on the basis of microdata that has been painstakingly harmonized to maximize its comparability. Its reputation as the gold standard of cross-nationally comparable inequality data is well deserved, but at present it provides only 232 country-years of data in 41 countries. More observations are available from other sources at the cost of sacrificing the benefits to comparability of harmonization. Eurostat, for example, provides 430 country-years (in 32 European countries) of inequality in household disposable income per adult equivalent. Even more observations are available from sources composed of observations calculated using different welfare definitions or income scales. The source data used to generate the SWIID, described in more detail below, draws on all of the sources depicted here as well as national statisti-

Figure 1: The Tradeoff Between Comparability and Coverage



Note: As underscored with contrasting color, only the LIS data is harmonized. D&S Accept refers to data presented in Deininger and Squire (1996); All the Ginis to data presented in Milanovic (2013); SEDLAC to CEDLAS (2013); WIID2c to UNU-WIDER (2008); and WIID3b to UNU-WIDER (2014). Other sources were accessed October 1, 2014.

cal offices and the scholarly literature. It now comprises more than 10,000 Gini indices in over 2800 country-years in 174 countries, but these Ginis are calculated on the basis of eleven different combinations of welfare definition and income scale.

This tradeoff suggests researchers have two principal options, neither very attractive. The first is to maximize comparability. This can be done either by insisting on only the harmonized data of the LIS or data from some other single source or, at some sacrifice, by using data generated using only a single basis of calculation. Either way, privileging comparability entails giving up on making many comparisons and throwing away most of the available information. This is true even for fairly recent years in the most data-rich part of

the world, Europe: depending on only Eurostat data to provide information about the context of inequality in which the first four waves of the European Social Survey were conducted (2002-2009) would result in missing data for more than one-fourth of the country-years in the sample (see Solt 2015). Even the least stringent approach, combining data sources that use a single combination of welfare definition and equivalence scale, yields just 1128 country-year observations of disposable-income per-adult-equivalent Ginis in the SWIID source data, only about 40% of the total country-years available.[1]

The second is to increase coverage by using more of the available data and making a global fixed adjustments to account for the average differences between statistics based on different calculations. (An often employed but even less defensible variant, of course, is to simply ignore the incomparability altogether; for a recent example, see Halter, Oechslin, and Zweimüller (2014).) Milanovic (2013, 8) stressed the incomparability of the observations included in his *All the Ginis* dataset and included a series of dummy variables indicating whether a particular observation was based on a welfare definition of gross income, net income, or expenditure and whether it was calculated using the unadjusted household or household per capita income as the equivalence scale. In line with the recommendations of Deininger and Squire (1996, 582), he advised using these dummies to make adjustments for each of these characteristics.

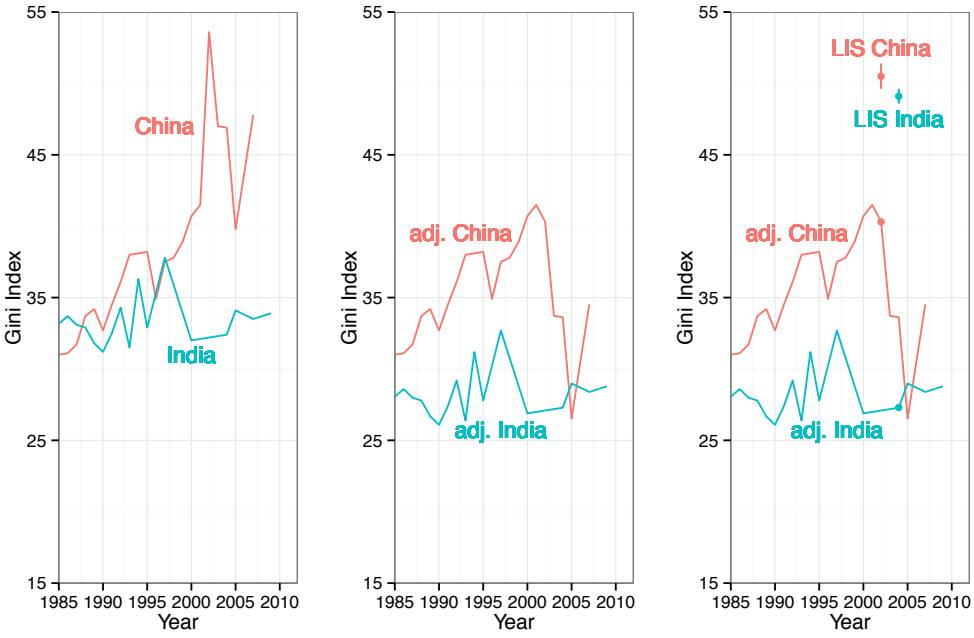
Although the recommendation to calculate such global fixed adjustments is straightforward to implement, it does not satisfactorily deal with the incomparability of these inequality statistics. The difference between inequality across household incomes and across household incomes per capita will depend on the relationship between income and household size; to the extent this relationship varies across countries and over time, a fixed adjustment will underestimate inequality in some country-years and overestimate it in others. Similarly, the difference between inequality in net income and in expenditures depends on patterns of savings and consumption across households; these patterns are known to be different across countries (see, e.g., Kirsanova and Sefton 2007). The difference between gross- and net-income inequality reflects the progressivity of the tax code and any variation in compliance by

[1] A third option, abandoning the Gini index and similar summaries of the entire income distribution and adopting instead a different conceptualization of income inequality that may allow for more and more comparable data to be brought to bear, has been taken by two prominent efforts, the University of Texas Inequality Project and the World Top Incomes Database. These efforts will be discussed further below.

income, both of which are also well understood to vary considerably across countries and years (see, e.g., Förster, Llena-Nozal, and Nafilyan 2014).

An example can best illustrate the severe limitations of combining observations calculated on different bases using only global fixed adjustments. Consider the question of whether income inequality is higher in China or India posed by Mukhopadhaya, Shantakumar, and Rao (2011, 102). The left panel of Figure 2 shows unadjusted data for the two countries since 1985 from the `Giniall` series of Milanovic's 2013 *All the Ginis* dataset. It suggests that inequality in both countries was similarly moderate until the mid-1990s, when levels in China increased dramatically and became much higher than those in India.

Figure 2: Inequality in China and India



Notes: The left panel depicts the unadjusted `Giniall` series (Milanovic 2013). The center panel applies global fixed adjustments to these data to account for differences in their bases of calculation. The right panel compares the adjusted series with data from the Luxembourg Income Study. Even with adjustments, the `Giniall` series yield very different conclusions than the comparable data provided by the LIS.

`Giniall` is the preferred series in the dataset, but even in this limited subsample it includes observations of inequality in gross income per capita, net income per capita, and expenditures per capita, so "an adjustment for each of these characteristics is desirable" (Milanovic 2013, 8). A linear regression of `Giniall` on the included dummies for consumption (versus income), household (versus per capita), and gross income (versus net income) across the entire dataset indicates that to make a consistent series of net-income per capita inequality, one should subtract 5.1 points from all consumption-based observations, add 2.4 points to all household-based observations, and subtract 13.3 points from all gross-income-based observations. The result is presented in the center panel of Figure 2. According to the adjusted data, inequality in China has been consistently (excepting only a single, quickly reversed dip in 2005) and considerably (by an average of about 7 points) higher than in India.

But actually comparable data reveals a very different picture. The right panel of Figure 2 adds the single data point available from the LIS for each country: from 2002 in the case of China, 2004 for India. In the adjusted data, the difference between these two country-years is 13 points. In the LIS, the difference is just 1.5 ± 1.0 points; inequality in China is distinguishably but only very slightly higher than in India for these two observations. The fixed adjustments are better than no adjustments in this case—in the unadjusted data depicted in the left panel this difference is over 21 points—but they still leave a great deal to be desired.

To make a single, global fixed adjustment in an attempt to account for differences in inequality statistics calculated on two dissimilar bases is to assume that these differences are always and everywhere the same, an assumption that is of course false. As shown in the next section, the SWIID takes advantage of an abundance of available income inequality data to relax this assumption and so maximize comparability.

Standardizing the Available Data

The starting point for the SWIID is two collections of Gini indices: the LIS data and the source data. The LIS data consists of two series, one of net (that is, post-tax, post-transfer) income inequality and one of market (pre-tax, pre-transfer) income inequality. The net-income inequality series is taken directly from the LIS Key Figures (LIS 2014*a*). The market income

inequality series is generated from the LIS microdata (LIS 2014*b*).[2] The quality and comparability of these LIS data are unparalleled, but, as noted above, their shortcoming for broadly cross-national work is coverage. From 1967, the year of the first LIS observation, until 2012, the year of the most recent, the LIS includes just 231 observations from 41 countries: there are twelve observations for Canada over these 45 years; seven other countries have only one.[3]

The source data has the reverse set of strengths and weaknesses. Although in the earliest versions of the SWIID it consisted of only the UNU-WIDER (2008) database, the source data has since expanded with each revision and now encompasses data provided by all of the major cross-national inequality databases, the national statistical offices of countries around the world, and dozens of scholarly articles. (The source data, annotated with the original sources, can be found in the SWIID replication materials available online.) As of version 5.0, the source data includes over ten thousand Gini indices, dating from 1960 to 2013.[4] Observations are only included in the source data if they are based on all or nearly all of the country's population and if there is

[2]The code employed to generate the market income inequality series is available in the online SWIID replication materials; I thank Tomas Hellebrandt of the Peterson Institute for International Economics for his valuable suggestions regarding the market-income inequality series in the LIS.

[3]I make two modifications to the coverage of the two LIS series. First, New Zealand's privacy laws have thus far prevented the country from contributing microdata to the LIS project; I treat as LIS data four observations of net-income inequality (1982, 1986, 1991, 1996) specially prepared by Statistics New Zealand (1999, 73) to be comparable with the LIS Key Figures. Second, the LIS had originally hosted data on Russia in 1992, 1995, 2000 from the Russian Academy of Sciences, but it apparently lost permission to use these data sometime in late 2011 or early 2012. This series was replaced in the LIS with data from the Russian Longitudinal Monitoring Survey for 2000, 2004, 2007, and 2010. The RLMS data, however, show a sharp decline in income inequality after 2004 (a fall of over 12% by 2010) that is not reflected in any other source I have located. I retain the original 1992 and 1995 RAS-based figures in the LIS series, but because the RLMS-based observations for Russia after 2004 seem to lack face validity, I reluctantly omit these two observations. I thank Louis Chauvel for our conversation on this second topic.

[4]That is, the source data includes about twice the number of Ginis as the entire revision 2c UNU-WIDER database (which includes many observations calculated on less than the entire population or without information regarding the welfare definition or equivalence scale employed), and more than 50% more than in the whole of the recently released and partially undocumented revision 3b of that source (UNU-WIDER 2014). Careful checking confirmed that the new, preliminary additions to these data include few useable Ginis not in the SWIID source data.

sufficient information to identify the equivalence scale and welfare definition employed in their calculation. Even with these restrictions, however, the differences in the way these statistics are calculated render them ill-suited for making direct comparisons.

The SWIID uses the two LIS series as baselines to which the source data are standardized. More precisely, the source data are used to generate model-based multiple imputation estimates of the many missing observations in the LIS series (on model-based multiple imputation, see Gelman and Hill 2007). The process begins by sorting the source data into eleven categories defined by the combination of welfare definition and equivalence scale used in their calculation.[5] Observations in the source data are classified as using one of three different welfare definitions: (1) net income, (2) market income, or (3) expenditure.[6] The source data are also classified by equivalence scale: (1) household per capita, (2) household adult equivalent, (3) household unadjusted, or (4) person.[7] As the 'person' equivalence scale is used only with information on the distribution of (pre- or post-tax) wage income, there are no observations with the expenditure-person combination of welfare definition and equivalence scale. The remaining combinations leave the aforementioned eleven categories. As the standard, the two series of LIS data, which are calculated on the basis of household adult equivalent (using the square root scale) for net and market income respectively, are treated as their own separate categories, bringing the total number of categories to thirteen. Rather than choose among sources, when more than one observation is available within a category for a particular country and year, these

[5]Earlier versions of the SWIID classified the source data using nineteen categories devised by Babones and Alvarez-Rivadulla (2007). I am grateful to Stephen P. Jenkins for pointing out that these categories were not entirely coherent. The old categories also included several for observations for which the welfare definition was unknown; as noted above, such observations are now excluded from the source data.

[6]Observations calculated on the basis of pre-tax, post-transfer gross income are omitted from the source data when market-income series are available for the same country; otherwise, they are at present classified as market income. I plan to split these observations out into their own separate classification when sufficient data become available.

[7]Several different definitions of 'household adult equivalent' appear in the source data, including the square root of household size (the definition used in the LIS Key Figures), the OECD scale, and several country-specific scales. The differences in the Gini indices based on these different definitions of adult equivalent, however, are typically very small, less than one point on the 0–100 scale. For this reason, I opt at present to treat them as a single group to facilitate the standardization process, although at the cost of slightly greater uncertainty.

observations are averaged.

The result of the categorization, then, is a dataset of country-year observations, each of which has data on inequality in one or more of the thirteen categories. What is needed to generate a series with data on all countries and years from the incomplete inequality variables in thirteen categories are the ratios between each pair of variables. If the ratio ρ_{ab} between the Gini index data in categories a and b were known, missing observations in a could be replaced simply by multiplying available data in b by ρ_{ab}. But as noted previously, the relationship between Gini indices with different reference units and income definitions will vary considerably from country to country and also over time depending on the extent of redistributive policies, details of tax law, patterns of consumption and savings, family structure, and other factors. In other words, ρ_{ab} is not constant but varies across countries i and years t. Further, ρ_{abit} is only directly calculable for those pairs of categories in those countries and years for which it is not immediately useful, that is, only when data is already available in both categories for that observation.

Those ratios ρ_{abit} that are directly calculable are valuable nevertheless because they provide information about what the ratios that are missing are likely to be. Because the factors that affect these ratios—redistributive policies, patterns of consumption, and so on—tend to change only slowly over time within a given country, the best prediction for a missing ratio will be based on available data on the same ratio in the same country in proximate years, thereby minimizing any differences in these factors. With this in mind, the ratios ρ_{abit} are predicted from the results of a series of models.

First, in those countries with sufficient data, predictions are generated by loess regression, which incorporates the maximum amount of information from proximate years by fitting a smooth curve point-by-point through the available data. Next, predictions are generated through a series of regression models. In order of increasing availability—but also increasing uncertainty as reflected in larger standard errors—$\hat{\rho}_{abit}$ was predicted as a function of (1) country-decade, (2) country, (3) region, and (4) advanced or developing world.[8] The predictions of all of these models are then combined for each ratio ρ_{abit}, assigning each country-year the available prediction with the smallest standard error.

[8]The earliest versions of the SWIID also predicted $\hat{\rho}_{abit}$ as a function of region-decades, but these models were discovered to not contribute significantly to the final estimates and were therefore dropped for the sake of simplicity.

These predictions $\hat{\rho}_{abit}$ alone, however, do not take advantage of all of the information available in the source data. An additional prediction of each conversion factor can be generated in a two-step process through other categories of data. That is, the ratio of the LIS net-income data (labeled category 1) to the data in category b can be calculated as the product of the ratio between data in category a and category b and the ratio of the LIS net-income data to data in category b: $\hat{\rho}_{1bit} = \hat{\rho}_{abit} \times \hat{\rho}_{1ait}$.

These two-step predictions improve upon the conversion factors predicted in one step in two ways. First, for some combinations of a and b, few or no observations of both categories of the Gini index are available, making modeling $\hat{\rho}_{abit}$ in one step impractical or impossible. Second, the uncertainty in the predicted conversion factor can often be reduced by averaging the one-step prediction with one or more two-step predictions.

Once all of the predicted ratios $\hat{\rho}_{1bit}$ are calculated, eleven series of estimates comparable with the LIS net-income series are gained by multiplying these predicted ratios by the available data in each of the eleven source-data categories. Because each of these comparable series is incomplete, they are combined into a single variable by assigning each observation with the estimate with the least uncertainty or, when the average of some or all of the available estimates yields an even smaller standard error, this average.

A final piece of information about the income inequality in a particular country and year is gained by noting that the distribution of income within a country typically changes only slowly over time: contemporary levels of inequality should generally be very similar to levels observed in the preceding year. With two exceptions discussed below, dramatic differences in the estimates of inequality for a given year and those preceding and following it likely reflect persisting errors in measurement. Allowing observations to be informed by the estimates for surrounding years works to minimize such errors. This is achieved by using the following five-year weighted moving average algorithm: $G_{it} = \frac{1}{6} \times (G_{it-2} + G_{it-1} + (2 \times G_{it}) + G_{it+1} + G_{it+2})$.

The first exception to the foregoing regards the Luxembourg Income Study data. Because of the very high quality of the LIS data, differences from one year to the next are unlikely to be caused by persistent measurement error, so observations from this source are therefore not adjusted with the moving average algorithm: all LIS observations are retained without change in the SWIID.[9] The second exception involves the countries of east-

[9]The LIS data still has its own measurement error, however, averaging 0.38 points in

ern Europe and the former Soviet Union during the collapse of communist rule. The sharp increases in inequality observed in most of these countries from 1990 to 1991 would appear to be due to the profound restructuring of these countries' societies and economies rather than measurement error. Applying the moving average algorithm to this region results in overestimates of inequality in 1989 and 1990 and underestimates in 1991 and 1992; therefore the algorithm is not used in these countries during these years.

Simply applying the moving-average algorithm to the net-income inequality variable, however, would lose the estimates of uncertainty associated with each observation. Therefore, the variable is re-generated one thousand times through Monte Carlo simulation and the moving-average algorithm applied to each simulation.

The foregoing steps yield estimates of LIS-compatible Gini indices of the distribution of net income in all country-years for which there is a Gini index in at least one of the eleven categories of source data. To generate estimates for additional observations, information on inequality across countries and over time from two other sources, the University of Texas Inequality Project (UTIP) and the World Top Incomes Database (WTID), is incorporated. Both of these sources seek to address the tradeoff between coverage and comparability by employing a single consistent, though more limited, conceptualization of income inequality: the UTIP measures differences in average pay between industrial classifications as reported by the United Nations Industrial Development Organization (Galbraith 2009), while the WTID consists of the share of taxable income reported on various top fractiles of personal tax returns (Atkinson, Piketty, and Saez 2011). This approach has the advantage of allowing many country-years to be observed. The WTID includes nearly one thousand country-years since 1960; the UTIP over four thousand. Each of these sources, however, has distinct drawbacks for making comparisons across countries in the level of income inequality. For the UTIP data, comparability is compromised by differences across countries in the share of all employment that is in agriculture or services rather than industry, the share of earnings differences that occurs within rather than between industrial classifications, and the share of income accruing to capital rather than labor. For the WTID, the problems include substantial differences across

the net-income series and 0.46 points in the market-income series. The standard errors of the LIS series were calculated by bootstrap using the LIS microdata; the code employed is available in the SWIID replication materials online.

countries in the definitions of taxable income and the tax unit as well as in the prevalence of tax avoidance and evasion across incomes; these differences led the dataset's compilers to use it only to compare the trends over time across countries, not levels (see Atkinson, Piketty, and Saez 2011, 4-5 and passim).[10] All of these differences, however, should change either slowly (e.g., the size of the industrial workforce) or relatively rarely (e.g., significant reforms to the tax code) making these sources appropriate for comparisons within a given country over time. On these grounds, the relationship within each country between the LIS-compatible Gini indices already estimated and any available UTIP or WTID data over time is estimated using loess regression.[11] These analyses are then used to predict a LIS-compatible estimate of the Gini index with standard error for each of those observations with UTIP or WTID data but without information in the source data; for observations with data in both the UTIP and the WTID, the estimate with the smallest standard error is used. Monte Carlo simulation is again used to generate one thousand simulated series drawn from a distribution with mean equal to the point estimate for each country-year and with a standard deviation equal to the standard error for that estimate.

In the final step in generating the SWIID's net-income inequality series, values for all post-1975 country-years still without estimates but between country-years with estimates were then interpolated for each simulation. The entire process was then repeated to generate a series standardized on the LIS household-adult-equivalent market-income data. Measures of absolute redistribution (the difference between the market-income and net-income Gini indices) and of relative redistribution (this difference divided by the market-income Gini and multiplied by 100, that is, the percentage by which market-income inequality is reduced) are calculated. Observations for these measures of redistribution are omitted for countries for which the source data do not include more than three observations of either market- or net-income inequality.[12] These four series of estimates—of net-income inequality, market-income

[10] I am grateful to Facuno Alvaredo for conversations underscoring how poorly cross-national differences in the WTID data correspond to cross-national differences in inequality measures from other sources.

[11] Due to differences in the scale and dispersion of these variables, all are logarithmically transformed in these analyses. To capture the uncertainty in the LIS-compatible Ginis, the analyses are repeated using ten Monte Carlo simulations and the results combined.

[12] In such cases, although the two inequality series each still constitute the most comparable available estimates, the difference between them reflects only information from

inequality, absolute redistribution, and relative redistribution—together constitute the SWIID.

Assessing the SWIID

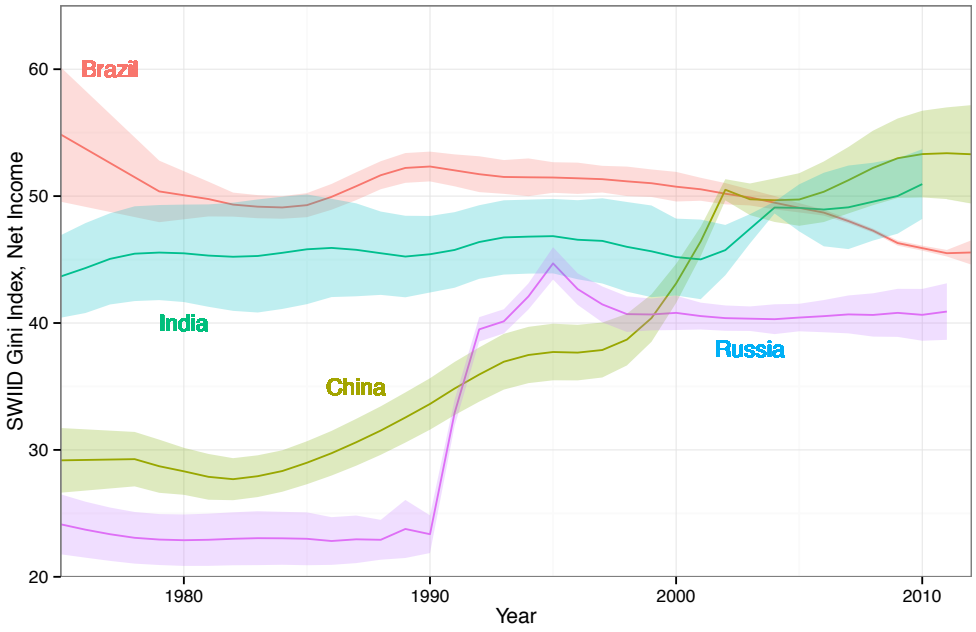
Version 5.0 of the SWIID dataset covers 174 countries, with estimates of net-income inequality comparable with the LIS Key Figures for 4631 country-years and estimates of market income inequality comparable with those obtained from the LIS for 4629 country-years. The SWIID's aim, as noted above, is to provide data for the broadest possible sample of countries and years that are made as comparable as feasible. On the first criterion, breadth of coverage, the SWIID bests all other inequality datasets. In fact, it more than doubles the country-year observations in Milanovic's (2013) *All the Ginis*, the next-largest income inequality dataset, and it is more than ten times the size of the Eurostat data, the largest collection calculated on the basis of a single welfare definition and equivalence scale.

This breadth of coverage allows researchers to make comparisons of countries around the world. As shown in Figure 3, the LIS-comparable data on China and India provided by the SWIID tells a very different story than the left and center panels of Figure 2. It reveals that estimated inequality in net incomes had been discernably higher in India than in China until the late 1990s, and with the exception of a few years early in the new millenium when inequality in China was briefly distinctively higher, the difference between the two countries' levels of inequality has not been large or clear since. This figure further compares India and China with the other two BRICs, Brazil and Russia. Notably, it shows that the recent downward trend in inequality in Brazil—long thought one of the most unequal countries in the world—has left that country with a more equal distribution of net income than either.

Of course, much hinges on the claim that the SWIID data are, in fact, comparable to the LIS. A preliminary test concerns the extent to which the SWIID avoids relying on the dubious assumption underlying fixed adjustments that differences in inequality statistics calculated on two dissimilar bases are constant across space and time. Figure 4 displays for each region the share of each type of adjustment used to calculate the SWIID estimates of net-income inequality from the source data. First, it is important

other countries, and treating it as meaningful independent information about redistribution cannot be justified.

Figure 3: Net-Income Inequality in the BRICs Countries, SWIID v5.0

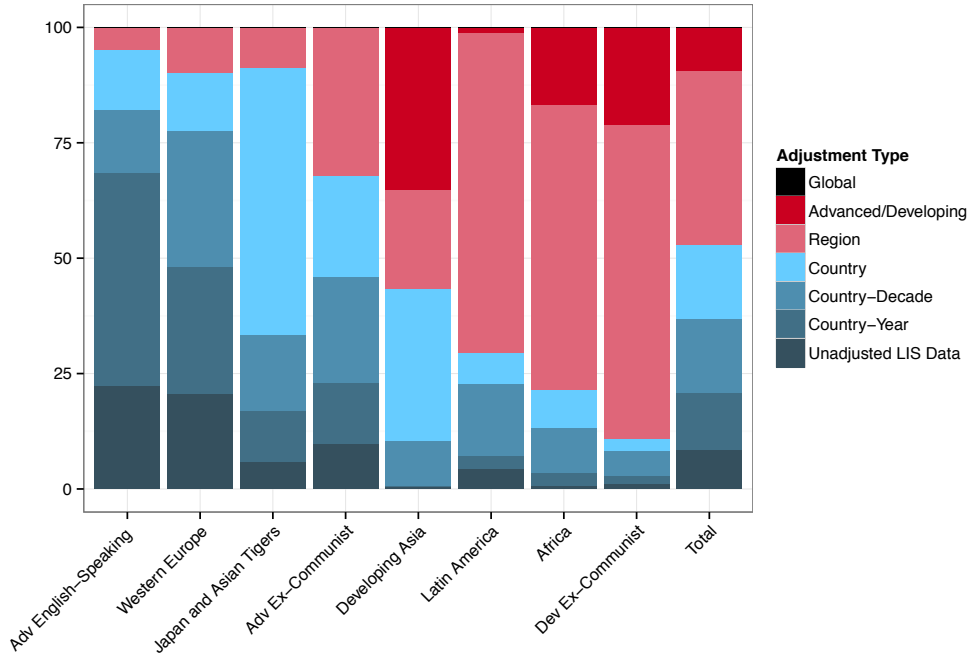


Notes: Solid lines indicate the mean estimate of net-income inequality across household adult-equivalents; shaded regions indicate the 95% confidence interval of these estimates.

to recognize that the SWIID completely avoids global fixed adjustments of the sort recommended by, e.g, Deininger and Squire (1996, 582). At worst, some SWIID estimates for poorer countries are based on relationships observed elsewhere in the developing world, but never on those seen in advanced countries (and for advanced countries, no estimates rely on information from countries outside their region). This alone constitutes a considerable advance over common practice.

The figure further shows that, where the available inequality data allows, the SWIID estimates are based on much more nuanced—and concomitantly much more realistic—adjustments. In the advanced English-speaking countries, for example, nearly seven in every ten estimates are taken directly from the LIS or are based on country-year-varying predicted ratios from loess regressions, as are about half of the estimates for the countries of western

Figure 4: SWIID Adjustment Types by Region



Europe. Only about 5% and 10%, respectively, of the estimates in these regions are based on ratios observed in other countries. A similar share of the estimates for Japan and the four Asian Tigers (Hong Kong, South Korea, Taiwan, and Singapore) are likewise based on at least within-country ratios, although the relatively small number of country-years in the LIS limits the extent to which these relationships were modeled as varying over time. Among the world's richer countries, only in ex-Communist central and eastern Europe are a substantial fraction of estimates based on regional averages rather than on information from within the country itself.

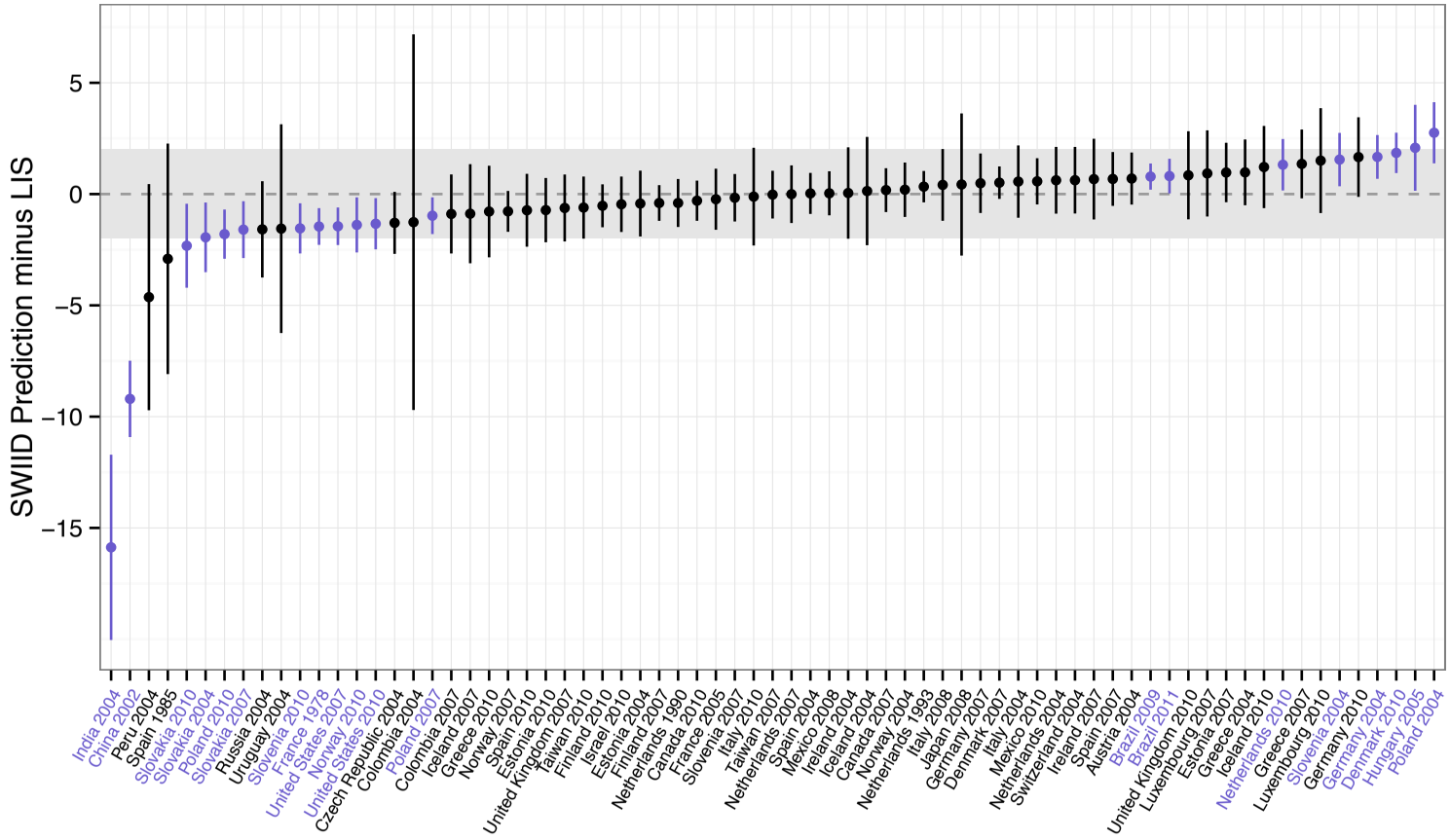
In the developing world, on the other hand, most SWIID estimates are based on ratios observed in other countries. This is a disappointing limitation imposed by the relative paucity of data available for developing countries. Even here, though, a substantial minority of estimates use within-country ratios, and most others rely only on the averages of regional neighbors. And

as the LIS expands to cover more of the developing world—it has announced that it will soon release data for four more Latin American countries as well as Egypt and Serbia—the SWIID will continue to improve in this respect. Overall, as depicted in the rightmost column of the figure, more than half of SWIID estimates are based on same-country information on the relationships between inequality statistics calculated on different combinations of welfare definition and equivalence scale.

The expansion of the LIS since the release of the SWIID provides an even more exacting test. By adopting the LIS as its standard, the SWIID means to provide estimates of what the LIS would show if a given country-year were in fact included in the LIS. Since 2008, the LIS has added data on 71 country-years that had been already included in the SWIID. If the SWIID succeeds in providing estimates of the highly comparable LIS figures, the differences between what the SWIID first predicted and the data that the LIS later released will not be substantively and statistically significant.[13]

[13]For discussions that inspired this test, I thank participants in the Expert Group Meeting on Reducing Inequalities in the Context of Sustainable Development, Department of Economic and Social Affairs, United Nations, New York, October 24-25, 2013.

Figure 5: SWIID Predictions of the LIS



Notes: The points represent the differences between values of the LIS net-income Gini index and the earlier SWIID estimates for these same observations, and the whiskers trace the associated 95% confidence intervals of these differences. The area within ±2 Gini points of zero is shaded gray; differences within this interval are not considered substantively significant. Differences that do not reach statistical significance are depicted in black; statistically significant differences are blue. For only 5 of 71 (7%) country-years are the differences between the predictions in the SWIID and the actual LIS values both substantively and statistically significant.

Figure 5 plots these differences, arranged with the SWIID underestimates of the LIS on the left and overestimates on the right. A first reassuring observation is that there is no overall tendency in either direction: the median difference is −0.02 Gini-index points. Differences larger than two points are considered substantively significant; this level is admittedly but necessarily arbitrary. In only 5 of these 71 country-years—that is, just 7%—are the differences between the LIS and the earlier SWIID predictions both substantively and statistically significant. This is an impressive record of out-of-sample prediction, and it lends considerable confidence that the SWIID has been providing data for a broad sample of countries and years that are comparable to the LIS, and so in turn across space and time.

The SWIID's record in predicting the LIS is all the more impressive in comparison to that of other cross-national income inequality datasets. Recall the two options implied by the tradeoff between comparability and coverage discussed above. The first option privileges comparability at the cost of coverage by using only those data generated using the same basis of calculation. The recently released, though as yet not fully documented, Version 3b of the UNU-WIDER (2014) dataset includes 951 country-years with Gini indices calculated on the basis of household-adult-equivalent disposable income, the same combination of equivalence scale and welfare definition as employed to calculate the LIS Key Figures.[14] A substantial number of these country-years, 174, overlap with the LIS, allowing a test of the comparability of these data similar to the one applied to the SWIID in Figure 5. Given that this approach sacrifices coverage sharply—having only about one-fifth of the observations in the SWIID—for the purpose of maximizing comparability, the results are disappointing: the difference between the UNU-WIDER data and the LIS is substantively and statistically significant in 37 country-years, 21% of the total.

Not surprisingly, the second option—to increase coverage at the cost of comparability by using data generated using multiple combinations of equivalence scale and welfare definition—fares even worse than the first. Beginning in 2010, the *All the Ginis* dataset adopted the LIS as its most preferred

[14]Combining these data—none of which are sourced from the LIS—with the LIS Key Figures provides the broadest country-year coverage of usable observations available for any combination of equivalence scale and income definition in the UNU-WIDER (2014) data. For the purposes of this test, when more than one observation of household-adult-equivalent disposable income inequality for a given country-year was available in this dataset, these observations were averaged.

source of income inequality data, using other sources to fill in observations of the `Giniall` series for which LIS data was not available (Milanovic 2010, 2-3). The result is considerably better coverage than that achieved via the first option—a total of 2043 country-years in the most recent version (Milanovic 2013)—but at the cost of mixing observations calculated on eight different bases. To assess the comparability of these data, for the 81 country-years with non-LIS `Giniall` data, I took the difference between `Giniall` (using the latest of the 2010, 2012, and 2013 revision of the dataset with non-LIS data for that country-year) and the Gini index of inequality in household per capita disposable income (*All the Ginis*' preferred combination of welfare definition and equivalence scale) calculated from LIS data. No fewer than 39 of the non-LIS `Giniall` country-years, or 48%, were substantively and statistically significantly different from the LIS data now available. Using the *All the Ginis* dummies to make global fixed adjustments within each revision as suggested by Milanovic (2013, 8) actually yields even fewer good predictions: fully 58% of the adjusted non-LIS `Giniall` observations are sustantively and statistically significantly different from the LIS.[15] The SWIID's approach of using the LIS as a standard and estimating its missing values with all of the available data, using as much information as possible from the same country and proximate years, does a much better job of providing comparable estimates than either of the two more straightforward options presented by the tradeoff of comparability and coverage.

Still, further examination of Figure 5 raises two issues that give some pause. First, though countries new to the LIS are generally predicted as well as others, the previous SWIID estimates for India in 2004 and China in 2002 (both from Version 3.1) badly understated the level of inequality in those two countries. These country-years constituted not only the first LIS observation in each country but also the first LIS observations of the developing countries in Asia. The relationships between the LIS data and other Gini indices calculated on different bases observed elsewhere in the developing world proved not to provide good estimates for this region.[16]

[15]Excluding those non-LIS Ginis based on the INDIE series in the 2013 revision, which Milanovic (2013) prefers to the LIS, yields very similar rates: 19 of the 41 differences (46%) between the unadjusted data and the LIS—and 25 of the 41 (61%) differences between the adjusted data and the LIS—are substantively and statistically significant.

[16]The difference between the first LIS observation in Africa, of South Africa in 2008, and the closest previous SWIID prediction for that country, for 2005, was just 4 ± 7 points, suggesting that these relationships are particularly distinctive in developing Asia.

Now that all regions of the world have at least some representation in the LIS, such large errors are unlikely to persist in the SWIID, but this points again to the continuing need to minimize reliance on information from other countries and particularly other regions.

Second, only 51 of these 71 differences, or 72%, have 95% confidence intervals that include zero. This suggests that the standard errors associated with the SWIID estimates have often been too small. Some corrections have been made in the code used to generate Version 5 of the SWIID to help ensure that the certainty of the estimates is not overstated. It is possible, however, that the lack of any consideration of the sampling errors of the Gini indices in the source data continues to be an important cause of overconfidence. Measures of sampling error—or even the information needed to calculate them—are only rarely provided in the sources of these data, but a reasonable estimate would likely be preferable to the present assumption that the sampling error is zero. The possibility of incorporating the sampling error in the source data will be considered in future versions of the SWIID.

The sum of these assessments, however, is very positive. The SWIID covers a broader sample of countries and years than any other income inequality dataset, allowing researchers to investigate differences in levels and trends in inequality that would otherwise go unexamined. It entirely avoids the dubious assumption that differences in inequality statistics calculated on two dissimilar bases are constant across space and time inherent in global fixed adjustments, and it minimizes reliance on information from other countries and regions. Finally, the SWIID has done a very good job of predicting LIS data before its release, lending confidence in its cross-national and over-time comparability, and its performance on this score far surpasses that of other cross-national income inequality datasets when tested similarly. In light of the foregoing, the SWIID is clearly the best source available for broadly cross-national work on income inequality.

Using the SWIID for Cross-National Research

The SWIID can be accessed in two ways. First, to facilitate straightforward comparisons of levels and trends in income inequality, the SWIID is now available as a user-friendly web application built using RStudio and Shiny. The web application allows users to graph the SWIID estimates of any of net-income income, market-income inequality, relative redistribution, or absolute

redistribution in as many as four countries or to compare these measures within a single country. Its output can be downloaded with a click for use in reports or articles.

Second, to perform statistical analyses, the SWIID data are available pre-formatted for use with the tools developed for analyzing multiply imputed data in Stata (the `mi estimate:` prefix) and in R (the `mitools` package). These tools automate the process of performing an analysis repeatedly using multiple Monte Carlo simulations and averaging the results (see King et al. 2001). Although these tools introduce a bit more complexity into an analysis, they are necessary for taking the uncertainty in the SWIID estimates into account. A review of how to use these tools, along with examples, is included in the download.

Whether using the web application or the pre-formatted data, most researchers will find the net-income inequality series to be the best suited to their needs conceptually. Market-income inequality, although accurately described as measuring the distribution of income before taxes and transfers are taken into account, cannot be considered 'pre-government': a wide range of non-redistributive government policies, from public education and job-training programs to capital-accounts regulations, also shape the income distribution (see, e.g., Iversen and Stephens 2008; Morgan and Kelly 2013). In addition to such market-conditioning policies, market-income inequality also includes the feedback effects of redistributive policies on households' decisions regarding savings, employment, and retirement. Where robust public pension programs are in place, for example, most households will save little for retirement; as a result, most elderly households will be without market income and market-income inequality will be exaggerated in comparison to settings in which public pensions are less complete (see, e.g., Bradley et al. 2003; Jesuit and Mahler 2010). By affecting market-income inequality, of course, market conditioning and policy feedback also mean that measures of absolute and relative redistribution such as those included in the SWIID do not map straightforwardly onto the broader concept of 'the effect of government on inequality.' Morgan and Kelly (2013) present an example of how market-income inequality, absolute redistribution, and net-income inequality can be used in tandem to investigate how governments affect inequality through both market conditioning and redistribution.

Both the SWIID web application and the pre-formatted data are available at the author's website, `http://myweb.uiowa.edu/fsolt`. All of the files needed to replicate the SWIID, including all of the source data, are available

there as well.

It bears underscoring that the SWIID represents a particular choice in the balance between comparability and coverage: it maximizes comparability for broadest possible coverage of countries and years. This makes the SWIID ideal for broadly cross-national work, but it is not the most appropriate choice for all research on income inequality. Greater comparability can often be achieved when one's scope of inquiry is narrower. Though it offers only limited coverage, the Luxembourg Income Study provides higher quality, superior comparability, and greater flexibility than the SWIID; these traits will continue to make it the preferred source for many cross-national studies. Those studying changes in inequality over time in a single country, further, will often find that examining the national sources found within the SWIID source data and becoming familiar with the exact assumptions and definitions they employ will better meet their needs. Approaches using all these data sources hold promise for advancing our understanding of economic inequality and its causes and consequences.

References

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2014. "The World Top Incomes Database." http://topincomes.g-mond.parisschoolofeconomics.eu. 23 July 2014.

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. 2011. "Top Incomes in the Long Run of History." *Journal of Economic Literature* 49(1):3–71.

Babones, Salvatore J., and María José Alvarez-Rivadulla. 2007. "Standardized Income Inequality Data for Use in Cross-National Research." *Sociological Inquiry* 77(1):3–22.

Bradley, David, Evelyne Huber, Stephanie Moller, François Nielsen, and John D. Stephens. 2003. "Distribution and Redistribution in Postindustrial Democracies." *World Politics* 55(2):193–228.

CEDLAS. 2013. "Socio-Economic Database for Latin America and the Caribbean (SEDLAC)." http://sedlac.econo.unlp.edu.ar/eng/statistics-detalle.php?idE=35. December 2013.

CEPALStat. 2014. "Social Statistics and Indicators Database." http://interwp.cepal.org/sisgen/ConsultaIntegrada.asp?idIndicador=250&idioma=e. October 1, 2014.

Deininger, Klaus, and Lyn Squire. 1996. "A New Data Set Measuring Income Inequality." *World Bank Economic Review* 10(3):565–591.

Eurostat. 2014. "Gini Coefficient of Equivalised Disposable Income, ilc_di12." http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di12&lang=en. September 16, 2014.

Förster, Michael, Ana Llena-Nozal, and Vahé Nafilyan. 2014. "Trends in Top Incomes and their Taxation in OECD Countries." OECD Social, Employment, and Migration Working Papers, No. 159.

Galbraith, James K. 2009. "Inequality, Unemployment and Growth: New Measures for Old Controversies." *Journal of Economic Inequality* 7(2):189–206.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Halter, Daniel, Manuel Oechslin, and Josef Zweimüller. 2014. "Inequality and Growth: The Neglected Time Dimension." *Journal of Economic Growth* 19(1):81–104.

Iversen, Torben, and John D. Stephens. 2008. "Partisan Politics, the Welfare State, and Three Worlds of Human Capital Formation." *Comparative Political Studies* 41(4-5):600–637.

Jesuit, David K., and Vincent A. Mahler. 2010. "Comparing Government Redistribution Across Countries: The Problem of Second-Order Effects." *Social Science Quarterly* 91(5):1390–1404.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95(1):49–69.

Kirsanova, Tatiana, and James Sefton. 2007. "A Comparison of National Saving Rates in the UK, US and Italy." *European Economic Review* 51(8):1998–2028.

LIS. 2014*a*. "Inequality and Poverty Key Figures." `http://www.lisdatacenter.org`. August 2014. Luxembourg: LIS.

LIS. 2014*b*. "Luxembourg Income Study Database." `http://www.lisdatacenter.org`. Multiple countries; August 2014. Luxembourg: LIS.

Milanovic. 2010. "Description of *All The Ginis* Dataset." World Bank, Research Department.

Milanovic. 2013. "Description of *All The Ginis* Dataset." World Bank, Research Department.

Morgan, Jana, and Nathan J. Kelly. 2013. "Market Inequality and Redistribution in Latin America and the Caribbean." *Journal of Politics* 75(3):672–685.

Mukhopadhaya, Pundarik, G. Shantakumar, and Bhanoji Rao. 2011. *Economic Growth and Income Inequality in China, India, and Singapore: Trends and Policy Implications*. London: Routledge.

OECD. 2014. "Income Distribution Database." `http://stats.oecd.org/Index.aspx?DataSetCode=IDD`. October 1, 2014.

Solt, Frederick. 2009. "Standardizing the World Income Inequality Database." *Social Science Quarterly* 90(2):231–242.

Solt, Frederick. 2015. "Economic Inequality and Nonviolent Protest." *Social Science Quarterly* 96.

Statistics New Zealand. 1999. *New Zealand Now: Incomes*. Wellington: Statistics New Zealand.

University of Texas Inequality Project. 2013. "The UTIP-UNIDO Data Set." `http://utip.gov.utexas.edu/data.html`.

UNU-WIDER. 2008. "World Income Inequality Database, Version 2.0c, May 2008." `http://www.wider.unu.edu/research/Database/en_GB/database`.

UNU-WIDER. 2014. "World Income Inequality Database, Version 3b, September 2014." `http://www.wider.unu.edu/research/WIID3-0B/en_GB/database`.