WIDER TECHNICAL NOTE | 7/2021

World Income Inequality Database (WIID)

WIID companion (May 2021): data selection

Carlos Gradín*

May 2021

wider.unu.edu

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). This technical note describes the first stage in constructing the new version (31 May 2021) of the companion datasets, i.e. the data selection process. It provides an overview of the approach followed by the selection of the series from different sources with information on income distribution and inequality that best represent each country and period. It also discusses the general criteria used and their implementation, which are illustrated with a few country examples.

Key words: selection, series, observations, 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-data-selection

Related publications:

Gradín, C. (2021). 'WIID Companion (May 2021): Integrated and Standardized Series'. WIDER Technical Note 2021/8. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-8

Gradín, C. (2021). 'WIID Companion (May 2021): Global Income 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/2021-7

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

When measuring inequality across countries and over time, the estimation of levels and trends can be affected by data choices (see a discussion on this matter by Atkinson and Brandolini (2001, 2009)). This technical note is dedicated to a detailed and transparent description of the selection of series in constructing the WIID Companion, among those available in the UNU-WIDER Income Inequality Database (WIID).[1] This note updates a previous version that refers to the March 2021 release of the data (Gradín 2021a). The necessary steps to integrate these series, which differ across welfare concepts and other relevant characteristics, to make them consistent over time and across countries are discussed in another technical note (Gradín 2021b).

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 for 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.

For example, the distributive information reported by the Luxembourg Income Study (LIS) for the Brazilian population between 2006 and 2016 for per capita net income (total household income divided by the household size) based on the Pesquisa Nacional por Amostra de Domicílios (PNAD) survey is a series, while the same information for net income per equivalent adult (where household income is divided by the square root of the household size instead) is another series. Thus, broadly, series for a given country can vary across source, resources, scale, geographic or population coverage, time period, survey, and more, depending on which inequality was calculated. All these aspects were considered when making data choices, which is described in more detail below. The selection was done on a case-by-case basis, but keeping in mind a few general selection criteria.

As stated earlier, there may be a variety of series available in the WIID to describe inequality in a country; in other cases, there is only one. Some examples are shown below before discussing the selection process in more detail.

2 Example of series in the WIID

A user interested in assessing inequality trends in Afghanistan, for example, will find information for three years in the WIID: 2008, 2012, and 2017. Information for those years refers to per capita consumption,[2] measured at the national level for the entire population, obtained from the European Union and the national statistic authority (NSA)—that is, the Central Statistics Organization of Afghanistan—based on the Living Conditions Survey (LCS). That is, all three observations (rows in the WIID Excel or Stata files, summarized in Table 1) are from the same

[1] Dataset, version March 2021, in UNU-WIDER (2021a); see user guide in UNU-WIDER (2021b), and see Pelanteri (2021) for a description of the main sources.

[2] Total household consumption divided by the number of household members. Household is the sharing unit, and the person is the reference unit.

series as earlier defined. Therefore, one could expect a high level of internal consistency in how observations were obtained.

With only one series (with one observation per year), Afghanistan provides an example for which further selection was not necessary when constructing the WIID Companion. This series refers to consumption and, for comparability, the WIID Companion will represent inequality of per capita net income and therefore will need to be adjusted in the next stage.

Table 1: Available series in the WIID: Afghanistan

Series	Year	No. year observations	Resource	Scale	Area coverage	Source
1	2008, 2012, 2017	3	Consumption	Per capita	All	NSA

Source: author's construction based on the WIID.

In other cases, observations for a country are obtained from different series. If they overlap over time, there will be cases with more than one observation stemming from multiple series to represent inequality in specific years.

For example, there are eight series available for Angola (as shown in Table 2). Seven of them provide some information for 2009, while only one series includes information for 1995 and another one for 2001 and 2019. They differ across resource, area coverage, survey, and source, as well as across the reported measures. For instance, series number 3 displays measures for 2009 using income (net/gross),[3] making it different from series 4, which uses consumption, even if both refer to the same source (Angola Instituto Nacional de Estatistica), scale (per capita) and area coverage (the entire country) and survey (Integrated Survey on the Welfare of Population).

From a user's perspective, it would thus require going through the available alternatives and selecting those most adequate for the purpose of their analysis. If interested in the Gini index, one could prefer using the estimate from the PovcalNet series, which refers to per capita consumption across the entire population (urban and rural). But one could also prefer to use the Gini index estimated by the NSA—which refers to consumption as well—or a different resource, such as per capita income in the country in rural areas or in urban areas. Which series and respective observations are the best often depends on the purpose of the user's analysis (e.g. interest in specific years, comparing with other countries, investigating differences across areas in the same country). However, for many purposes, users are often interested in obtaining the longest possible series for a given country. In the case of Angola, the PovcalNet series for per capita consumption provides the longest country-level trend (spanning from 2001 to 2019).

Other aspects to consider when selecting a series are whether they report the Gini index only or also income or consumption shares, considering that these can be reported at different levels of detail (deciles in the PovcalNet series; quintiles in the NSA series; more limited in the case of the research study), or what resources one is more interested in. In many cases there will be trade-offs between the various preferences to ponder in our choice. In this case, the PovcalNet series was selected.

[3] This indicates that it is not clear or well defined whether income is expressed in net or gross terms. It is possible, for example, that the source only says 'income', or that the survey questionnaire requests gross income but there is the general belief that respondents in the formal sector report their take-home income (while there is no distinction between net and gross in the informal sector).

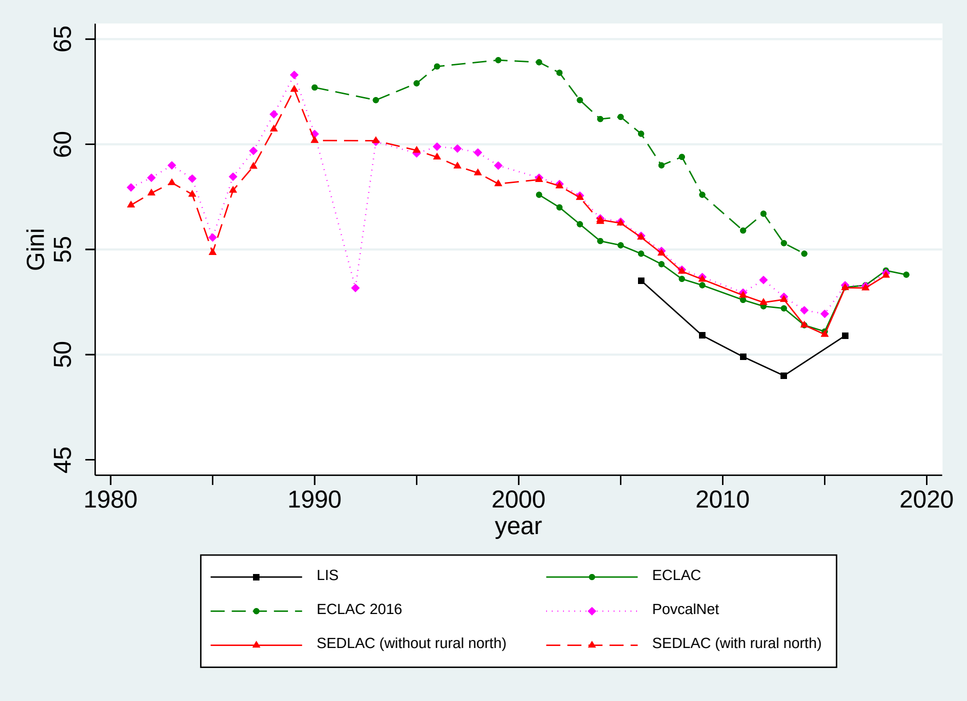
Table 2: Available series in the WIID: Angola

Series	Years	Observations	Aggregate measure	Resource	Scale	Area coverage	Source
1	1995	1	Gini, d1 and d10	Income (net/gross)	Per capita	Urban	Research study
2	2001, 2009, 2019	3	Gini, d1–d10	Consumption	Per capita	All	PovcalNet
3	2009	1	Gini, q1–q5	Income (net/gross)	Per capita	All	NSA
4	2009	1	Gini, q1–q5	Consumption	Per capita	All	NSA
5	2009	1	Gini, q1–q5	Income (net/gross)	Per capita	Rural	NSA
6	2009	1	Gini, q1–q5	Consumption	Per capita	Rural	NSA
7	2009	1	Gini, q1–q5	Income (net/gross)	Per capita	Urban	NSA
8	2009	1	Gini, q1–q5	Consumption	Per capita	Urban	NSA

Source: author's construction based on the WIID.

Other countries with richer information can be more complex than the presented case, with multiple options to reasonably represent the trend in inequality for specific periods. For example, Figure 1 displays different available series of Gini in per capita income in Brazil in recent years obtained from Economic Commission for Latin America and the Caribbean (ECLAC), Socio-economic Database for Latin America and the Caribbean (SEDLAC), PovcalNet, and LIS, with many years overlapping. In the case of Brazil, the WIID Companion will use a combination of series to cover this period: LIS first, complemented by ECLAC and SEDLAC in this order.

Figure 1: Available series in the WIID (per capita income): Brazil



Note: the LIS series is for 'income net', the others are for 'income net/gross'. All displayed series are per capita.

Source: author's construction based on the WIID.

3 Data selection criteria in the WIID Companion

The selection used to build the WIID Companion was based on series rather than on single observations. Choosing entire series or fragments of series that are necessary to cover specific periods rather than individual observations enables us to generate a more consistent approach. Various series may differ in many aspects, some of them not well documented in the variables included in the WIID. For example, as reflected in Figure 1 for Brazil, the old and new series for ECLAC for Latin American and the Caribbean (LAC) have different Gini values (the same applies to income shares), even when they have the same key variables (resource, scale, etc.). This is due to some methodological changes undertaken by ECLAC that, for example, were used to correct incomes for their underestimation compared with the National Accounts. Similarly, ECLAC and SEDLAC series for LAC may reflect different levels of inequality due to methodological choices in producing their estimates. The trends among these two sources are in general highly consistent, even if not identical.

For the selection of series, the construction of the WIID Companion did not follow any explicit algorithm (a mechanical sequence in which predefined preferences are applied), which imposes the same criteria for each country in the selection process. Instead, by using guiding principles for the selection process, each country was analysed separately to decide which series best fulfils these principles. While one series might represent the best option for one country and period, it may not be the best for other countries and/or periods.

For example, while LIS may be a good series to represent inequality in many countries, and will be the preferred source in general where it is available, it was excluded in the cases of the Dominican Republic, Guatemala, and Romania as in these cases other series may be a better reference, as explained below. Hence, the availability of series for each country was evaluated considering the selection criteria across them.

The selection process was followed in a sequence. That is, first, the main series for the country was identified. The series was then extended backwards and forwards with other series (or fragments) until the maximum time period was covered. The purpose was to cover the longest possible period, not necessarily using all survey years available for a country if that would imply mixing different series over a similar period, raising issues of year-to-year comparability.

In the following, we describe the set of selection criteria used in the construction of the WIID Companion. One overarching purpose is to provide rich information about the distribution of net per capita income among the entire population residing in a country for the longest possible period, with maximum comparability over time and across countries. Therefore, information that is as close as possible to that definition (or can be converted to that with the minimum possible manipulation) was given priority. Table 3 summarizes the key criteria, while their implementation is discussed in the next section.

Overall, via a targeted selection process, the WIID Companion includes about 11 per cent of the more than 20,000 observations available in the WIID (i.e. 2,384 observations). To facilitate the identification of WIID Companion observations in the WIID database, a new variable, *wiidcompanion*, was added that takes a value of 1 if the observation is part of the WIID Companion. This and the fact that the *id* variable that uniquely identifies observations is the same in the WIID and in the WIID Companion allow users to easily identify with precision the origin of all observations and the characteristics of each series or fragment of a series included in the WIID Companion in terms of the reported information before any adjustment to integrate a country series takes place.

Table 3: Summary of the main selection criteria to construct the WIID Companion

Source	LIS was generally taken as the main source, whenever available. Other sources that were given high priority (depending on the time period covered) are some regional sources (ECLAC and SEDLAC for LAC; Eurostat for EU countries), as well as national sources (NSA), PovcalNet, and other World Bank sources. Research studies have also been considered, especially when reporting long-term series. Some old compilations very often linked to international organizations (ECLAC, World Bank, Asian Development Bank) have been considered, especially for the earliest periods.
Population and geographic coverage	Largest possible geographic and population coverage. Only in some countries and years partial information was used—that is, for some or all urban areas (e.g. several LAC countries), excluding specific areas of a country (e.g. rural Amazonas in Brazil) or population groups (e.g. underrepresentation of blacks in South Africa, exclusion of immigrants in Oman, Qatar, and Saudi Arabia) due to lack of better information.
Resource	Income, particularly net income if available; income (net/gross), gross income, or consumption otherwise, depending on availability. Earnings have been excluded in this selection process.
Scale	Per capita, otherwise equivalized or with no adjustment for household size.
Distributional information	Preference for series with more detailed information about income shares (by deciles or, at least, quintiles). Only a few observations do not report the Gini index.

Source: author's construction

4 Implementation of selection criteria: data sources

4.1 LIS

In the selection process, the maximization of international comparability was prioritized. At first, if available, priority was generally given to the LIS harmonized data. These series are directly estimated from microdata through LISSY. For consistency with other sources, the year in LIS observations included in the current version of the WIID refers to the survey year (the last one if between two years). This is unlike LIS microdata, which generally uses the income reference year instead (in some cases, for example, it refers to the calendar year or the previous 12 months before the interview takes place). For example, the database for Spain 2004 in LIS microdata, based on the Survey on Income and Living Conditions 2005, will be relabelled as Spain 2005 in the WIID, while Spain 1990, based on the Household Budget Survey 1990/91, will be relabelled as Spain 1991 in the WIID.

The main advantage of LIS with respect to other sources is its great consistency across countries (57 in the current version) and over time because of the post-harmonization process of samples, key variables, and definitions, carefully undertaken by the LIS team. The main limitation is the limited time and geographical coverage. Initially, the LIS series only included high-income economies (now 35), but it has undertaken an effort to progressively extend the coverage to middle-income countries (14 upper-middle and 6 lower-middle are now present), and 2 low-income countries (Somalia and Sudan).[4] Therefore, LIS has a very good representation of European countries, the United States, Canada, and Australia. LIS also includes Russia and other former USSR countries, several LAC countries, particularly Chile and Mexico, but also Brazil, Colombia, the Dominican Republic, Guatemala, Panama, Paraguay, Peru, and Uruguay. LIS series include several MENA (Middle East and North Africa) countries, mainly through the series harmonized by the Economic Research Forum (ERF-LIS), including Egypt, Iraq, Israel, Jordan,

[4] A discussion of the challenges and solutions involved in incorporating middle or low-income countries to LIS, see discussion in Checchi, Cupak, and Munzi (2021).

Tunisia, and the West Bank and Gaza. LIS also includes Côte d'Ivoire, Somalia, South Africa, and Sudan (ERF-LIS series) in the SSA (sub-Saharan Africa) region, and China, India Japan, the Republic of Korea, Taiwan, and Vietnam in Asia.

Regarding the time frame, the longest LIS series are for countries like the United Kingdom (1969–2019), Canada (1972–2017), Germany (1973–2017), and the United States (1975–2019). In LAC countries, the longest series are for Mexico (1984–2018) and Chile (1990–2017). Other key series tend to be shorter, such as that for Brazil (2006–16), India (2005–12), and China (2003–14), increasing the need to use other sources to complete the time series.

In the selection process, some LIS series were excluded: the Dominican Republic, Guatemala, and Romania. LIS has only one observation for the Dominican Republic, where ECLAC has a substantially longer series (the level of the Gini index seems to be different in both series). The LIS series for the Gini index of per capita net income (2006, 2011, and 2014) in Guatemala shows a discontinuity between 2011 and 2014, with a large drop from 55.4 in 2011 to 45 in 2014 that is not matched by other sources (e.g. fall from 51.4 to 48.3 with SEDLAC, or from 55.8 in 2006 to 53.5 in 2014 with ECLAC 2019). ECLAC was also used as the main series in this case. In the case of Romania, LIS only has data for 1995 and 1997, while Eurostat allows us to consistently cover a longer and more recent period.

In the case of Egypt, the WIID Companion uses the ERF-LIS series for 2000–15 (based on the Household Income, Expenditure and Consumption Survey (HIECS)) and does not use the LIS single value for 2012 (based on the Egypt Labor Market Panel Survey (ELMPS)). The latter, with a Gini index for per capita net income of 54.6, seems to be inconsistent with the former (33.1 in 2011 and 31.9 in 2013).

4.2 Regional sources: ECLAC, SEDLAC, Eurostat

As secondary sources, for countries and periods in which LIS is not available or deemed appropriate, some regional sources are then used.

ECLAC and SEDLAC are used for LAC countries. There are two main different series for ECLAC in the WIID. The oldest (ECLAC 2016) has been discontinued and the newest one launched in 2019. The WIID preserves the older series because not all observations can be replaced by the new one, and they are substantially different. ECLAC undertook important methodological changes that make both series distinct from each other (for example, regarding the imputation of income underestimation). ECLAC is the main series used for LAC countries when LIS was not available, but SEDLAC was the main series in Guatemala, and also an important source of information for other countries, especially Argentina, Brazil, Costa Rica, Panama, Uruguay, and Venezuela, complementing ECLAC.

For European countries (including Malta, part of MENA in the World Bank's regional classification), the WIID Companion also uses Eurostat series. Most series are based on the European Union Statistics on Income and Living Conditions (EU-SILC) and the European Community Household Panel (ECHP), with values estimated from microdata. Reported values by Eurostat on its website are also used for countries or years for which these could not be obtained from microdata, particularly Turkey and North Macedonia, but also Bulgaria, Cyprus, Lithuania, Malta, and Romania. Eurostat series also include two observations from the European Commission 2006 (Croatia and Cyprus in 2003).

4.3 National statistical authorities

In other cases, the WIID Companion uses series provided by NSAs when well-established series are available. This applies to countries such as the United States (United States Census Bureau) or Canada, but also Bulgaria (Bulgaria National Statistical Institute), Bangladesh (Bangladesh Bureau of Statistics), China (National Bureau of Statistics of China), Jamaica (Planning and Statistical Institutes of Jamaica), New Zealand (Ministry of Social Development), Pakistan (Pakistan Bureau of Statistics), Sri Lanka (Sri Lanka Department of Census and Statistics), Taiwan (Taiwan Directorate General of Budget, Accounting and Statistics) and a few others.

4.4 World Bank

In many developing regions, the World Bank series constitutes the main available source, given the absence or insufficiency of the other sources, particularly in SSA and South Asia. The majority of World Bank series that have been selected in the WIID Companion come from PovcalNet, which can be retrieved from the website or using the built-in Stata module, but a significant amount of information also comes from related historical repositories or studies, like Deininger and Squire (2004, unpublished), Jain (1975), and Fields (1989) for Brazil, or the India Database, among others. These are particularly relevant for the earliest years in many cases, for which other sources are unavailable. Part of the information provided by PovcalNet is redundant due to other sources included in the WIID Companion (e.g. LIS or SEDLAC) and so is not used.

4.5 Other research studies

Other research studies, whether independent or linked to international organizations, are especially important for the earliest years in LAC countries (e.g. Farne (1994) for Chile and Paukert (1973) for several countries, among many others), Europe and Central Asia (e.g. Brandolini (1999) for Italy; Alexeev and Gaddy (1993), Atkinson and Micklewright (1992), Milanovic (1998), and Milanovic and Ying (1996) for eastern countries), Africa (e.g. Gradín and Tarp (2019) for Mozambique; Lachman and Bercuson (1992) for South Africa; and Paukert (1973) for various countries), and East Asia (e.g. Ravallion and Chen (2007) for China; Mizoguchi and Takayama (1984) for Japan).

4.6 Other sources

Other sources used in the WIID Companion include other UN series, mainly from UNICEF for Bulgaria, North Macedonia, Poland, and some former USSR countries (Georgia, Kazakhstan, Lithuania, Tajikistan, and Turkmenistan), Altimir (1986) for Argentina, or the International Labour Organization (ILO 1984) for some SSA countries, along with the Asian Development Bank (ADB) series (including Dowling and Soo (1983) for China) and the Inter-American Development Bank (1999, 2016) for Suriname and Barbados.

As a result, a larger share of observations in the WIID Companion comes from PovcalNet (26 per cent) and LIS (20 per cent), but the sources are highly fragmented because around 13 per cent of observations come from different research studies, and another 11 per cent from various NSAs, for example, highlighting the difficulty of obtaining the majority of consistent series from the same source (Table 4).

Table 4: Data sources in the WIID Companion, by region (percentage observations)

	North America	LAC	Europe and Central Asia	MENA	SSA	South Asia	East Asia	All
LIS (microdata)	59.3	11.5	30.6	33.3	3.8	2.0	9.4	20.5
LIS	59.3	11.5	30.6	16.0	3.0	2.0	9.4	19.4
ERF-LIS				17.4	0.8			1.1
Eurostat			22.1	11.1				9.5
Microdata			18.3	9.0				7.8
SEDLAC		14.1						2.9
United Nations		44.6	4.2	3.5	1.9		1.4	11.6
ECLAC		43.2						9.0
UNICEF			3.8					1.2
NSA	38.3	7.1	7.2	6.9	1.5	34.3	23.8	11.2
OECD							1.4	0.2
World Bank		8.7	25.8	33.3	77.8	57.6	41.0	31.1
PovcalNet		5.8	24.6	28.5	66.5	27.3	33.2	26.0
Research studies	2.5	13.5	10.1	11.8	14.7	4.0	20.5	12.4
Other international Organizations		0.4			0.4	2.0	2.5	0.6
Total	100	100	100	100	100	100	100	100

Note: regions are North America, Latin America and the Caribbean, Europe and Central Asia, Middle East and North Africa, sub-Saharan Africa, South Asia, and East Asia and the Pacific.

Source: author's construction based on the WIID.

5 Implementation of selection criteria: coverage

Generally, priority was given to the largest possible geographic and population coverage. In most cases, the series in the WIID Companion refer to the whole-country population, including urban and rural areas in all regions, and nationals and immigrants. There are a few cases (3.6 per cent; Table 5) in which the WIID Companion uses information for the urban population. This is because for some Latin American countries earlier observations only refer to urban areas (Argentina, Bolivia, Chile, the Dominican Republic, Ecuador, Panama, Paraguay, Uruguay).[5] The geographic coverage of household surveys in these countries has been progressively extended from representing only the capital or a few cities (e.g. the main 15 cities) to reach all urban areas or the entire country in the most recent years. Only in Argentina, a highly urbanized country, the area coverage remains urban even in recent years. In the next stage, these cases were adjusted so final values are representative of the entire country or, in the case of Argentina, the entire urban population (even if the trend was determined by the available coverage).

Also, in a few cases (1 per cent) small parts of the territory are excluded in a few years, such as rural areas in the north of Brazil in the earliest years, West Irian, or Maluku in Indonesia (1964 and 1970), nomadic areas in Iraq (2004), or non-continental Portugal (1973, 1980, and 1990).

Lastly, almost all series apply to all population groups. Exceptions include a few observations in Oman, Qatar, or Saudi Arabia, reporting only information about nationals, therefore excluding the large immigrant communities. In the case of South Africa, although the corresponding population coverage variable indicate that information refers to the entire population, it is well known that

[5] Urban observations are also used in Eritrea for 1997 and Suriname for 2004.

during apartheid the population of African descent was heavily underrepresented in any official statistics. In this latter case, the scale of inequality can be partially corrected in the next stage, while keeping the reported trend.

The WIID Companion does not include any series that only apply to rural populations or small regions/areas of a country.

Table 5: Original population and area coverage in the WIID Companion (percentage observations)

	Geographic area			
Population	All	'Almost' all**	Urban	Total
All	95.3	0.9	3.6	99.9
Other*	0.2			0.2
Total	95.5	0.9	3.6	100

Note: * Qatari, Omani, Saudi. ** Except specific country areas in Brazil (rural north), Indonesia (West Irian or Maluku), Iraq (nomadic areas), and Portugal (islands) in specific years.

Source: author's construction based on the WIID.

6 Implementation of selection criteria: resource

Another important criterion was to use income as a resource measure, although information on consumption is also used whenever consistent information on income is not available (Table 6). In this first version of the WIID Companion, no series based on individuals' earnings as a resource was considered. Although income inequality is to a large extent the result of earnings inequality, information on earnings is typically based on a different rationale, referring to a particular sub-group of a population—the active workforce—not considering the household as sharing units, being a concept more closely related to labour market dynamics than the general distribution of income among a population (e.g. Peichl and Pestel 2014). The integration of information on earnings, although feasible, is not straightforward.

Among income series, priority was given to those referring to net income (that had been calculated after discounting direct taxes and social contributions) but in some cases also gross income (before such discounts) or consumption is used. Series with the ambiguous label of 'net/gross' were also used, considering that the concept is likely to be close to net income in practice.

Some sources—such as Eurostat, SEDLAC, ECLAC, UNICEF, OECD, or the US Census Bureau—only provide information on income; LIS provides information on both (but the majority of series refer to income), while World Bank series can refer to either income or consumption. Income tends to be the most common resource except for South Asia and SSA, where consumption is predominant. In some cases, it was necessary to combine information on income and consumption for the same country over time. For instance, earlier series for SSA, MENA, or Eastern Europe are based on income, while later observations utilize consumption. In other cases, most series primarily display former measures based on consumption, while information on income has become available more recently. This applies, for example, to India, Egypt, and Côte d'Ivoire, with income distributions provided by LIS.

As reported in Table 6, 69 per cent of all observations in the WIID Companion refer to different income concepts (net, 38; net/gross, 22, gross, 10), the other 31 per cent to consumption, reflecting the regional asymmetries discussed above.

Table 6: Original resource concept by region in the WIID Companion (percentage observations)

Region	Income				Consumption	Total
	Net	Net/gross	Gross	All Income		
North America	59.3		40.7	100		100
Latin America and the Caribbean	14.3	71.0	5.8	91.1	8.9	100
Europe and Central Asia	65.2	5.0	7.0	77.2	22.8	100
Middle East and North America	37.5	5.6	3.5	46.6	53.5	100
Sub-Saharan Africa	11.3	8.3	7.9	27.5	72.6	100
South Asia	2.0	12.1	17.2	31.3	68.7	100
East Asia and the Pacific	27.4	20.2	20.2	67.8	32.1	100
Total	38.6	21.5	10.3	70.4	29.7	100

Source: author's construction based on the WIID.

7 Implementation of selection criteria: equivalence scale

The priority was to use series in which income or consumption are expressed in per capita terms (Table 7). That is, series that take the person as the reference unit but also consider that people share resources within their households (household or family as the sharing unit). Per capita income is used to adjust for household size and is obtained by dividing the total household amount by the number of household members. This implies assuming there are no economies of scale in cohabiting, and is the most common practice, particularly in developing countries.

Table 7: Original equivalence scale by region in the WIID Companion (percentage observations)

Region	Per capita	Equivalized	No adjustment or missing	Total
North America	59.3	9.9	30.9	100
Latin America and the Caribbean	82.9	0.4	16.7	100
Europe and Central Asia	83.2	7.3	9.5	100
Middle East and North	81.2	3.5	15.3	100
Sub-Saharan Africa	82.0	0.4	17.7	100
South Asia	63.6	0.0	36.4	100
East Asia and the Pacific	60.4	10.3	29.4	100
Total	77.9	5.1	17.0	100

Source: author's construction based on the WIID.

Some of the most important sources tend to report per capita values, such as PovcalNet from the World Bank and most UN sources such as ECLAC or UNICEF; SEDLAC provides information on per capita and equivalized values, while other sources, such as LIS or EU-SILC, allow obtaining per capita estimates directly from microdata, along other scales. When official Eurostat estimates are used instead, they tend to use the modified OECD equivalence scale (where the number of equivalent adults is obtained by weighting as 1 the first adult, 0.5 the rest of the adults, and 0.3 those 14 or younger). Equivalence scales that account for the economies of scale of cohabiting are also quite common in many research studies, especially of industrialized countries. The OECD and some NSAs also report estimates using the square root of household size as the equivalence scale. Equivalized income values were also given consideration whenever per capita values were not available. Finally, some cases will use total household income or consumption (with no adjustment for household members), given that these are common among some older series and several NSAs.

As Table 7 indicates, 77 per cent of all observations in the WIID Companion refer to per capita values, 17 per cent to non-adjusted total household values, and only 5 per cent to different equivalence scales (mainly modified OECD and square root). Once again, the expected regional asymmetries emerge.

After combining resource and scale (Table 8), the most common welfare concepts turn out to be per capita net income (30 per cent), per capita consumption (29 per cent), and per capita income (net/gross) (17 per cent), as well as no adjusted (total household) gross income (8 per cent).

Table 8: Original 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	49.5	4.8	0.1	2.8	0.2	1.6	15.9	70.2
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				0.3	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	17.3	100

Note: Other: OECD, supplemental poverty measure, unknown equivalence scale.

Source: author's construction based on the WIID.

8 Implementation of selection criteria: distributional information

Many series in the WIID report information on income shares. Some series provide richer information (deciles and top and bottom 5 per cent), others only report information by quintiles, while several provide different combinations of income shares. The priority was to use series with more detailed information about the distribution. Income shares are considered when at least the complete set of quintiles is available—ideally, the full set of deciles and top and bottom 5 per cent. Doing so allows us to estimate the entire synthetic distributions at the percentile level within countries, which are used to compute a variety of inequality measures. These can also be used for further analysis, particularly to facilitate aggregation across countries to estimate the global income distribution in further stages. As a result (Table 9), almost 90 per cent of the observations in WIID Companion report a complete set of income shares. The most common case reports deciles (with or without the bottom and top 5 per cent), accounting for 80 per cent of the total, with another 10 per cent reporting at least quintiles (with or without the bottom and top 5 per cent).

There are only a few cases in the WIID, mostly from NSAs and research studies, in which only income or consumption shares are reported, and neither the Gini index nor other measure of inequality are reported. These series were also given consideration because inequality measures can be estimated based on the reported income shares (see Gradín 2021b). The only cases without a reported Gini included in the WIID Companion (0.6 per cent of all observations) are a few observations from various NSA series for Azerbaijan, Brunei, Fiji, Japan, Nauru, Oman, and

Singapore, and from research studies for Eritrea (Arneberg and Pedersen 2001) and Kuwait (Al-Qudsi 1981).

Table 9: 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

Source: author’s construction based on the WIID.

9 Implementation of selection criteria: quality considerations

When analysing inequality over time and across countries, it must be noted that there is a necessary trade-off between the quality and amount of information used regarding the length of series, number of countries included in the comparison, and other factors. The construction of the WIID Companion is no exception.

The WIID has a variable for quality. It is worth mentioning that this variable displays a subjective evaluation during the selection process in WIID’s history over more than 20 years and hence does not necessarily use the same criteria consistently across series. However, it has been widely used as a reference for the quality of the data. According to this variable, roughly 10 per cent of all observations in the WIID dataset are labelled as ‘low’ or ‘unknown’ quality (Table 10). However, this applies disproportionately to low-income countries, with 21 per cent of observations from low-income countries being of ‘low’ or ‘unknown’ quality. Furthermore, this proportion increases when looking at country observations before 1980, amounting to 59 per cent for countries overall and nearly all observations for low-income countries.

Hence, removing all observations labelled as low quality, although tempting, would leave us with only a small set of observations for earlier years and a strong bias regarding countries, regions, and income groups, whereby high-income countries would be overrepresented. It is not at all clear that any explicit (or implicit by omission) imputation of these missing observations in making inferences about inequality trends, levels, determinants, effects, etc. would be of higher quality than the removed observations. It is hence up to the user to make informed choices about the inclusion of earlier years or countries with poor data in their analysis where quality of information tends to be lower. Ultimately, any user can filter observations in the WIID Companion using the mentioned quality variable or their own assessment.

Table 10: ‘Quality’ variable in the WIID and the WIID Companion (percentage observations)

	High		Average		Low/unknown		Total
	WIID	Companion	WIID	Companion	WIID	Companion	
Total	73.7	51.3	16.6	35.5	9.7	13.2	100
By region							
North America	93.9	96.3	4.7		1.3	3.7	100
Latin America and the Caribbean	69.1	75.2	16.3	9.5	14.6	15.3	100
Europe and Central Asia	82.1	58.8	13.2	33.0	4.7	8.2	100
Middle East and North Africa	74.3	54.2	12.8	31.2	12.8	14.6	100
Sub-Saharan Africa	31.1	6.8	37.0	71.4	32.0	21.8	100
South Asia	16.3	18.2	42.1	68.7	41.7	13.1	100
East Asia and the Pacific	51.1	28.8	32.1	52.6	16.8	18.6	100
By income Group							
High	84.8	77.0	10.2	11.7	5.0	11.4	100
Upper-middle	58.3	43.1	25.2	41.0	16.5	15.9	100
Lower-middle	41.1	22.7	35.7	65.5	23.2	11.9	100
Low	32.2	10.2	44.6	75.0	23.2	14.8	100
By year							
Before 1980	20.7	21.9	20.8	23.5	58.5	54.6	100
After 1980	79.2	56.6	16.2	37.7	4.6	5.5	100

Source: author’s construction based on the WIID.

A second quality issue to consider is that the selected series may have outliers. The presence of outliers or dubious trends also played a role in the selection process of series. The general principle applied in the construction of the WIID Companion was, however, to avoid removing them if they are part of the selected series.

Another quality issue refers to the heterogeneity of the information selected in the WIID Companion, particularly, but not only, in terms of the welfare concepts. Another technical note in this series (Gradín 2021b) describes how the reported information was adjusted to facilitate more consistent comparisons over time and across countries.

References

Alexeev, M., and C. Gaddy (1993). ‘Income Distribution in the U.S.S.R. in the 1980s’. *Review of Income and Wealth*, 39(1): 23–36. https://doi.org/10.1111/j.1475-4991.1993.tb00435.x

Al-Qudsi, S. (1981). ‘Income Distribution in Kuwait’. *Journal of the Social Sciences*, 18–31.

Altimir, O. (1986). ‘Estimaciones de la distribución del ingreso en la Argentina, 1953–80’. *Desarrollo Económico*, 25(100): 521–66. https://doi.org/10.2307/3466844

Arneberg, M.W., and J. Pedersen (2001). *Urban Households and Urban Economy in Eritrea*. Oslo: Fafo Institute for Applied Social Science.

Atkinson, A.B., and A. Brandolini (2001). 'Promise and Pitfalls in the Use of 'Secondary' Datasets: 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

Atkinson, A.B., and A. Brandolini (2009). 'On Data: A Case Study of the Evolution of Income Inequality Across Time and Across Countries'. *Cambridge Journal of Economics*, 33(3): 381–404. https://doi.org/10.1093/cje/bel013

Atkinson, A.B., and J. Micklewright (1992). *Economic Transformation in Eastern Europe and the Distribution of Income*. Cambridge: Cambridge University Press.

Brandolini, A. (1999). 'The Distribution of Personal Income in Post-War Italy: Source Description, Data Quality, and the Time Pattern of Income Inequality'. *Giornale degli Economisti e Annali di Economia*, 58: 183–239.

Dowling, J.M. Jr., and D. Soo (1983). 'Income Distribution and Economic Growth in Developing Asian Countries'. Asian Development Bank Economic Staff Paper 15. Manila: Asian Development Bank.

Farne, S. (1994). 'Apertura Comercial y Distribución del Ingreso: La teoría y las Experiencias de Chile, México, y Uruguay'. *Universitas Economica*, 9(1): 71–104.

Fields, G.S. (1989). 'A Compendium of Data on Inequality and Poverty for the Developing World'. Ithaca, NY: Department of Economics, Cornell University.

Gradín, C. (2021a). 'WIID Companion (March 2021): Data Selection'. WIDER Technical Note 2021/4. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-4

Gradín, C. (2021b). '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., and F. Tarp (2019). 'Investigating Growing Inequality in Mozambique'. *South African Journal of Economics*, 87(2): 110–38. https://doi.org/10.1111/saje.12215

Inter-American Development Bank (1999). 'Integration and Regional Programs Department Datasheets'. Washington, DC: Inter-American Development Bank.

Inter-American Development Bank (2016). *Barbados Living Conditions Survey*. Washington, DC: Inter-American Development Bank.

International Labour Organization (ILO) (1984). *Rural–Urban Gap and Income Distribution (A Comparative Sub-Regional Study): Synthesis Report of Seventeen African Countries*. Addis Ababa: ILO, Jobs and Skills Programme for Africa.

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

Lachmann, D., and K. Bercuson (eds) (1992). 'Economic Policies for a New South Africa'. IMF Occasional Paper 91. Washington, DC: International Monetary Fund. https://doi.org/10.5089/9781557751980.084

Milanovic, B. (1998). *Income, Inequality, and Poverty during the Transition from Planned to Market Economy*. Washington, DC: World Bank.

Milanovic, B., and Y. Ying (1996). *Notes on Income Distribution in Eastern Europe*. Washington, DC: World Bank.

Mizoguchi, T., and N. Takayama (1984). *Equity and Poverty under Rapid Economic Growth: The Japanese Experience*. Tokyo: Institute of Economic Research, Hitotsubashi University.

Paukert, F. (1973). 'Income Distribution at Different Levels of Development: A Survey of Evidence'. *International Labour Review*, 108(2): 97–125.

Peichl, A., and N. Pestel (2014). 'Earnings Inequality'. IZA Policy Paper 89. Bonn: Institute for the Study of Labor (IZA).

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

UNU-WIDER (2021a). 'World Income Inequality Database (WIID)'. 31 May 2021 version. 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