WIDER TECHNICAL NOTE | 8/2021

World Income Inequality Database (WIID)

WIID companion (May 2021): integrated and standardized series

Carlos Gradín*

May 2021

Abstract: This document is part of a series of technical notes describing the compilation of a new companion database that complements the World Income Inequality Database (WIID). A previous note described the selection of income distribution series. Since these series may differ across welfare concepts and other methods used, this technical note describes the second stage, constructing integrated and standardized country series. It provides an update on the procedure used for the May 2021 version of the data. It discusses all the necessary adjustments conducted to construct the final series for each country, with consistent estimates of the distribution of net income per capita over the entire period for which information is available. This is mainly divided into two stages. First, integrating country series by interlinking series that overlap over time, then using a more general regression-based approach.

Key words: integrated series, adjustments, inequality, WIID, WIID Companion

JEL classification: C82, D31, E01, O15

Acknowledgements: The author is grateful for valuable contributions provided by Dmitry Dobrikov and Annalena Oppel in preparing parts of this technical note.

Supplementary material, such as code files, are freely available to download from the webpage of this technical note: https://www.wider.unu.edu/publication/wiid-companion-may-2021-integrated-and-standardized-series.

Related publications:

Gradín, C. (2021). 'WIID Companion (May 2021): Data Selection'. WIDER Technical Note 2021/7. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-7

Gradín, C. (2021). 'WIID Companion (May 2021): Global Distribution'. WIDER Technical Note 2021/9. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-9

Gradín, C. (2021). 'Trends in Global Inequality Using a New Integrated Dataset'. WIDER Working Paper 2021/61. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/999-0

* UNU-WIDER, Helsinki, Finland; gradin@wider.unu.edu

This study has been prepared within the UNU-WIDER project World Income Inequality Database (WIID).

Information and requests: publications@wider.unu.edu

https://doi.org/10.35188/UNU-WIDER/WTN/2021-8

United Nations University World Institute for Development Economics Research

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the Institute or the United Nations University, nor the programme/project donors.

1 Introduction

A previous technical note (Gradín 2021b) explained the selection of the series[1] from the UNU-WIDER Income Inequality Database (WIID)[2] that were totally or partially included in the WIID Companion, so there is only one observation per year and country, covering the longest possible period of time. However, the original information may refer to different welfare concepts or differ across other aspects, such as methods, surveys, or population coverage.

About 31 per cent of country–year observations in the WIID Companion originally refer to the target welfare concept (net income per capita), such as those obtained from the Luxembourg Income Study (LIS) and European Union Statistics on Income and Living Conditions (EU-SILC; microdata), or from other reported sources like TransMonEE (UNICEF) for Eastern Europe, the National Bureau of Statistics (NBS) of China, and several research studies (Table 1). In other cases (17 per cent), the scale is also per capita, and the measure of resources refers to income net/gross, which is deemed to be close enough to net income, such as in series coming from ECLAC or SEDLAC in LAC[3] countries, or from the World Bank. This means that 48 per cent of observations do not need to be converted to a different measure of resources or scale. Yet, more than half of observations need some type of adjustment due to the measure of resource (gross income or consumption), scale (per equivalent adult, no adjustment, unknown), or both. Furthermore, some observations have other inconsistencies in the selected series due to changes in geographical coverage of surveys over time, while others may differ in other aspects, such as methods or surveys.

Table 1: Original welfare concept (resource and equivalence scale) in the WIID Companion (percentage observations)

Resource	Per capita	Equivalized					No adjustment (or missing)	Total
		Total	OECD	Mod. OECD	Square root	Other		
Income	49.6	4.8	0.1	2.8	0.2	1.7	15.9	70.3
Net	31.4	4.4	0.1	2.8	0.2	1.3	2.8	38.6
Net/gross	16.5						5.0	21.5
Gross	1.6	0.4				0.4	8.1	10.1
Consumption	28.3	0.3				0.3	1.1	29.7
Total	77.9	5.1	0.1	2.8	0.2	2.0	17.0	100

Note: Other: OECD, supplemental poverty measure, unknown equivalence scale. Original (reported) observations, before any adjustment.

Source: author's construction based on the WIID.

This technical note describes the adjustments undertaken to integrate the various country series in the WIID Companion coming from different sources and to standardize them so that they are

[1] With 'series' here we generally refer to information on the distribution of a welfare concept for a population over time, with some internal consistency in terms of source, survey, population and geographical coverage, or methods. The welfare concept is obtained by pooling a measure of resources (e.g. income or consumption) from the sharing unit (e.g. household) that are available to the reference unit (e.g. person), either in total or after adjusting by household needs (e.g. per capita or per equivalent adult). The statistics that are reported may be the mean, median, aggregate measures of inequality, especially the Gini index, and income shares, mainly by deciles or quintiles.

[2] Dataset in UNU-WIDER (2021a); see user guide in UNU-WIDER (2021b) for a description of the main sources.

[3] Economic Commission for Latin America and the Caribbean; Socio-economic Database for Latin America and the Caribbean; and Latin America and the Caribbean, respectively.

more consistent across countries and over time. This note updates a previous version that refers to the March 2021 release of the data (Gradín 2021a). The purpose of the WIID Companion is to describe the distribution of national net income per capita in all cases. The note explains the general process, providing country examples. The Stata codes used for this process and illustrative country graphs (for the Gini index) are also made available with this technical note.[4]

2 The approach

The procedure used to obtain the WIID Companion is divided into two phases.[5] The process is done with the reported Gini index, as well as with each of 100 percentile income shares in the synthetic distribution that was estimated using the Shorrocks–Wan algorithm based on the reported information on available income shares (quintiles, deciles, bottom and top 5 per cent) as detailed in Appendix A. After all these adjustments, from these percentiles, various inequality indices are estimated.[6]

In the first phase, various series are integrated into one within each country. The series that overlap by at least one year are interlinked, so that older series are generally adjusted (shifted up or down using a common factor) to match the most recent one in the overlapping year. This implicitly corrects the levels of the older series for differences in methods, coverage, etc. with respect to the next one, while keeping the information about its trend. This is the preferred method for adjustment as it takes into account more country-specific factors (some of them unobserved) than other possible adjustments, increasing the consistency among various series for the same country.

However, it is not possible to apply the previous approach when there is no overlap among series. Furthermore, in some cases, after applying it, the resulting series, although internally consistent, does not refer to the target welfare concept (per capita net income), compromising comparisons with other countries (or the same country in periods that could not be adjusted). For these reasons, in a second phase the remaining series that still refer to a different welfare concept (resource and/or scale) are converted to reflect the distribution of per capita net income. This standardization is done using the predictions from regressions that relate the different income concepts in the LIS sample.

In what follows, the approach is explained in more detail.

[4] Access supplementary material on the technical note's webpage: https://doi.org/10.35188/UNU-WIDER/WTN/2021-8.

[5] For a discussion of the main challenges involved in the construction of a cross-country database on income distribution, see, for instance, Atkinson and Brandolini (2001) and Anand and Segal (2008), or the papers in the journal special issue introduced by Ferreira et al. (2015), with a specific paper about the WIID (Jenkins 2015).

[6] Note that there are two different Gini indices in most cases. One adjusted/converted directly, the other estimated from adjusted/converted percentiles. The former is probably more suitable for comparison based only on the Gini index, while the latter is more suitable when comparisons involve the entire distribution or the use of several indices.

3 Phase 1: integrating income distribution by interlinking overlapping series (adjusted values)

3.1 General case

The WIID Companion sequentially combines the available series for a specific country whenever they overlap over time. One series, taken as a reference, is extended backwards by integrating it with series for preceding periods, or forward. The general procedure is explained here, while the next subsection provides examples to illustrate how this works in practice.

Let us label by I_t^k and AI_t^k the reported and adjusted values for any statistic of interest (Gini index or income percentile) at year *t*, for series $k = 1,\ldots,K$, where series are ordered from earliest to most recent period covered.

A reference series generally represents one that best represents inequality in the country, particularly for the most recent periods, or is closest to the desired inequality concept and deemed to be more comparable with other countries. The reference series is labelled as k'. Normally, the reference series is the most recent, $k' = K$, but in a few cases is not, $k' < K$.

The earliest point in time at which two selected series for a country overlap is referred to as the integration point. For all series preceding the reference one, $k < k'$, if series k and $k+1$ overlap in more than one year, then observations of k beyond the integration point $t_{k,k+1}$ will generally be removed unless they do not conflict with the observations of the (integrated) reference series. For series after the reference series, $k > k'$, all observations earlier than the integration point will also be removed.

In some cases, in which there is no overlap by only one year, I assumed they still overlapped. This may introduce some error, but probably smaller than using the more general approach in phase 2. For example, the observation from PovcalNet in 2017 for Kosovo is changed to 2018 so it makes it easier to match the Eurostat observation for that year and convert the entire PovcalNet consumption series to net income.[7] We have also to consider that, due to the lack of uniform criteria, various sources may refer to the same survey with different years (i.e. the reference income year, the first or the last calendar year of the survey). The WIID tries to harmonize this using the last survey year as the reference when there is enough information, but this is not always the case. Therefore, it may happen that two observations from two different series in two consecutive years may still refer to the same survey (but were labelled differently in the original sources).

The levels of the reference series are not adjusted in this phase, while adjustments for the rest of the series are made sequentially when they overlap, using the differential between the value for one series and the adjusted value of the preceding series at the integration point, $t_{k,k+1}$:

$$AI_t^{k'} = I_t^{k'} \text{ (reference series)}$$

$$AI_t^k = I_t^k + \left(AI_{t_{k,k+1}}^{k+1} - I_{t_{k,k+1}}^k \right), \text{ for } k < k' \text{ (before the reference)}$$

[7] A similar approach was used in specific cases in Canada, Haiti, Hong Kong, Ireland, the Netherlands, Republic of Korea, Romania, and Slovakia.

$$AI_t^k = I_t^k + \left(I_{t_{k,k-1}}^k - AI_{t_{k,k-1}}^{k-1}\right), \text{ for } k > k' \text{ (after the reference)}$$

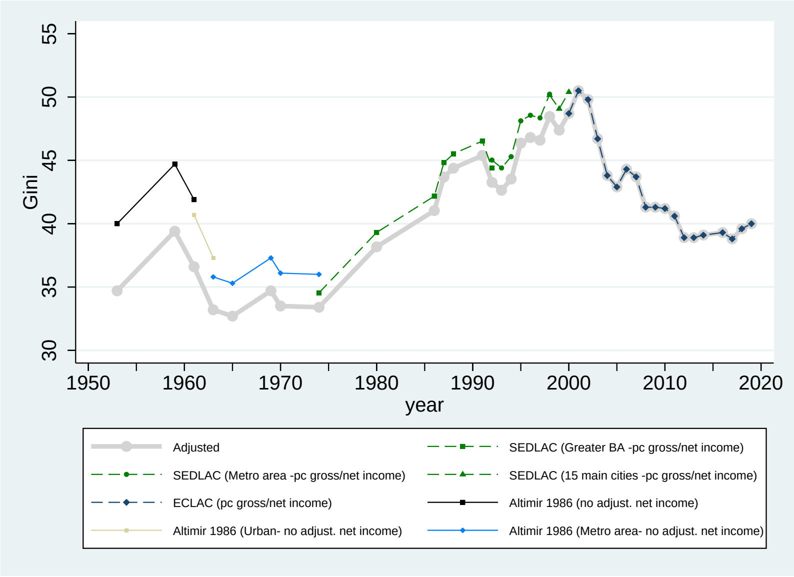
When a series is adjusted to match another one in the integration point, the main variables of the former (resource, scale, geographic and population coverage) are updated with the values of the latter. That is, if a series for total consumption per household is adjusted to match a per capita net income series, the resulting integrated series is considered to be all per capita net income. Similarly, if an urban series is adjusted to match a country series, the resulting series will be considered to be at the country level. Otherwise, if series k and $k+1$ do not overlap, the information of series k is not updated and $AI_t^k = I_t^k$.

Overall, this approach implies that the implicit correction factors for overlapping series are country-specific. It generally corrects not only for differences in the welfare concepts, but also for any differences in methods, population or geographic coverage, etc. across selected series, assuming that such differences are constant over time. However, it is important to notice that the latter is a simplifying assumption and might not hold true in real terms. In the following, a few examples are shown for the adjustment of the Gini index during this phase.

3.2 Examples

Figure 1 shows that the Gini inequality trend in Argentina can be represented by seven series. The most recent provided by ECLAC functions as the reference series. All seven series overlap and therefore there are six integration points: 1961, 1963, 1974, 1992, 1998, and 2000. Note that observations from a preceding series beyond the integration point with the next series are not represented in Figure 1.

Figure 1: Adjusting selected series for Argentina in phase 1 (integration), Gini index



Note: the ECLAC series is taken as the reference. This series is sequentially extended backwards with preceding series from SEDLAC and Altimir (1986) to construct the adjusted series. This will be the final series in WIID Companion in this case.

Source: author's construction based on the WIID.

The reference series is used to cover the period 2000–19 in Argentina. These are ECLAC estimates for the country, which refer to per capita income (net/gross) and cover all urban populations (in

Argentina rural areas are not covered by surveys, and urban areas have been progressively included in the survey design).

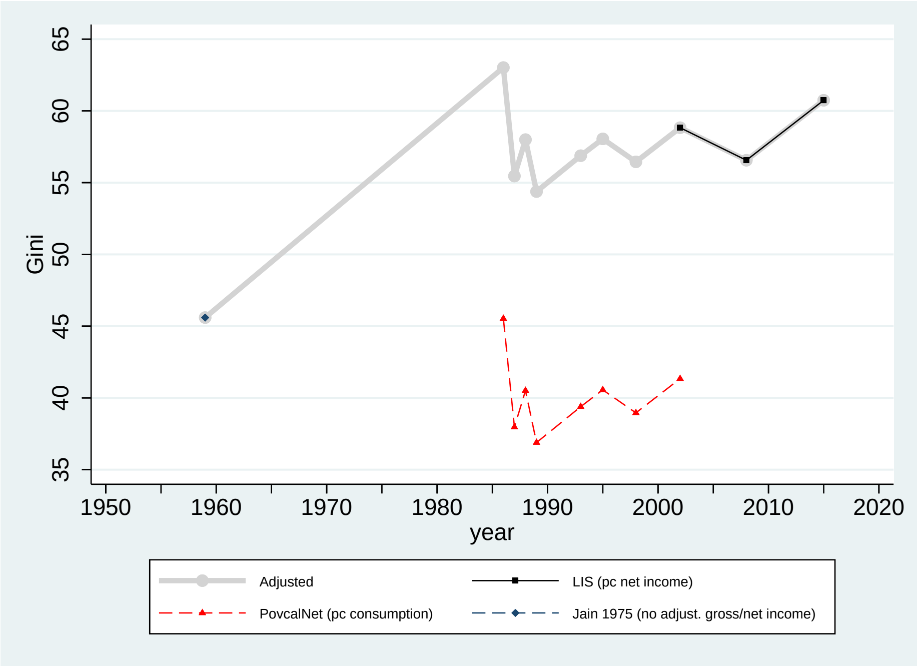
To extend the series backwards, the immediately preceding series for 1998–2000 from SEDLAC for the same income concept is used first, but with a reduced geographical coverage: 28 main cities only. These two series are integrated in 2000, adjusting the value of the Gini index for the older series (50.39) to match the most recent one (48.7). This adjustment factor (−1.69 Gini points) is added to the values of the SEDLAC series for 1998 to 2000 (but 2000 is then removed as it would be a duplicate of the existing observation from ECLAC). Therefore, the adjusted series will reflect the levels implied by the ECLAC series, while preserving the trend from SEDLAC for the years in which ECLAC is not available. This adjustment corrects for the difference in geographical coverage using actual information for the impact of extending this coverage in Argentina, but also for any other unobserved difference in how SEDLAC and ECLAC estimate inequality for the country.

Similarly, the already adjusted two series are extended to include the next preceding series, also from SEDLAC, but now with an even smaller geographical coverage of 15 main cities between 1992 and 1998, taking advantage of the fact that both series overlap in 1998. The 15-city SEDLAC series has a Gini value that year of 50.22, while the 28-city SEDLAC series, after adjustment, has a value of 48.46 (reported 50.15, −1.69 adjustment), implying that the necessary adjustment for the 15-city SEDLAC series should be −1.76 (= −1.69 to 0.07). The new adjusted series is then connected with the preceding series from SEDLAC for Greater Buenos Aires, covering the period 1974–92, and so on, until all overlapping series are connected.

The resulting Gini series is represented by a solid grey line in Figure 1, covering the entire 1958–2019 period. The gap between each reported and adjusted value reflects the accumulated adjustments as shown above. In fact, this will be the final step in the case of Argentina, given that here income (net/gross) for urban Argentina is considered as being close enough to national net income, and no further adjustment will be made in the second phase in this particular case.

Similarly, Figure 2 shows the integration of the LIS net income Gini series for 2002–15 in Côte d'Ivoire with the preceding fragment of the consumption series from PovcalNet for 1995–2002, both in per capita terms. The latter needs to be substantially shifted upwards to match the former in the integration point (2002). This is just reflecting the fact that inequality in this country is much higher if measured with income than with consumption. This can be confirmed by comparing the LIS observations for both resources: Gini is 58.8 for per capita net income in 2002, compared with 48.2 for per capita consumption. The correction made here, however, is also accounting for other possible differences in how PovcalNet or LIS obtain their income distributions (other than the measure of resources). For example, in 2002 the Gini estimate from the World Bank for per capita consumption was 41.3 in 2002, significantly lower than the value estimated in LIS for the same welfare concept. This integrated series will reflect the level of LIS and combine the trends from LIS and PovcalNet (World Bank) and does not need any additional correction since it already refers to net income in per capita terms. An additional observation for 1959 from Jain (1975) is also integrated in the series but is not adjusted in this phase due to the lack of overlapping. The latter has a missing equivalence scale. It will be assumed that it most likely refers to total per household and the value will be slightly corrected for scale in the second phase.

Figure 2: Adjusting selected series for Côte d'Ivoire in phase 1, Gini index



Note: the LIS income series is taken as the reference. This series is sequentially extended backwards with the preceding consumption series from PovcalNet and Jain (1975) to construct the adjusted series (solid grey line). Only the latter value (Jain 1975) will also be converted in the second phase (from total income to net income).

Source: author's construction based on the WIID.

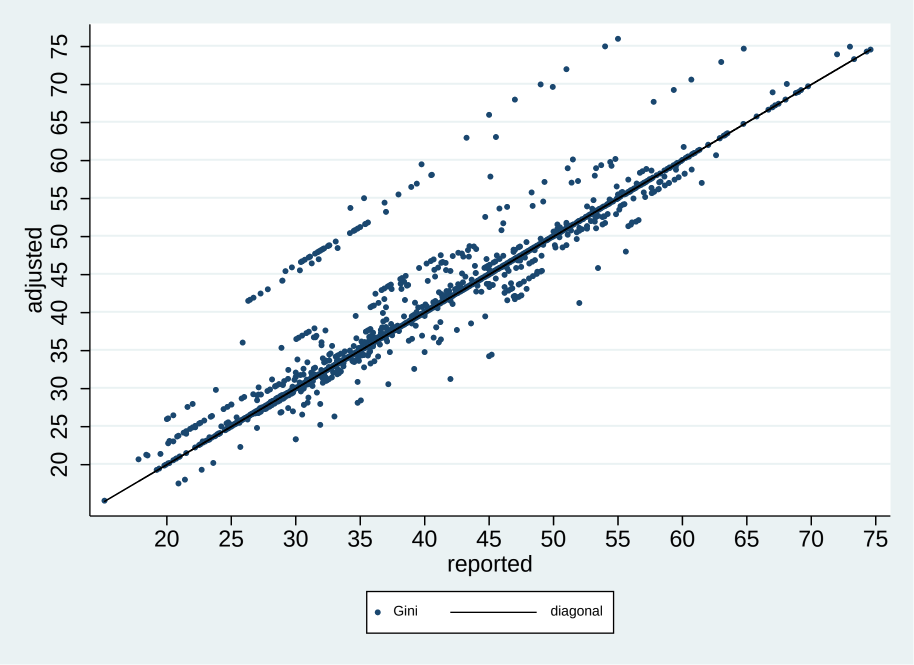
Figure 3 maps all reported and adjusted values for the Gini index—that is, before and after the adjustment takes place in this first phase. Table 2 summarizes these changes in average values of the Gini index among all observations, as well as only among those affected (665 observations, about 28 per cent of the sample).

The adjustments undertaken in this phase imply that the average Gini among those adjusted rises by 2 Gini points, while there is also an increase of the average Gini among all observations, adjusted or not, of 0.6 Gini points. The increase varies greatly, being largest for sub-Saharan Africa (SSA) and South Asia (12.2 and 12.6 Gini points). Pre-apartheid South Africa experiences the largest adjustment (21 Gini points higher). But large adjustments are also necessary in Burkina Faso (19.7), Sudan (19.5), India (16.2), and Côte d'Ivoire (17.5) when the resource shifts from consumption to income. Intermediate adjustments are done in Middle East and North African (MENA) countries (4.1), including up to 5.1 to convert consumption to income estimates in Egypt. The adjustments imply an increase, on average, of less than 1 Gini point in the other areas, but there are also cases of countries with large adjustments (the Dominican Republic, Montenegro, Haiti, Paraguay, Peru, Georgia, Denmark, Bolivia, Russia, Canada, and Turkey exhibit adjustments of 5 points or more). The adjustments imply a reduction of the average Gini in East Asia, driven by the lower inequality reported by LIS as compared with NBS in China (−3.7).[8] Countries in other areas also exhibit large negative adjustments of 5 Gini points or higher (including Hong Kong,

[8] In the case of China, note that LIS and NBS exhibit in 2014 very similar distributions of quintile income shares (and therefore of inequality measures estimated based on them), but the Gini index estimated from LIS is considerably lower than the one reported by NBS (respectively, 43.2 and 46.9).

Argentina, New Zealand, and Venezuela). In general, it is clear that the impact of adjustment is much greater for low-income countries.

Figure 3: Reported versus adjusted Gini index (after phase 1)



Source: author's construction based on the WIID.

Table 2: Average reported and adjusted Gini (phase 1) by region and income group

	All (*N* = 2,418)			Adjusted (*N* = 683)		
	Reported	Adjusted	Change	Reported	Adjusted	Change
Region						
North America	37.1	37.3	0.2	36.2	36.6	0.4
Latin America and the Caribbean	48.5	48.5	0.0	50.0	49.8	-0.2
Europe and Central Asia	31.8	32.2	0.4	31.2	32.7	1.5
Middle East and North Africa	38.6	39.4	0.8	36.9	41.0	4.1
Sub-Saharan Africa	47.4	49.0	1.6	48.4	60.6	12.2
South Asia	36.0	41.7	5.7	32.4	45.0	12.6
East Asia and the Pacific	37.4	37.2	-0.2	39.5	38.5	-1.0
Income group						
High	34.5	34.4	-0.1	35.5	35.2	-0.3
Upper-middle	41.8	42.3	0.5	43.2	44.8	1.6
Lower-middle	40.7	42.7	2.0	38.2	46.1	7.9
Low	42.8	43.6	0.8	42.3	60.1	17.8
Total	38.6	39.2	0.6	38.9	40.9	2.0

Note: 14 observations without reported Gini are excluded.

Source: author's construction based on the WIID.

4 Phase 2: standardization: conversion to per capita net income (final income distributions)

4.1 Adjusted distributions that need to be converted

After values of overlapping series have been adjusted in phase 1 and the key information about resource or equivalence scale updated, the Gini values and income shares of many observations are already expressed in terms of per capita net income (or income net/gross).

As Table 3 shows, 47 per cent of the adjusted values already refer to inequality in net income per capita, while another 12 per cent refer to income net/gross (13 per cent in the case of shares). These values, near 60 per cent in each case, do not need to be converted in the second phase as they are already standardized. The aim of this second phase is to convert all the remaining distributions into net income per capita. Adjustments in the first phase still leave 23 (25) per cent of observations with per capita consumption, 5 per cent of observations with total household (unadjusted) gross income, and around another 5 per cent with equivalized net income, among other combinations.[9]

Table 3: Resource and equivalence scale among adjusted Gini values (after applying phase 1)

	Scale						Total	
	Per capita		Equivalized		No adjustment			
	N	%	*N*	%	*N*	%	*N*	%
Gini								
Income (net)	1,142	47.2	108	4.5	33	1.4	1,283	53.1
Income (net/gross)	294	12.2			90	3.7	384	15.9
Income (gross)	44	1.8	2	0.1	114	4.7	160	6.6
Consumption	558	23.1	6	0.2	27	1.1	591	24.4
Total	2,038	84.2	116	4.8	264	10.9	2,418	100
Percentiles								
Income (net)	1,021	46.9	108	5.0	27	1.2	1,156	53.1
Income (net/gross)	290	13.3			42	1.9	332	15.2
Income (gross)	27	1.2			107	4.9	135	6.2
Consumption	537	24.7	4	0.2	14	0.6	555	25.5
Total	1,875	86.1	112	5.1	191	8.9	2,178	100

Note: in the case of Hong Kong (13 year observations), only the Gini index was adjusted in phase 1 for equivalence scale, not the percentiles, which needed to be converted in the second phase.

Source: author's construction based on the WIID.

[9] About 3 per cent of observations for gross income coming from former socialist countries, or in current SSA and South Asia, will only be converted to account for differences in scale, but not in resources, due to the lack of a good reference for such conversions, and assuming gross and net are similar under those circumstances.

4.2 Final income distributions (integrated and standardized series)

The conversion factors used to standardize the adjusted income distributions are obtained using a regression approach. It exploits the information about how income distributions using different welfare concepts are empirically related by country, region, and/or income group. These regressions use a sample of 3,919 country–year observations in LIS, which includes 493 cases with inequality estimated using per capita net income, which will be paired with observation using other welfare concepts (e.g. 157 cases with per capita consumption or 349 cases for total household gross income). The sample is described in detail in Appendix B.

Using the LIS sample, the corresponding statistic (Gini index or each percentile) for net income per capita is used as the dependent variable. The corresponding statistic for other welfare concepts estimated for the same country and year are used as explanatory variables along with all possible interactions between dummies, reflecting the resource, the equivalence scale, and a geographical unit for each country grouping. That is, if $I_{i,t}^{r,s}$ indicates the LIS value for the statistic of interest in country *i* at period *t* for resource *r* (= net income, gross income, consumption) and equivalence scale *s* (= per capita, no adjustment, OECD-modified, square number of household members), then the estimated model is:

$$I_{i,t}^{net\ income,pc} = \beta_0 + \beta_g I_{i,t}^{r,s} + \beta_{r,s,g} r_{i,t} \# s_{i,t} \# group_i + u_i$$

where # stands for all possible variable interactions. The location variable *group_i* may be alternatively obtained by (1) country, (2) the interaction between geographical region and income group,[10] (3) region, or (4) income group, provided the combination is part of LIS. There are four different regressions for each statistic (i.e. 404 in total: 4 for Gini and 400 for percentiles).

All adjusted values obtained in the first phase that already refer to either per capita net income or income net/gross are not updated in this second phase. All other values are converted.[11] The final statistic, FI_{it}, is obtained by replacing the adjusted statistic (after applying phase 1) by the predicted value in the corresponding regression, depending on the country grouping that applies.[12] That is:

$$FI_{i,t} \begin{cases} AI_{i,t}^{r,s} & \\ \hat{\beta}_0 + \hat{\beta}_{r,s,g} + \hat{\beta}_g AI_{i,t}^{r,s} & \end{cases} = \begin{matrix} \text{'income net, per capita', or 'income net/gross, per capita'} \\ \text{otherwise} \end{matrix}$$

If the country that needs the conversion belongs to the LIS sample, the prediction for that will be used for years not covered by LIS. If the country does not belong to the LIS sample, but belongs to one of the groups defined by country income group and geographical region covered by LIS,

[10] Economies are classified by geographical regions and income groups as defined by the World Bank. However, there is a separate category for non-EU former communist countries in the region.

[11] An exception was made, as indicated earlier, for observations for gross income coming from former socialist countries or in current SSA. These will only be converted to account for differences in scale, but not in resources, due to the lack of a good reference for such conversions, and assuming gross and net are rather similar under those circumstances. In all areas, unknown equivalence scales are generally treated as no adjustment for household size (the most likely case in the context of those observations).

[12] In the case of percentiles, these are re-normalized if necessary to ensure monotonicity. Furthermore, negative and zero values are assigned a very small value close to zero.

this is the predicted value assigned.[13] If it does not belong to any of these, then it is assigned the value based on the predicted region or income group covered by LIS. With this approach, all adjusted values are finally converted to net per capita income.

Note that had the restriction $\hat{\beta}_g = 1$ been imposed, the model would be equivalent to just adding to the adjusted statistic the corresponding average gap between the income distributions of both welfare concepts among LIS observations in the relevant country grouping (i.e. $I_{LIS}^{net\ income,pc,g} - I_{LIS}^{r,s,g}$, where LIS stands for the weighted mean in the LIS sample for country grouping *g*). Thus, the estimated model, with no restriction imposed on $\hat{\beta}_g$, is basically doing this, but in a more flexible way, allowing the magnitude of the correction to vary with the original value of the adjusted statistic (e.g. if $\hat{\beta}_g < 1$, the conversion will imply a larger change for smaller original values).

Note also that this model can be easily expanded by using other exogenous country-level time-varying variables like GDP, public sector size, etc. However, this was explicitly avoided here because their inclusion may compromise the use of these converted values in regressions where those factors are also used as either the dependent or the explanatory variables (e.g. regressions on the relationship between inequality and growth).

4.3 Regression results

The goodness of fit, R^2, of the regressions for the Gini index vary between 91 per cent (regression at the income group level) and 99 per cent (at the country level). This is shown in Table 4, along with the estimated coefficients for the Gini index, showing that in general $\hat{\beta}_g < 1$, and therefore the level of Gini correction depends on the original Gini value (i.e. $\hat{\beta}_g$ varies between 0.70 and 0.96, depending on the country grouping).

Table 4: Regression coefficients for the adjusted Gini index, by country grouping

Variable	Country grouping			
	Country	Region # Income group	Region	Income group
Gini	0.77***	0.70***	0.71***	0.96***
	0.03	0.02	0.02	0.01
Intercept	15.34***	27.11***	10.11***	17.17***
	1.1	0.78	2.57	0.42
R^2	99	96	94	91
N	3,426	3,426	3,426	3,426

Note: coefficients for interactions with resource, scale, and country grouping omitted. Standard errors in parentheses.

Source: author's construction based on the WIID.

[13] Turkey will be assigned to the MENA region in this process.

Table 5 reports, as an example, the detailed results for converting per capita consumption when the country grouping is the combination of region and income group.

To better understand how the correction is applied, let us consider the example of the Republic of Congo, a lower-middle-income SSA country (Table 6).[14] The Gini for per capita consumption in 2005 (PovcalNet) is 47.33, a value that has not been adjusted in phase 1 and needs to be converted in phase 2. The corresponding predicted value for per capita net income is 60.18—that is, 12.85 Gini points higher. This is the result of applying the conversion, based on the region and income group (represented in the LIS sample by Côte d'Ivoire): 27.11 (intercept) + 0.70 × 47.33 (per capita consumption Gini) − 0.24 (lower-middle-income country in SSA). This large adjustment just reflects that, on average, the Gini index for per capita net income in Côte d'Ivoire, the only country in the same region and income group in the LIS sample, is 13.5 Gini points higher than for per capita consumption (58.7 vs 45.2). Similarly, the estimated Gini value for 2011 is 61.31 (+12.37). The value for the Gini index in 1958 needs to use the conversion from total household income instead (i.e. the corresponding coefficient for total household net income is -10.27), changing from 41.90 to 46.30 (i.e. +4.40 Gini points). Table 6 also shows the same type of conversion for the income share of the 50th percentile in the two years with information on income shares (from the original 0.65 per cent of total income to predicted 0.51 in 2011, i.e. −0.14).

Table 5: Regression coefficients for the Gini index, by country grouping

Intercept		27.11***
Gini		0.70***
Income group	Region	Per capita consumption
High Income	Europe and Central Asia	-17.26***
	Middle East and North Africa	-13.55***
	East Asia and the Pacific	-14.30***
Upper-middle income	Latin America and the Caribbean	-5.49***
	Europe and Central Asia	-11.61***
	Middle East and North Africa	-13.40***
	Sub-Saharan Africa	-6.08***
	East Asia and the Pacific	-14.22***
	Non-EU former socialist countries	-14.49***
Lower-middle income	Middle East and North Africa	-12.84***
	Sub-Saharan Africa	-0.24
	South Asia	-2.61***
	East Asia and the Pacific	-14.87***
Low income	Sub-Saharan Africa	-0.58***

Note: omitted: consumption # no adjustment # low-income # sub-Saharan Africa.

Source: author's construction based on the WIID.

Table 6: Example of prediction: Republic of Congo

Year	Original	Gini index	p50

[14] Note that in the case of this country, the only conversion actually needed is from total to per capita income in 1959.

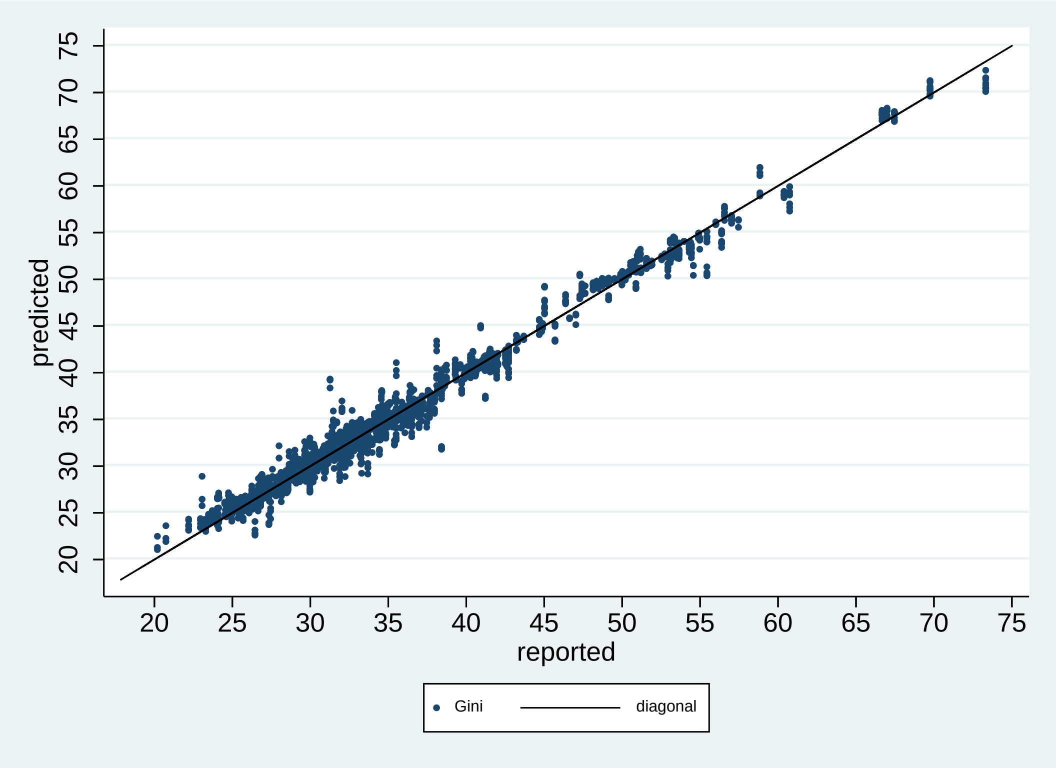
	welfare concept						
		Reported	Per capita net income (prediction)	Differential	Reported	Per capita net income (prediction)	Differential
1958	Total household income	41.90	46.30	4.40	–	–	–
2005	Per capita consumption	47.33	60.18	12.85	0.63	0.50	-0.13
2011	Per capita consumption	48.94	61.31	12.37	0.65	0.51	-0.14

Source: author's construction based on the WIID.

More generally, Figure 4(a–d) maps the predicted values for all observations in the LIS sample (in-sample predictions), regardless of the original welfare concept, by country grouping.[15] In line with the high R^2, in-sample predicted values are highly correlated with the actual values, with variation depending on the country grouping considered: between 94 per cent (income group) and 99 per cent (country), as reported in Table 7. Also, the different country grouping predictions are highly correlated among them.

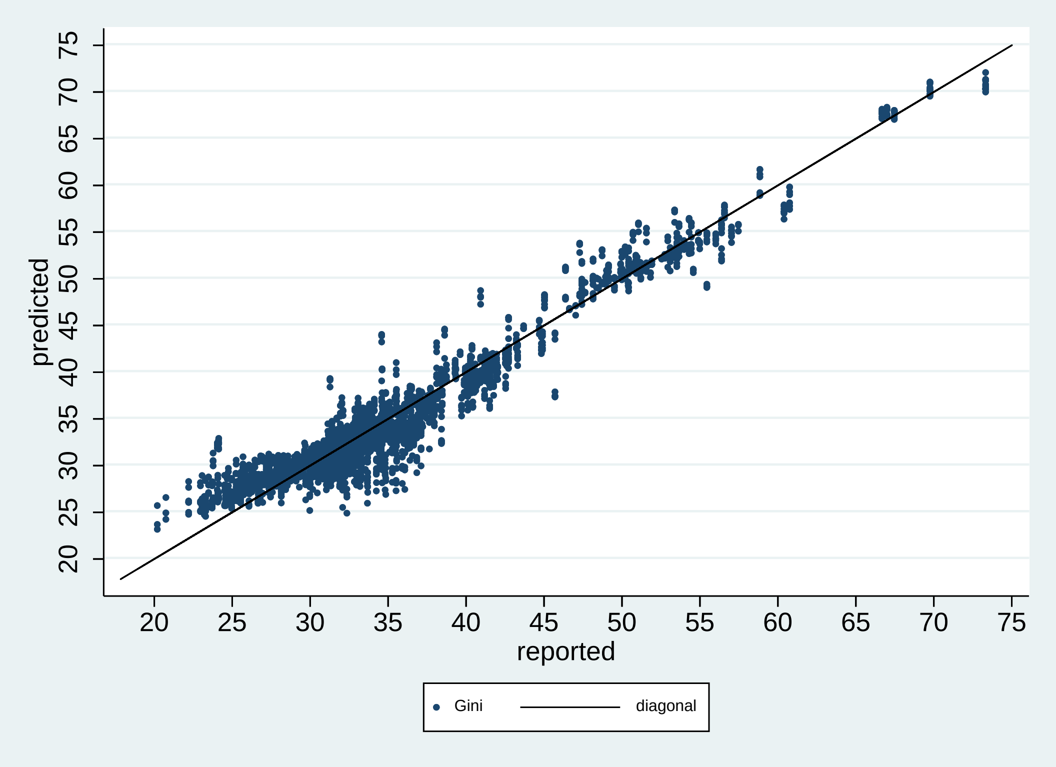
Figure 4: In-sample predictions for all welfare concepts, by country grouping

(a) Country

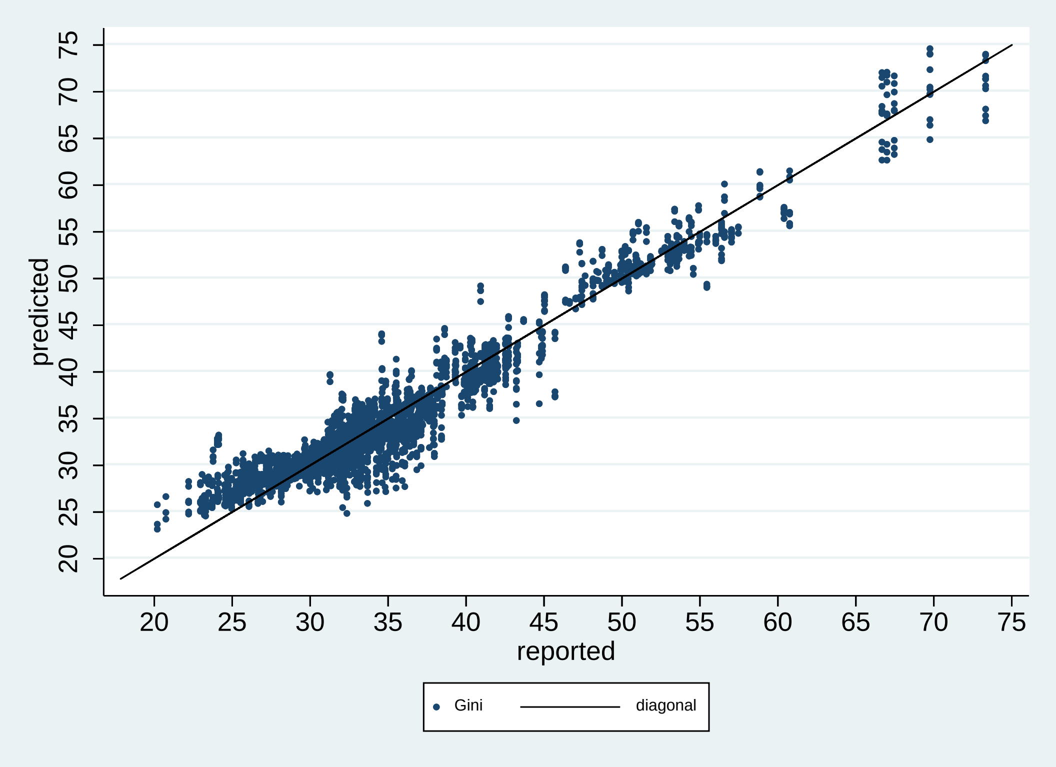


[15] In-sample predictions are shown here for illustration purposes only. Unfortunately, as mentioned above, there is no database (other than LIS) with systematic and consistent information on income and on consumption for a large number of countries that could be used to run out-of-the-sample predictions. Similarly, the LIS sample cannot be used for such purposes because that would imply removing some of the few key observations for developing countries.

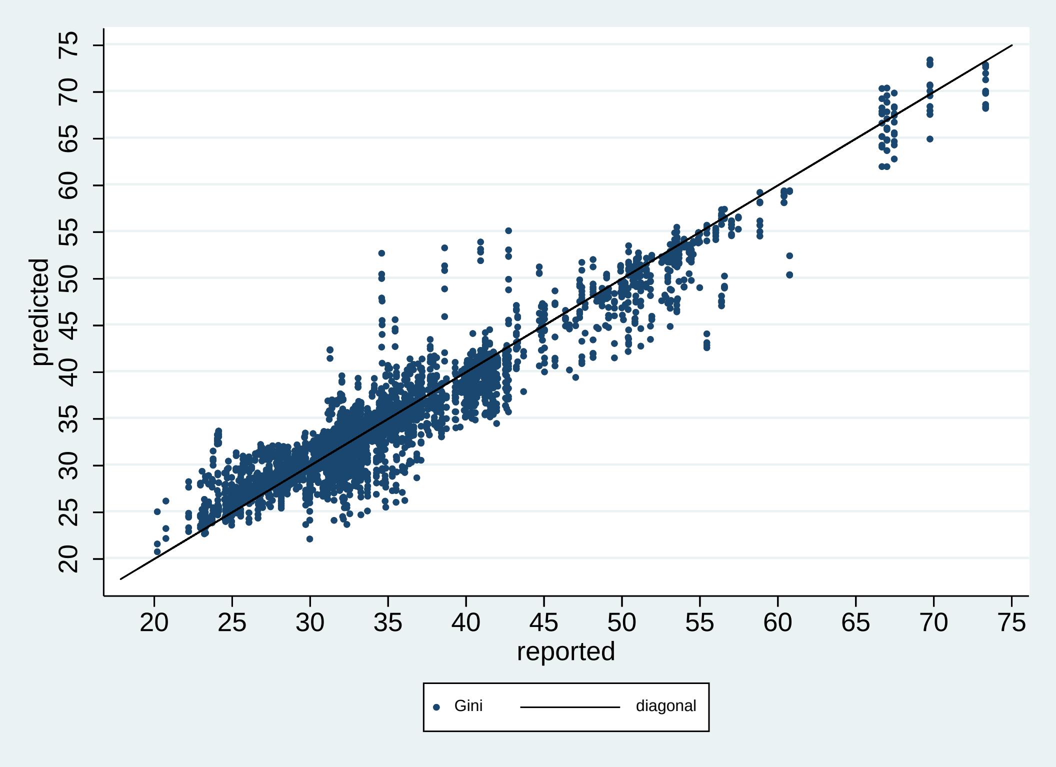
(b) Region and income group



(c) Region



(d) Income group



Source: author's construction based on the WIID.

Table 7: Correlation among actual and predicted values (LIS sample)

	Actual	Country	Region and income group	Region
Country	99.4			
Region and income group	97.7	98.4		
Region	97.1	97.8	99.4	
Income group	95.2	95.8	97.4	97.1

Source: author's construction based on the WIID.

The overall impact on the Gini index of the conversions conducted in this second phase is substantial and in the same direction as in the previous phase, with the largest increases in the average Gini values for SSA and South Asian countries (11 and 9 Gini points; Table 8).

Table 8: Average reported and final (converted) Gini (phase 2) by region and income group

	All (*N* = 2,405)			Converted (*N* = 943)		
	Adjusted	Converted	Change	Adjusted	Converted	Change
Region						
North America	37.3	37.3	0.0			
Latin America and the Caribbean	48.5	49.5	1.0	44.7	50.0	5.3
Europe and Central Asia	32.2	32.8	0.6	33.6	35.3	1.7
Middle East and North Africa	39.4	40.2	0.8	40.8	42.3	1.5
Sub-Saharan Africa	49	57.8	8.8	45.7	56.8	11.1
South Asia	41.7	47.5	5.8	37.6	46.8	9.2
East Asia and the Pacific	37.2	37.1	-0.1	36.0	35.9	-0.1
Income group						

	All (*N* = 2,405)			Converted (*N* = 943)		
	Adjusted	Converted	Change	Adjusted	Converted	Change
High	34.4	34.2	-0.2	36.1	35.1	-1.0
Upper-middle	42.3	43.8	1.5	38.6	42.0	3.4
Lower-middle	42.7	46.3	3.6	39.6	45.4	5.8
Low	43.6	53.8	10.2	41.6	53.2	11.6
Total	39.2	40.8	1.6	38.8	43.0	4.2

Source: author's construction based on the WIID.

5 Overall effect of integrating and standardizing income distributions

This section summarizes the effects of the entire process (phase 1—integration and phase 2—standardization). Table 9 summarizes the number of observations that have been changed in each phase. The Gini index was adjusted in the first phase for a total of 683 country–year observations (28 per cent of the total), while percentiles were adjusted in 580 cases (27 per cent). Similarly, about 39 per cent of all observations were converted in the second phase (943 for Gini; 844 for percentiles). Note that some observations were changed in both phases (82 and 73), while others remained unaffected after both phases (861 or 36 per cent; and 771 or 35 per cent). In the second phase, the most common country grouping was the combination of region and income group (715 or 76 per cent; and 658 or 78 per cent), followed by region and country (income group alone was never used).

Table 9: Summary of changes (number of country–year observations) from reported to final values (Gini and percentiles) in each phase

		Adjustment (integration, phase 1)					
		Gini			Percentiles		
		Not adjusted	Adjusted	All	Not adjusted	Adjusted	All
Conversion (standardization, phase 2)	Not converted	874	601	1,475	826	507	1,333
	Converted	861	82	943	771	73	844
	Country	94	5	99	64	4	68
	Region and income group	654	61	715	602	56	658
	Region	113	16	129	105	13	118
	Income group						
	All	1,735	683	2,418	1,597	580	2,177
Total adjusted or converted				1,545			1,351

Note: country region and income group as defined by the World Bank.

Source: author's construction based on the WIID.

Table 10 summarizes the impact of each phase on the average Gini values by region and income group. The average Gini for the entire sample rises from the original 38.6 to 39.2 after the first phase, and to 40.8 finally, with the largest increases among SSA and South Asian countries (near 10.4 and 111.2 Gini points) and, therefore, among low-income countries. Table 11 reports the impact by original measure of resources and equivalence scale, with the largest increase among those observations originally expressed in consumption terms (7.6).

Finally, Figure 5 compares the reported and the final Gini values. It becomes clear that the majority of adjustments increase the Gini index (74 per cent), but there is still a significant number of cases in which the Gini is actually reduced after the adjustments made in both phases.

Table 10: Summary of changes in Gini, phases 1 (adjusted) and 2 (final): number of observations (*N*) and average Gini index

	All					Adjusted or converted			
	N	Gini				*N*	Gini		
		Reported	Adjusted (phase 1)	Final (phase 2)	Total change		Reported	Final	Total change
Region									
North America	81	37.1	37.3	37.3	0.2	33	36.2	36.6	0.4
Latin America and the Caribbean	504	48.5	48.5	49.5	1.0	279	48.3	50.0	1.7
Europe and Central Asia	963	31.8	32.2	32.8	1.0	522	31.9	33.7	1.8
Middle East and North Africa	144	38.6	39.4	40.2	1.6	90	39.2	41.8	2.6
Sub-Saharan Africa	266	47.4	49.0	57.8	10.4	242	46.1	57.5	11.4
South Asia	99	36.0	41.7	47.2	11.2	97	35.7	47.4	11.7
East Asia and the Pacific	361	37.4	37.2	37.1	-0.3	281	37.3	36.8	-0.5
Income group									
High	1,002	34.5	34.4	34.2	-0.3	475	35.8	35.2	-0.6
Upper-middle	807	41.8	42.3	43.8	2.0	558	40.6	43.4	2.8
Lower-middle	481	40.7	42.7	46.2	5.5	392	39.1	46.0	6.9
Low	128	42.8	43.6	53.8	11.0	119	41.7	53.6	11.9
Total	2,418	38.6	39.2	40.8	2.2	1,544	38.8	42.3	3.5

Note: see Appendix C for details of changes by country.

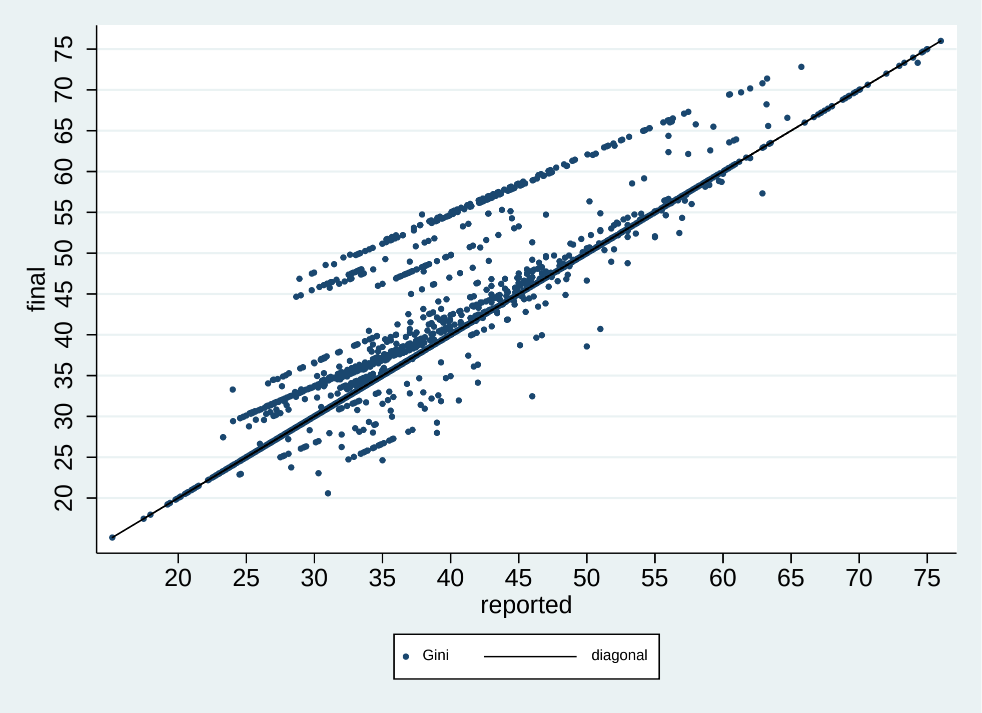
Source: author’s construction based on the WIID.

Table 11: Average reported and final Gini by resource and scale

	Adjusted or converted			All		
	Reported	Final	Change	Reported	Final	Change
Resource						
Income (net)	34.7	35.4	0.7	35.1	35.4	0.3
Income (net/gross)	46.5	46.4	-0.1	46.3	46.2	-0.1
Income (gross)	38.3	37.4	-0.9	38.3	37.5	-0.8
Consumption	37.6	45.2	7.6	37.6	45.2	7.6
Equivalence scale						
Per capita	38.9	43.9	5.0	38.5	41.3	2.8
Equivalized	33.2	34.7	1.5	33.2	34.7	1.5
No adjustment	40.5	40.7	0.2	40.5	40.7	0.2
All	38.8	42.3	3.5	38.6	40.8	2.2

Source: author’s construction based on the WIID.

Figure 5: Reported versus final Gini index



Source: author's construction based on the WIID.

6 Estimating inequality indices from integrated synthetic income distributions

The synthetic distributions estimated at the percentile level from reported information on income shares using the Shorrocks–Wan approach (Appendix A) were integrated in phases 1 and 2, along with the reported Gini index. In a final step, various inequality measures are estimated based on these integrated income distributions (Appendix C). These include summary relative measures of the entire distribution such as the Gini index, the entropy family, and the Atkinson family, as well as income shares of key population groups like the bottom 40 per cent of the population, the top 20 or 10 per cent, and income share ratios derived from them, such as the Palma index (top 10/bottom 40) or the S80S20 ratio (top 20/bottom 20). Measures of absolute inequality are also estimated, with summary measures such as the absolute Gini index or the standard deviation. Absolute measures of inequality use information on mean incomes estimated from GDP per capita in 2017 PPP (purchasing power parity) for the global distribution.[16]

Note that, as a result, there are two measures of the relative Gini index of inequality. In one the reported Gini index is directly integrated. This measure is available even when there is no information for income shares, and takes into account the possibility of negative or zero incomes. Another measure is provided in which the Gini index is indirectly estimated from the integrated synthetic distributions, with other measures, and, therefore, cannot be estimated whenever such information is not available. Furthermore, the estimated Gini index is based on the assumption of all incomes being positive, for the sake of comparability with the other indices (some of them are

[16] The construction of the integrated series for mean income is discussed in another technical note in this series (Gradín 2021c). Note that due to the lack of enough consistent information for country survey mean incomes, we are using here GDP per capita instead.

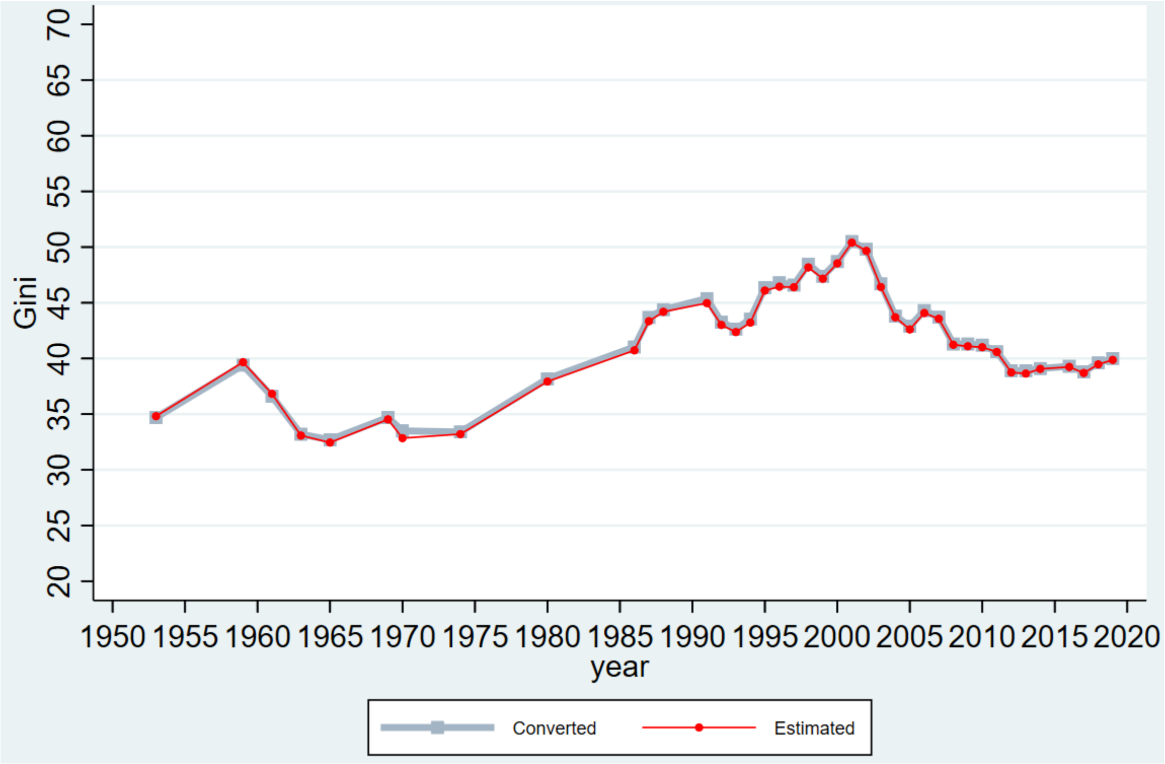
only well defined if incomes are positive). For that purpose, the construction of the synthetic distributions replaced negative and zero incomes with an arbitrary small amount. In general, both Gini indices will be very similar, particularly when the base information is of high quality, given the accuracy required to estimate this index from aggregate information (see Appendix A). However, there are a few cases in which they may substantially diverge. This is due to inconsistencies in the original data from some sources.[17] These inconsistencies are likely due to the fact that the Gini and income shares were actually originally obtained from different distributions (e.g. one per capita and the other total household income). In any case, the series based on the reported Gini is more likely to be useful if one is only interested in the Gini index, having the advantage of a larger coverage. The estimated Gini, however, has the advantage of greater comparability with other inequality indices and with the detailed distribution, and is therefore the one recommended for more general analyses of changes in the distribution.

Figure 6(a–g) displays the integrated series of the Gini index obtained in both ways for a few country examples. It highlights the similarity, in general, but also the fact that some discrepancies may emerge in some cases, as well as the fact that the series based on the reported Gini index is denser and/or longer, particularly for the earliest periods (e.g. Brazil, China, Republic of Congo, and South Africa). Note that the cases of China and South Africa are different from the rest, because the estimated and converted values in the same year may come from different series. This is due to the fact that, to cover some periods, the series that best represents the Gini index does not have information on income shares (e.g. Ravallion and Chen (2001) for China between 1981 and 2001, and several research studies for South Africa between 1960 and 2005). The lack of information is only partially covered using PovcalNet in South Africa since 1993 with information on income shares, and PovcalNet and the World Bank's Poverty Monitoring Database in China (1981–98).

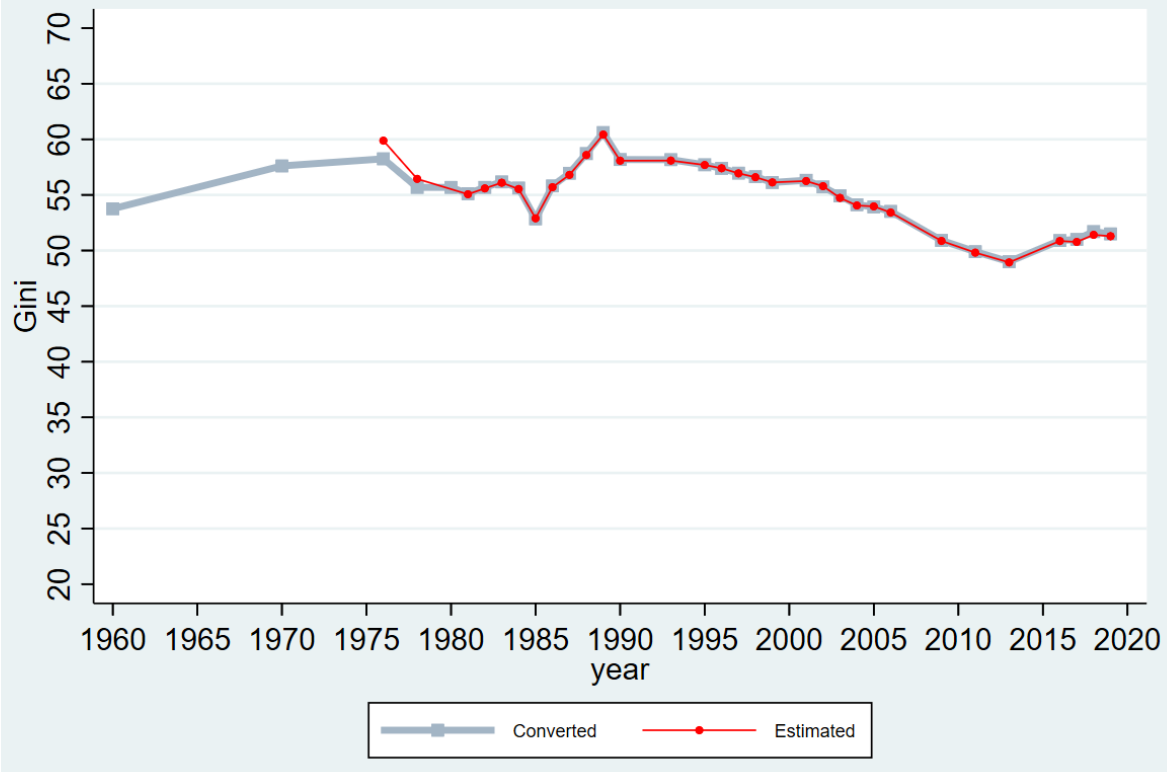
[17] See, for example, the discussion in Anand and Segal (2008) or in Shorrocks and Wan (2009).

Figure 6: Final series for the Gini index (converted) and Gini index estimated from converted income percentiles (estimated)

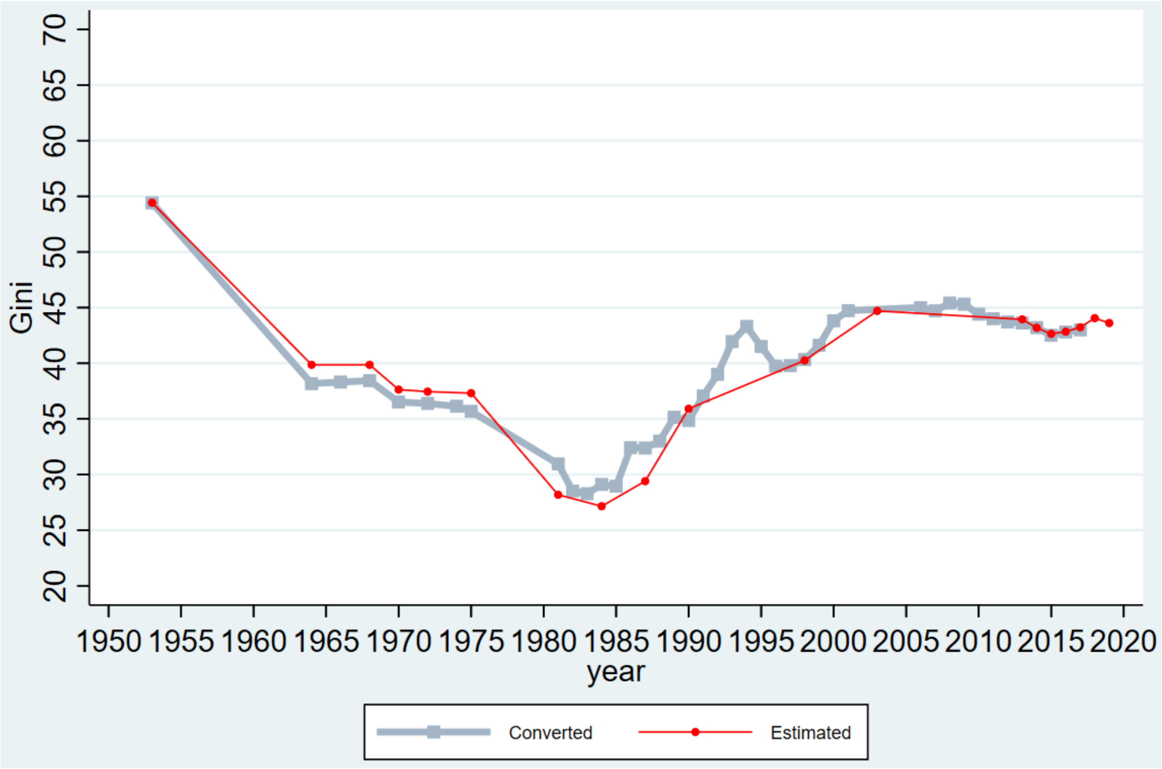
(a) Argentina



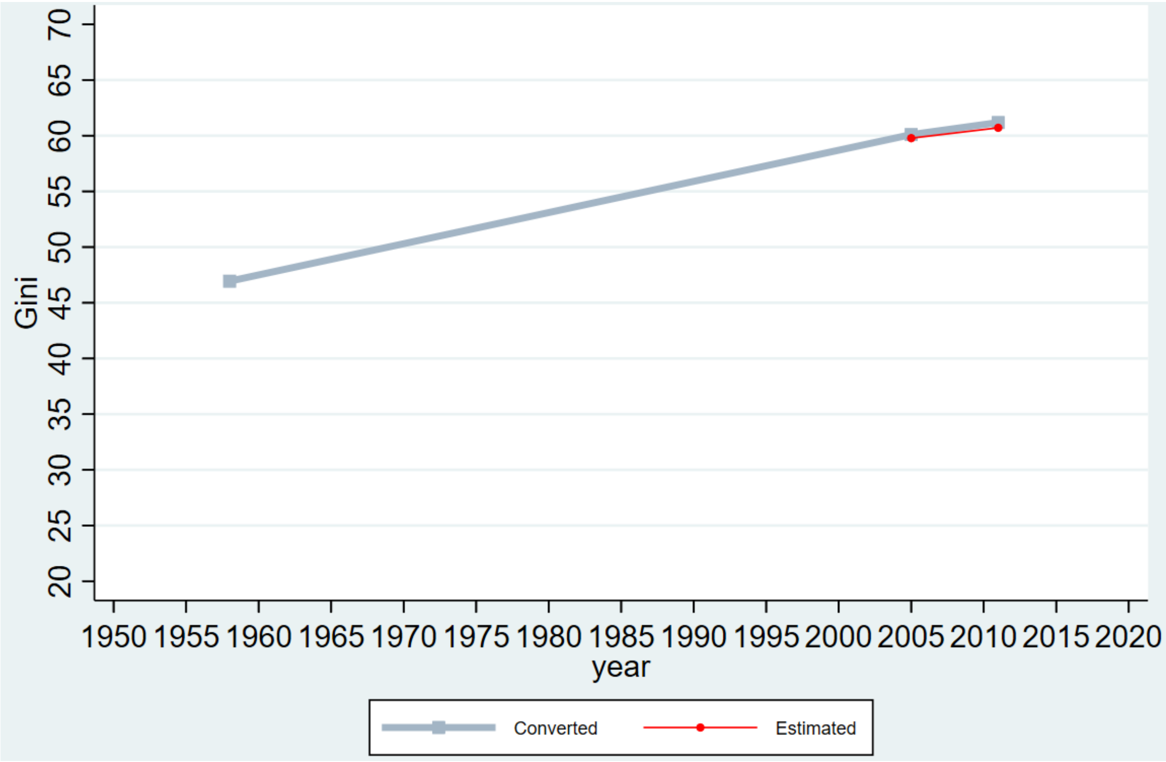
(b) Brazil



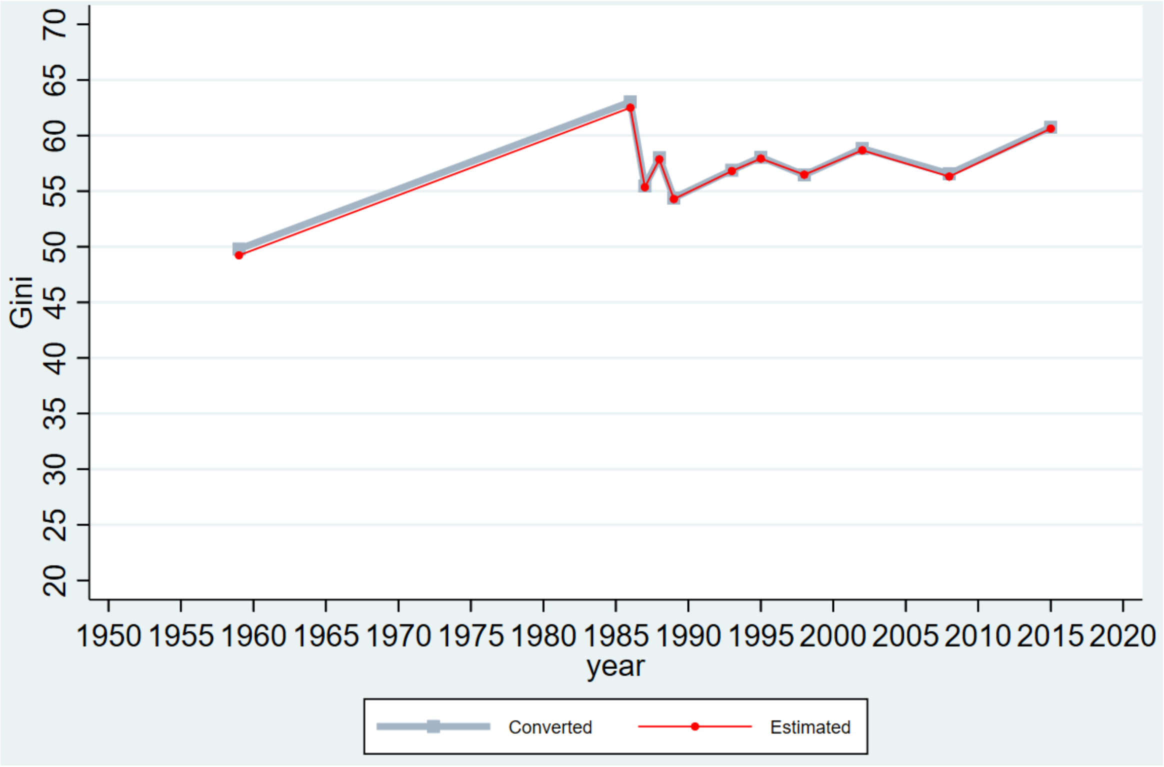
(c) China



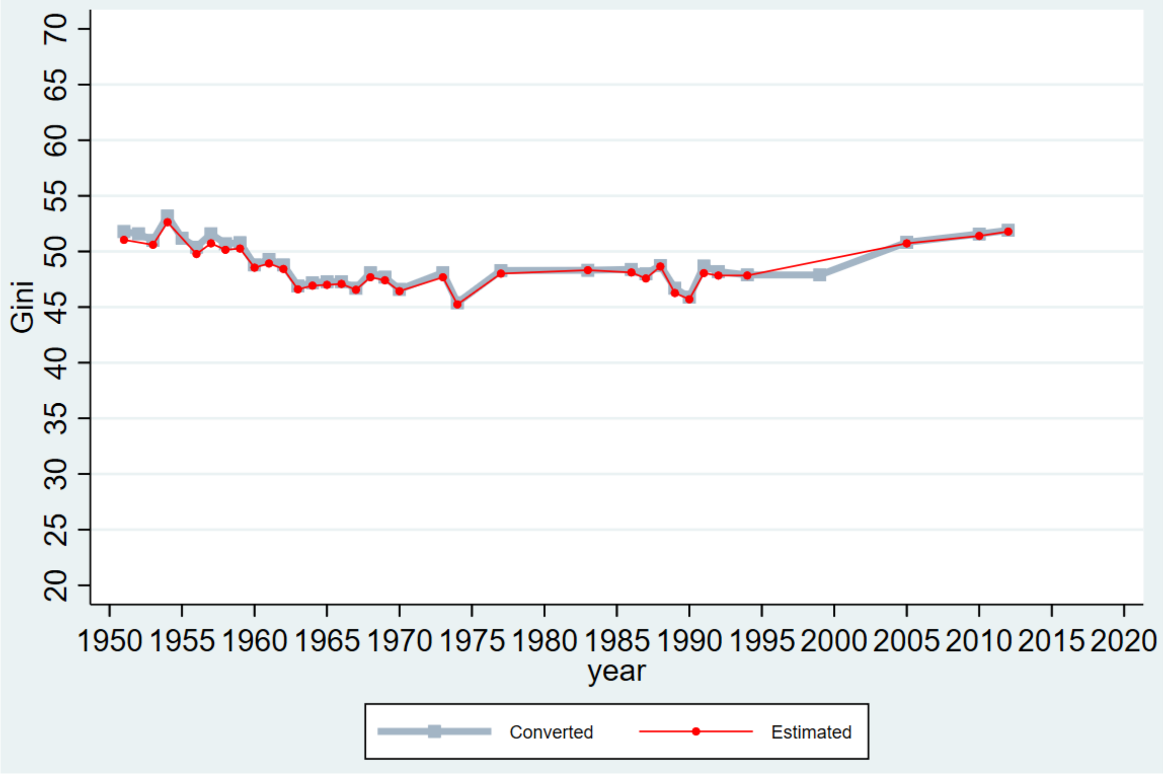
(d) Republic of Congo



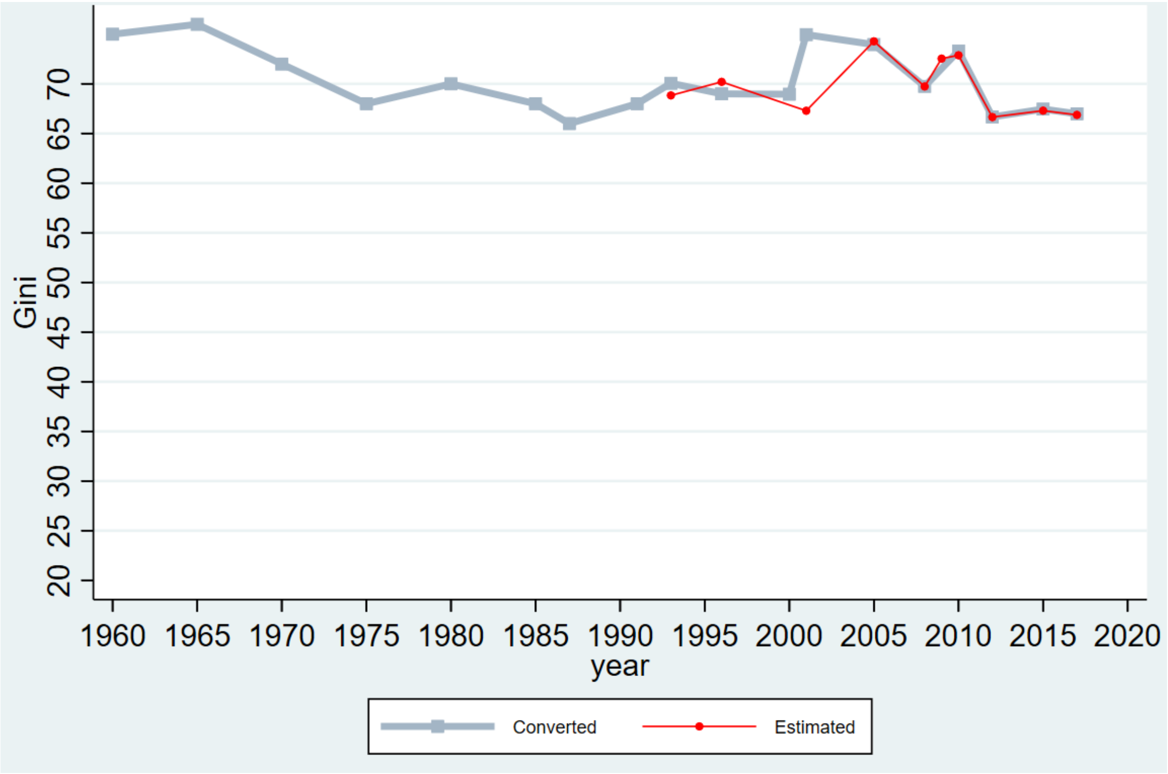
(e) Côte d'Ivoire



(f) India



(g) South Africa



Source: author's construction based on the WIID.

7 Final considerations

A previous technical note in this series (Gradín 2021b) has already stressed important quality considerations regarding the use of this database, in particular in relation to the selection of available information. It highlighted the existence of a clear trade-off between the quality and amount of information used in any analysis (time span considered, number of countries, especially in the developing world or earliest years).

In general, information after 1980 will not only be denser in terms of countries and people covered, but also of higher quality in many cases. This section discusses other aspects that should be considered regarding the construction of the database, which refer to the manipulation (or not) of the original data to produce the final estimates for income shares and inequality measures.

First, the objective of this database, so far, is to integrate and standardize distributive values obtained (almost entirely) from household surveys, not to correct for any flaws typically associated with the use of survey-based information.[18] Therefore, no adjustment or correction was made if the information reported by the original source already verifies this desired consistency over time and across countries. That is, at this stage, there is no attempt to correct for measurement problems related to the mis-estimation of certain incomes or the misrepresentation of people at the extremes of distributions. Notwithstanding this, the possibility of complementing the database with such

[18] Note, however, that in some cases, particularly European countries, there is a growing tendency to retrieve the information for incomes provided in surveys from administrative records.

corrections that may increase the reliability of estimates is left for future developments of the database.

Second, as discussed in Gradín (2021b), there is no correction for outliers. The general principle is to not smoothen the series, so they are presented as they would be typically obtained from household surveys. Therefore, outliers will be kept if they are already in the original series, unless it is obvious that they are too extreme and therefore not reliable. This is because some events, such as an economic crisis or structural reform, may justify some sharp trends, and to assess whether an outlier is legitimate or not requires specific knowledge of the context in which it occurs. Particular attention was paid to avoid outliers that are not part of any series, but result from merging two different and heterogeneous series.

References

Altimir, O. (1986). 'Estimaciones de la Distribución del Ingreso en la Argentina, 1953-1980'. *Desarrollo Económico*, 25(100): 521–66. https://doi.org/10.2307/3466844

Anand, S., and P. Segal (2008). 'What Do We Know about Global Income Inequality?''. *Journal of Economic Literature*, 46(1): 57–94. https://doi.org/10.1257/jel.46.1.57

Araar, A., and J.-Y. Duclos (2013). *User Manual, DASP: Distributive Analysis Stata Package*, version 2.3. Quebec City: Université de Laval.

Atkinson, A.B., and A. Brandolini (2001). 'Promise and Pitfalls in the Use of "Secondary" Data-Sets: Income Inequality in OECD Countries as a Case Study'. *Journal of Economic Literature*, 39(3): 771–99. https://doi.org/10.1257/jel.39.3.771

Ferreira, F.H.G., N. Lustig, and D. Teles (2015). 'Appraising Cross-National Income Inequality Databases: An Introduction'. *Journal of Economic Inequality*, 13: 497–526. https://doi.org/10.1007/s10888-015-9316-0.

Gradín, C. (2021a). 'WIID Companion (March 2021): Integrated and Standardized Series'. WIDER Technical Note 2021/5. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-5

Gradín, C. (2021b). 'WIID Companion (May 2021): Data Selection'. WIDER Technical Note 2021/7. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-7

Gradín, C. (2021c). 'WIID Companion (May 2021): Global Distribution'. WIDER Technical Note 2021/9. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-9

Jain, S. (1975). *Size Distribution of Income: A Compilation of Data*. Washington, DC: World Bank.

Jenkins, S.P. (2015). 'World Income Inequality Databases: An Assessment of WIID and SWIID'. *Journal of Economic Inequality*, 13: 629–71. https://doi.org/10.1007/s10888-015-9305-3

Ravallion, M., and S. Chen (2007). 'China's (Uneven) Progress Against Poverty'. *Journal of Development Economics*, 82(1): 1–42. https://doi.org/10.1016/j.jdeveco.2005.07.003

Shorrocks, G., and G. Wan (2009). 'Ungrouping Income Distributions Synthesis Samples for Inequality and Poverty Analysis'. WIDER Working Paper 16. Helsinki: UNU-WIDER.

UNU-WIDER (2021a). 'World Income Inequality Database (WIID)'. Version 31 May 2021. Available at: https://www.wider.unu.edu/database/world-income-inequality-database-wiid.

UNU-WIDER (2021b). 'World Income Inequality Database (WIID): User Guide and Data Sources'. Available at: https://www.wider.unu.edu/sites/default/files/WIID/WIID-User-Guide-31MAY2021.pdf

Appendix A: Estimating percentile distributions based on aggregate income shares

Most of the series included in the WIID Companion inform about country income distributions by reporting information about the income share of different population groups. As explained in Gradín (2021b), not all observations provide the same level of aggregation. Most cases report at least the full set of income shares by decile, others by quintile; in both cases these values may be complemented by the share for the bottom and/or top 5 per cent (Table A1).

Table A1: Original distributional information in the WIID Companion (percentage observations)

Available income shares	%
None/incomplete	10.0
Complete set of income shares	90.0
Full (deciles + top 5% + bottom 5%)	30.0
Deciles and top 5%	2.4
Deciles and bottom 5%	0.0
Deciles	48.1
Total with at least deciles	80.6
Quintiles and top 5%	2.2
Quintiles and bottom 5%	7.2
Quintiles	0.1
Total with at least quintiles	9.5

Note: original (reported) observations, before any adjustment.

Source: author's construction based on the WIID.

Several inequality measures can be estimated directly using this aggregate information, but those estimates would be lower bounds because they omit any possible inequality within these population groups (e.g. one must assume income is uniformly distributed about all individuals in the same decile, quintile, or ventile). The importance of this omission is obviously greater in the cases in which the distribution is more aggregated (i.e. only quintile shares are reported).

To avoid that, a more detailed synthetic distribution is estimated in the WIID Companion in all cases (90 per cent) that report at least a full set of quintiles, using the available aggregate information. This produces for each country–year a set of 100 observations (percentiles) reporting their corresponding relative average income, which is later (after all the necessary adjustments) used to more accurately estimate distributional measures that are not reported in the original source. This procedure also facilitates the interlink of series or comparisons among distributions with originally different degrees of aggregation.

This process of disaggregating grouped data is done using a parametric approach, applying the Shorrocks–Wan algorithm (Shorrocks and Wan 2009). This approach is implemented using the UNGROUP Stata command included in the Distributive Analysis Stata Package (DASP) produced by the Université de Laval (see Araar and Duclos 2013).

The procedure allows individual income observations to be reconstructed from any feasible grouping pattern. The method first assumes that incomes (*y*) follow a log-normal distribution Φ

with mean $\mu = 1$ and variance σ^2, which can be directly estimated using the procedure described by Shorrocks and Wan (2009): $\Phi\left(\frac{\ln(y)-(1-\sigma^2)}{\sigma}\right)$.[19]

One particular feature of this method is that, in a second phase, it ensures that the characteristics of the synthetic sample exactly match the reported values (e.g. the same income share by decile as the reported ones).[20] With this procedure, the average relative income of 100 observations (percentiles) is obtained.[21]

Shorrocks and Wan (2009) have shown that this technique is capable of reproducing individual data from grouped statistics with a high degree of accuracy. As an illustration, Figures A1–A3 compare various inequality measures (Gini, MLD, and T-Theil) estimated from aggregate data at the maximum level of disaggregation in the WIID (the full set of deciles and bottom and top 5 per cent) and their values estimated directly from microdata in the LIS sample (net per capita income). It is clear that the estimates are very accurate, particularly for the Gini index, but also largely for MLD and Theil indices (Table A2).

Table A2: Correlation between reported and estimated inequality measures, LIS sample

Type	Gini	MLD = GE(0)	Theil = GE(1)
Linear	100	98.5	98.9
Rank	100	97.8	98.5

Note: reported values are computed using full microdata; estimated values are computed using synthetic distributions obtained with the Shorrocks–Wan method.

Source: author’s construction based on the WIID.

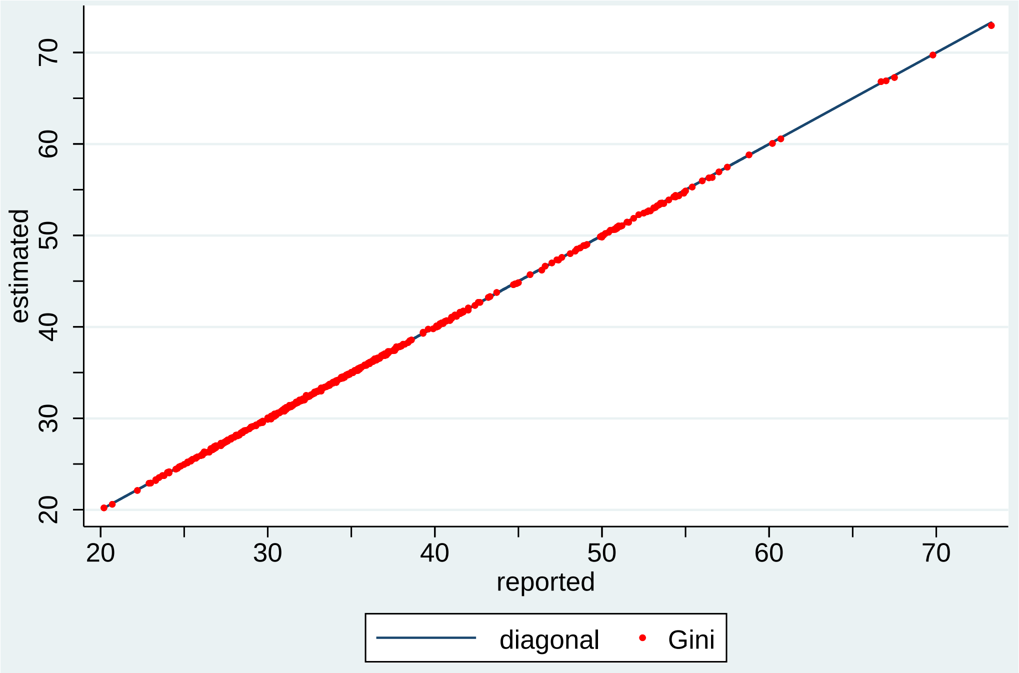
[19] As in the original paper, estimations using the beta Lorenz curve as an alternative are problematic because of a large proportion of observations with negative values.

[20] There might be differences, however, in the presence of negative values in the original distribution. However, in income distributions these tend to be rare. In the case of obtaining negative values in the synthetic distribution, these have been replaced by arbitrarily low values, facilitating the computation of inequality measures that are only defined on positive incomes, as well as for the sake of cross-country comparability.

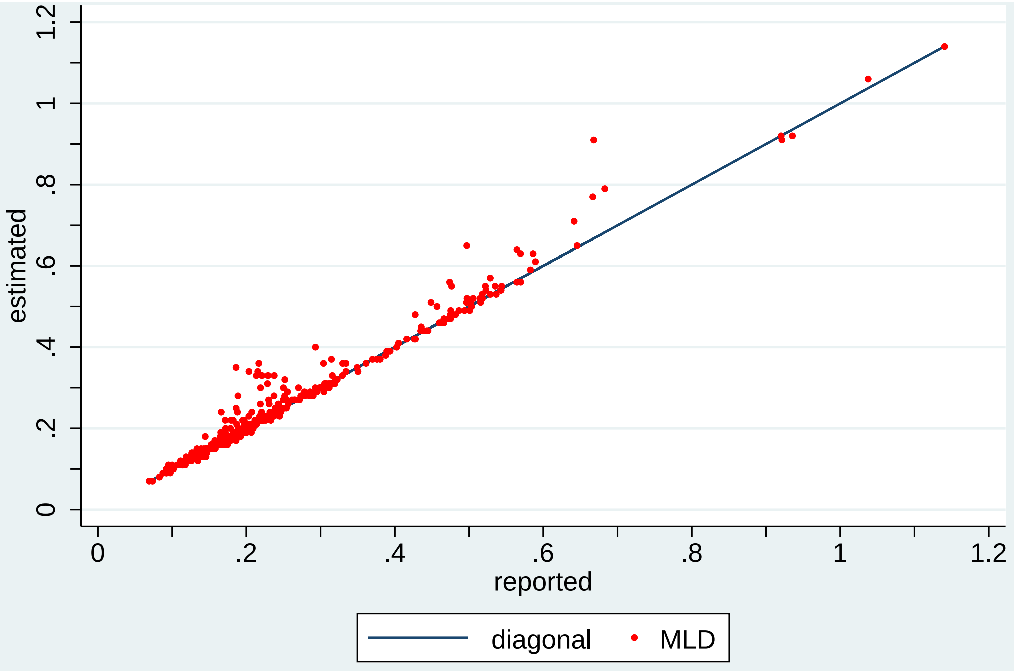
[21] For better accuracy and for practical reasons, a larger number of observations (10,000) was estimated, later aggregated to 100 in each country–year observation.

Figure A1: Measures of inequality: reported versus estimated from synthetic income distributions (LIS sample)

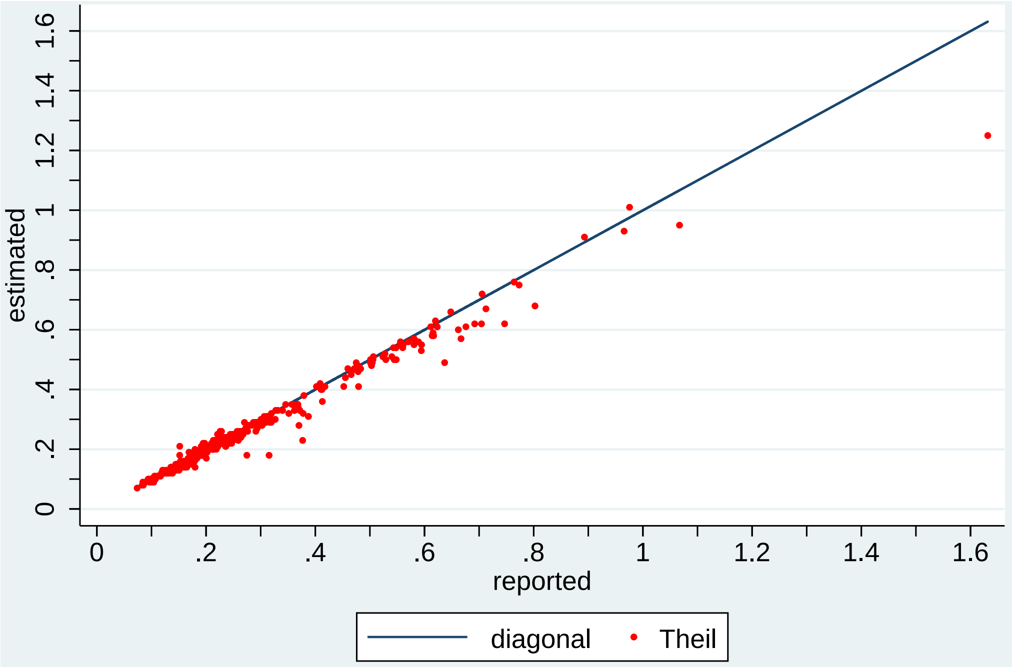
(a) Gini



(b) MLD



(c) Theil



Note: reported values are computed using full microdata; estimated values are computed using synthetic distributions obtained with the Shorrocks–Wan method.

Source: author's construction based on the WIID.

Appendix B: LIS sample

The LIS sample used here has 3,926 country–year observations for the entire income distribution and the Gini index (Table A3), of which 403 are for the target welfare concept (per capita net income) and the remaining 3,426 refer to other welfare concepts (621 for consumption per capita, for instance, or 349 for gross income per household). In each regression the latter are paired with the corresponding value for per capita net income estimated for the same distribution. Each country–year observation is weighted so that the sum of weights for each country is 1 (to avoid countries with more year observations contributing disproportionally to the results).

In this weighted sample, the average value for per capita net income tends to be higher than for consumption (a differential of 3.7 Gini points) and lower than for gross income (−2.7), as could be expected, but the difference varies depending on whether consumption and gross income are originally expressed in other equivalence scales, in which case the differential accounts for both difference in resource and in scale (Table A4). Figure A2(a–c) maps the corresponding Gini values for per capita income and per capita gross income and consumption estimated for the same distributions, highlighting the high degree of heterogeneity across observations, which can only be partially reduced by accounting for region and income group as done in the regression described in this technical note.

Table A3: LIS sample composition by resource and scale

	Scale				Total*	Total**
	Per capita	OECD modified	Squared household size	No adjustment		
Income (net)	493	457	493	493	1,443	1,936
Income (gross)	349	315	349	349	1,362	
Consumption	157	150	157	157	621	
Total*	506	922	999	999	3,426	
Total**	999					3,919

Note: figures refer to the unweighted number of country–year observations. Cases in which gross income and net income Gini values were identical were removed from the sample. * Total number of observations in the regressions (i.e. excluding cases for per capita net income). ** Total number of LIS observations used, including per capita net income.

Source: author's construction based on the WIID.

Table A4: Average Gini values by welfare concept, LIS sample

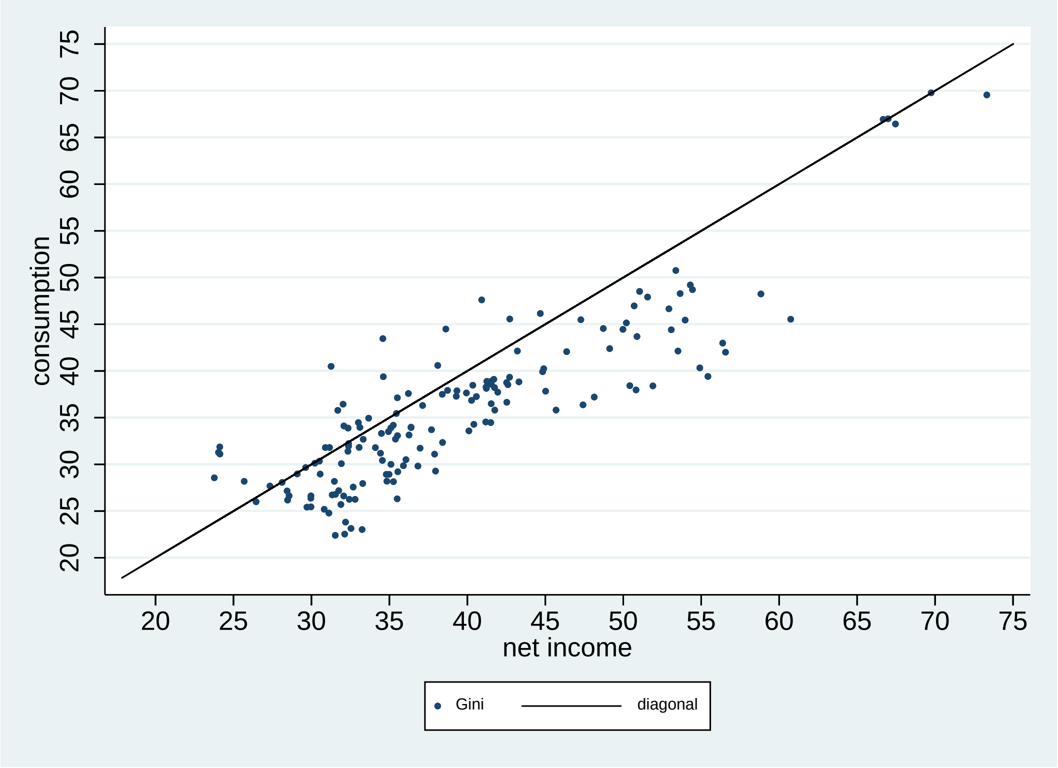
Sample for each equivalence scale	Gross income		Consumption	
	Average	Differential*	Average	Differential*
Per capita	39.2	2.7	36.7	-3.7
OECD-modified	37.2	0.7	34.2	-6.2
Square root	37.2	0.7	33.9	-6.5
No adjustment	42.0	5.5	37.7	-2.7
All scales	38.9	2.4	35.6	-4.8
Per capita net income	36.5		40.4	

Note: * differential between each Gini index and the one for per capita net income. Observations weighted so that each country has the same weight. Country–year observations are weighted so that the sum of weights is 1 for each country. Each average is computed on observations with information for net income and the other resource (gross income or consumption, respectively).

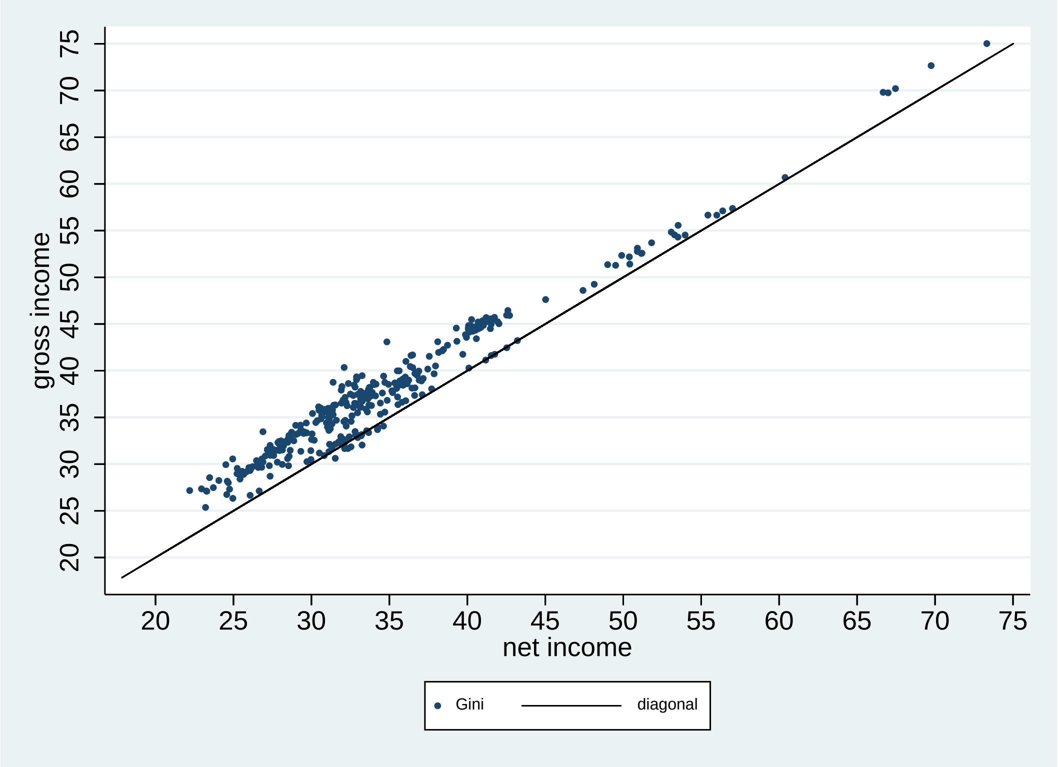
Source: author's construction based on the WIID.

Figure A2: Comparing different welfare concepts in the LIS sample (cases with information for net income and consumption, or net income and gross income)

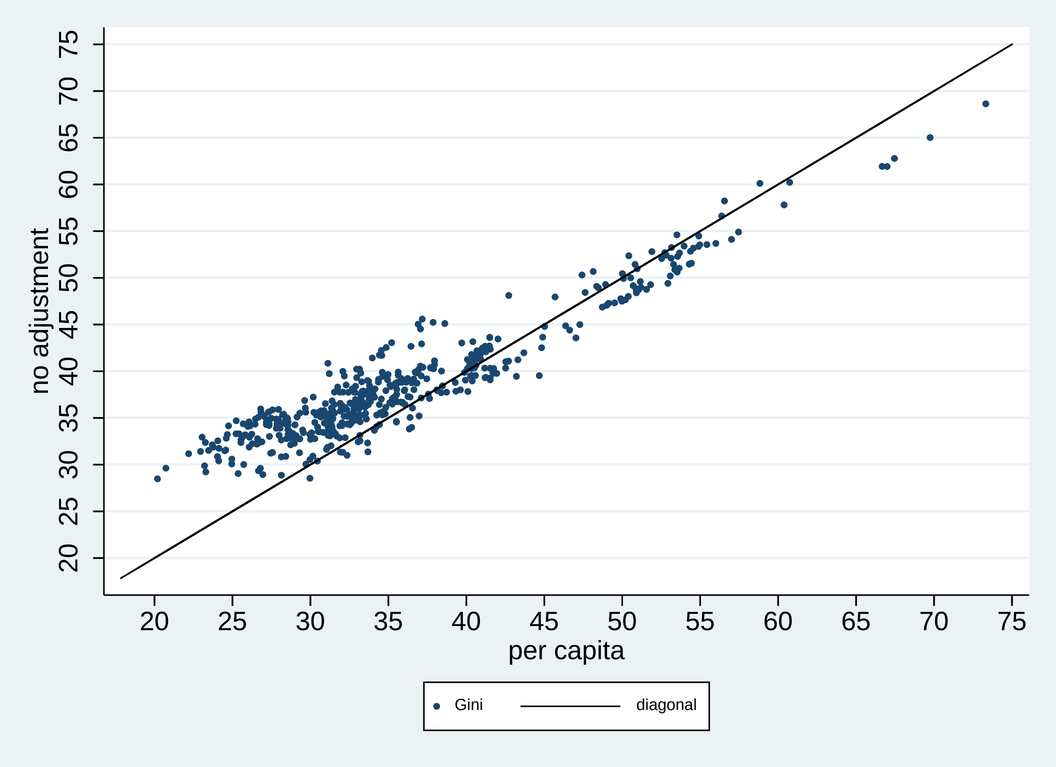
(a) Net income versus consumption (per capita)



(b) Net versus gross income (per capita)



(c) Per capita versus total household (net income)



Source: author's construction based on the WIID.

As is well known, the LIS sample is strongly biased in favour of high-income countries (Table A5). However, it is possible to take advantage of the increasing availability of middle-income countries and a few low-income countries to make the conversions. That means that the estimates for low- and middle-income countries will heavily rely on the limited available information for countries in the same region and/or income group. There is a much better representation of upper-middle-income countries in the LAC region (Brazil, Colombia, the Dominican Republic, Guatemala, Mexico, Paraguay, and Peru) and in Europe and Central Asia (Georgia, Russia, Serbia). In other cases, information is more limited. For example, for SSA, there is information about two low-income countries (Sudan and Somalia), one lower-middle-income (Côte d'Ivoire), and one upper-middle-income (South Africa). Final values for most other countries in the region will be converted using information on these sampled countries. Similarly, India will be used for South Asia (lower-middle), and China (upper-middle) or Vietnam (lower-middle) for East Asia. The same applies to Egypt, Tunisia, and the West Bank and Gaza (lower-middle) and Iraq and Jordan (upper-middle), which will be used for MENA countries and for Turkey.

Alternatively, the LIS sample could be expanded to use information from sources other than LIS. Although this could be explored in future versions, the advantage of only using LIS is the high quality of estimates using exactly the same surveys, which reduces the risk that the empirical gap observed between inequality in different resources and scales in a country in one year is due to other unobserved factors. This is especially relevant if one compares values obtained from different sources. Consider, for example, the problems involved in comparing the Gini index for per capita income obtained from LIS in Côte d'Ivoire and the same for consumption as obtained by PovcalNet, as previously discussed. The observed distributive gap between two different sources may reflect not only the difference in the resource indicator used to measure household economic capacity, but also different methods used by both institutions (surveys, corrections for price differences across rural and urban areas, correction for outliers or non-response, etc.). There are only a few values in WIID in which the inequality is observed in the same country and year for different inequality concepts, and the information comes from the same source (other than LIS).

Table A5: LIS sample composition by country region and income groups

	Income group				
Region	High	Upper-middle	Lower-middle	Low	Total
North America	323				323
Latin America and the Caribbean	82	292			374
Europe and Central Asia	1863	28			1891
Former Soviet countries (non-EU)		91	66		91
Middle East and North Africa	209	69	21	7	344
Sub-Saharan Africa		55	14		83
South Asia			14		14
East Asia and the Pacific	274	18	115		306
Total	2,751	553	66	7	3,426

Note: figures refer to the unweighted number of country–year observations (observations with per capita net income excluded).

Source: author’s construction based on the WIID.

Appendix C: Inequality summary measures

If $y=\left(y_1,y_2,\ldots,y_n\right)$ represents the income distribution of a country with *n* groups of equal size (e.g. 100 percentiles), each one representing one-*n*th of the population and $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ is the mean income of the same population, then:

Relative measures

Gini index

$$Gini\left(y\right)=\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{y_i-y_j}{\overline{y}}\right)=\frac{n+1}{n}-\frac{2}{n}\sum_{i=1}^{n}\left(\frac{n+1-i}{n}\right)\frac{y_i}{\mu}$$

General entropy

$$GE_\alpha\left(y\right)=\begin{cases}\frac{1}{\alpha\left(\alpha-1\right)}\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\overline{y}}\right)^{\alpha}-1\right] & \text{if }\alpha\neq 0,1\\ \frac{1}{n}\sum_{i=1}^{n}\ln\frac{\overline{y}}{y_i} & \text{if }\alpha=0\\ \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\mu}\ln\frac{y_i}{\overline{y}} & \text{if }\alpha=1\end{cases}$$

where GE_0 is also known as the mean log deviation (or M-Theil index); GE_1 is known as the L-Theil index; and $GE_2=\frac{1}{2}CV^2$, where CV is the coefficient of variation:

$$CV\left(y\right)=SD\left(y\right)/\,\overline{y}$$

Atkinson

$$A_\varepsilon\left(y\right)=\begin{cases}1-\left[\sum_{i=1}^{n}\frac{1}{n}\left(\frac{y_i}{\overline{y}}\right)^{1-\varepsilon}\right]^{\frac{1}{1-\varepsilon}} & \text{if }\varepsilon>0;1-\varepsilon\neq 1\\ 1-\prod_{i=1}^{n}\left(\frac{y_i}{\overline{y}}\right)^{\frac{1}{n}} & \text{if }\varepsilon=1\end{cases}$$

Absolute measures

Absolute Gini index

$$AbsGini\left(y\right)=\overline{y}Gini\left(y\right)=\frac{n+1}{n}-\frac{2}{n}\sum_{i=1}^{n}\left(\frac{n+1-i}{n}\right)y_i$$

Standard deviation

$$SD(y) = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\overline{y}} - 1\right)^2}$$