WIDER TECHNICAL NOTE | 9/2021

World Income Inequality Database (WIID)

WIID Companion (May 2021): global income distribution

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). This technical note describes the construction of the global distribution companion datasets, version 31 May 2021. While previous technical notes described the selection of country income distribution series and the integration and standardization process to overcome the heterogeneity in original welfare concepts and other methods, I describe here all the necessary additional steps and assumptions made to construct the global dataset.

Key words: global income distribution, inequality, WIID, WIID Companion

JEL classification: C82, D31, E01, O15

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-global-income-distribution.

https://www.wider.unu.edu/publication/wiid-companion-may-2021-global-income-distribution

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): 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). '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/2021-9

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.

1 Introduction

The main objective of the global dataset presented here is to estimate an annual series, from 1950 to 2019, of the global distribution of per capita net household income across all people in the world, regardless of the country of residence. This annual series should be constructed in such a way that it is possible to produce regular updates and revisions whenever new or better information becomes available.

The detailed global distribution is made of distributions representing 100 percentile groups of each country in the world every year. These are based on the UNU-WIDER Income Inequality Database (WIID) and, more particularly, on the WIID Companion cross-country dataset, an unbalanced panel of the synthetic distributions of per capita net income at the percentile level and the corresponding inequality measures for all available countries, with information for the longest possible period.[1] The original information comes primarily from PovcalNet, the Luxembourg Income Study (LIS), Eurostat, research studies, and historical compilations, among other sources.

I obtained this cross-country dataset in two stages. First, making a selection of the country series from the WIID that best represents each country's income distribution (as described in Gradín 2021b). Second, given that the original information contained in these series vary across welfare concepts and other methodological aspects, the WIID Companion uses the minimum necessary adjustments to integrate these series in a consistent way over time and across countries (as explained in Gradín 2021c). The country series used to estimate global income distributions are also included in the dataset to allow better tracking of the source of any distributional changes that are observed globally.

These country income distributions are finally aggregated at the global percentile level (representing 1 per cent of the world population), including mean incomes and income expressed as a share of total income. It also reports various distributional measures, including most common relative inequality indices (such as Gini, the general entropy family, and the Atkinson family), absolute inequality measures (absolute Gini and the standard deviation), aggregate income shares of different population groups (such as the bottom 40 or top 10 per cent), and ratios based on them (such as the Palma ratio or S20S80). The aggregation combines within-country income distributions at the percentile level with information about their populations, as estimated by UNDESA, while an integrated series of per capita GDP from various sources is used as a proxy for countries' mean incomes.

To give additional insights to better understand global inequality levels and trends, the dataset also estimates the income distribution by a country's geographical region and by income groups, and provides estimates of the between-country and within-country components of global inequality for each summary measure.

In this technical note, I explain the necessary steps to aggregate across countries the information on the synthetic income distributions in the WIID Companion (cross-country dataset) to produce the global distribution and various components of it. This note updates a previous version that refers to the March 2021 release of the data (Gradín 2021a). The construction of such a dataset necessarily involves strong assumptions to tackle the missing information. This is done here in a

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

transparent and simple way, with explicit assumptions, producing a distribution that is fully consistent with what is observed at the country level. Stata codes that allow replicating the steps described here are included with this technical note, while the necessary codes for previous steps are also released with the corresponding technical note.

The following sections are focused on the necessary steps:[2]

1. determining the list of countries for which annual information is going to be provided;
2. constructing an annual series of country populations and mean incomes;
3. constructing an almost balanced panel of annual within-country synthetic distributions; and
4. aggregating the income distributions across countries into the global income distribution and estimating various global distributive measures.

2 Country list

The first issue that needs to be addressed is to determine the list of countries that make up the global distribution. There were significant changes in country borders during the studied period, mostly after the collapse of communist regimes in Eastern Europe, for which distributive information is rather poor. In constructing this global dataset I use current existing countries, completing their information for earlier periods, when they integrated other states, if necessary, with the trend based on the original entities. For example, Czechoslovakia, Yugoslavia, or the USSR for former socialist countries, Ethiopia for Eritrea, or Sudan for South Sudan.[3]

The database contains information for 209 economies (see Table A1 in Appendix A), including some autonomous territories often listed in international databases. Of these, 186 economies have distributive information (income percentile shares) for at least one year.

3 Population and per capita income

Our main references for determining the population of a country in a specific year, as in the WIID Companion, are the estimates and projections made by UNDESA for total population (both sexes combined) by country, reported annually since 1950, in thousands (UNDESA 2019). In the few

[2] 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), 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).

[3] Due to the lack of distributive information on the German Democratic Republic in the WIID, to cover the period before unification I use in this case the combined information assigned to Germany for population (UNDESA) and mean income (Maddison Project), and the distributive information from the Federal Republic of Germany.

cases in which the necessary information was not obtained from this source, it was completed with other sources.[4]

The construction of a global dataset also requires annual information on country mean incomes. This in turn will determine where each country percentile falls in the global distribution when aggregating across countries, therefore affecting between-country absolute and relative inequalities in general, as well as within-country absolute inequalities.

There are two main approaches to estimating country mean incomes: (1) using a macro aggregate from National Accounts, like per capita GDP or GNI; or (2) using the survey per capita income or consumption. The main problem of the latter is that the information on survey income in the WIID is incomplete: not all country–year observations with income shares also report mean income. Furthermore, the mean income or consumption typically comes in local currency and with different reference periods, needing several adjustments that should take into account monetary reforms, evolution of exchange rates, inflation rates, etc., for which information is also incomplete. Considering that I am reconstructing an annual series for global income shares even for non-survey years, it seems to be more appropriate to use a macro aggregate because this information is estimated more widely and consistently.

For that reason, to collect information for as many countries as possible since 1950, I integrate information primarily coming from the per capita GDP series expressed in US$2017 PPP (purchasing power parity), recently published by the World Bank (2021), with complementary trends adjusted from other sources. These are the Maddison Project Database (2020) (multiple benchmarks, version November 2, 2020) and the Penn World Tables (real GDP per capita expenditure-side at chained PPPs, PWT 9.1, Feenstra et al. 2015), both originally expressed in 2011 PPP instead. GDP is used as a proxy for income, given it is more widely available than GNI.

Therefore, the reference is GDP per capita based on PPP (constant 2017 international dollars: NY.GDP.PCAP.PP.KD) obtained from the World Development Indicators (World Bank 2021) for 1990 onwards. This is defined in the metadata spreadsheet in the following way:

> PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the country plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2017 international dollars.

The integration of the different series for earlier years (1950–89) or for countries for which information in the World Development Indicators (WDI) is incomplete is done by taking advantage of overlaps with other series for most countries. Therefore, I incorporate the relative trend observed in the other sources, while keeping the scale observed in the WDI. This allows us to integrate the two series even if they are originally expressed in different PPPs. The Maddison series is rescaled to match the WDI series in 1990, adjusting all values before 1990 using the same ratio. The same applies to the PWT series whenever a country is not represented in the Maddison

[4] As described in the WIID user guide, for example, for some historical entities we take the values from the French Institute for Demographic Studies (INED), for Kosovo from the WDI, and for the West Bank and Gaza from its statistical authority.

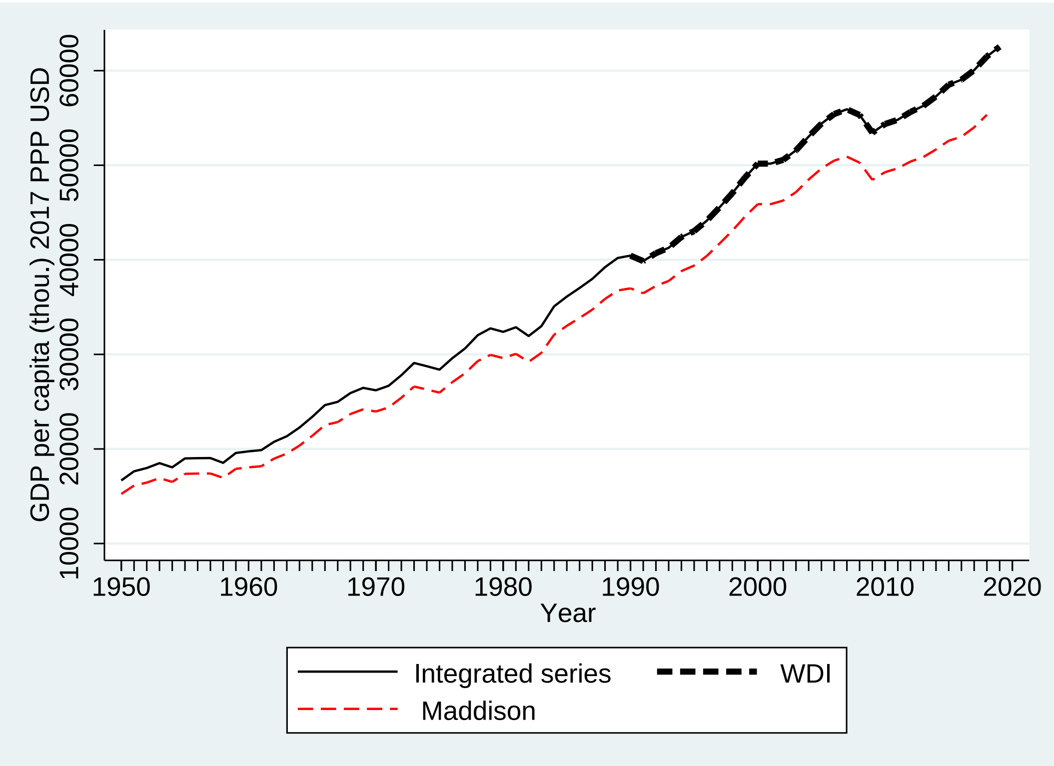
Project either (mainly small islands or territories like the Cayman Islands, Fiji, the Maldives, and others).

In the case of countries with no information at all in the WDI series, but with information in the Maddison Project (e.g., Cuba, North Korea,[5] Syria, Venezuela, Yemen), since the previous exercise cannot be done, I alternatively use the ratio between the GDP for the target country and that for the United States in 1990 in both sources to adjust the Maddison Project's values to be comparable to those in the WDI. The same applies for some countries, again mainly small islands, with information in the PWT but not in the WDI or the Maddison Project. In the case of countries with some years still missing, we impute the same growth rate as in their corresponding region and income group.[6]

As a result, we have information for the period 1950–2019 about per capita income and population for all countries listed in Table A1 in Appendix A. Figure 1 shows examples of the integrated series.

Figure 1: Examples of GDP per capita integrated series

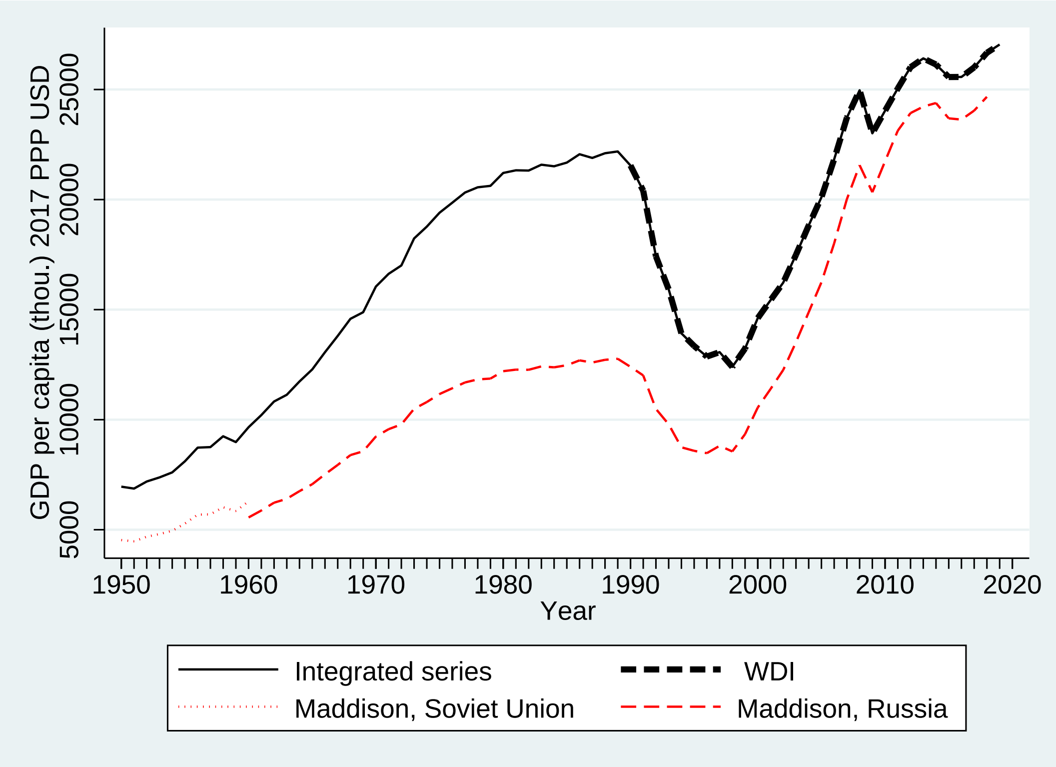
(a) United States (WDI since 1990; Maddison Project before 1990)



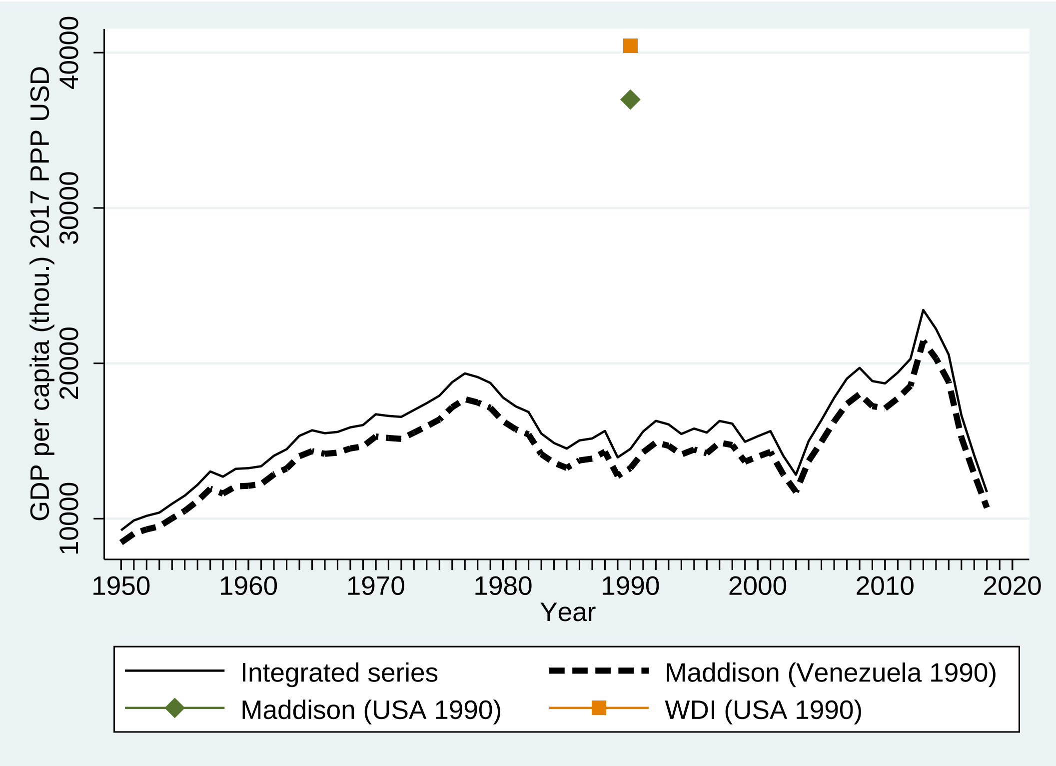
[5] In the case of North Korea, I interpolate per capita income between 1943 and 1990 (with almost constant income), give the discontinuity in the Maddison Project series.

[6] Fiji (before 1960); Bhutan, Brunei, Macao, and various Caribbean islands (before 1970); Guyana, North Korea, and various Pacific islands (before 1990); San Marino (before 1997); Palau and Timor-Leste (before 2000); Kosovo (before 2003); Nauru (before 2004). In the case of the Maldives (the only country in its region and income group), GDP is kept constant before 1970 instead. In the cases of Curacao, Kosovo, Liechtenstein, the Marshall Islands, and Monaco, the lack of information before 1960 is replaced, keeping constant their information in 1960. Similarly, the information for Anguilla and Montserrat is extended from 2017 to 2019.

(b) Russia (WDI since 1990; Maddison Project for Russia (1960–89) and Soviet Union (before 1960))



(c) Venezuela (Maddison Project); United States 1990 is used as a reference for adjustment



Source: author's construction based on the WDI and the Maddison Project.

4 Panel of within-country synthetic income distributions

The most important issue to address in the construction of a global database is the lack of distributive information for many country–year observations, given that this information is scarce and irregular in many cases.

The global dataset is based on the cross-country dataset in the WIID Companion. This dataset estimates synthetic income distributions at the percentile level based on aggregate income shares originally reported in the WIID, mostly at least at the decile level (in some cases at the quintile level), which may include or not the top and bottom vintiles.[7] These income distributions originally referred to different welfare concepts (such as per capita net income, per capita consumption, total household gross income) and were made consistent over time and across countries using a two-stage regression-based standardization process explained in Gradín (2021b). Table A1 reports the initial and final years of information for synthetic income distributions for all listed countries in the WIID Companion global dataset, as well as the number of survey years used.

One possibility to cope with this issue is to use only countries with information available each year (or in the closest years falling in specific time bins defined ad-hoc, e.g. ±2 years away from the comparison year). This would significantly reduce the number of countries, especially in early years, and would produce a highly unbalanced panel, with varying composition, unless the missing information is imputed in some way. Otherwise, the implicit assumption is that countries that are missing in the sample have the same distribution as those included, and this in a context in which the set of omitted countries changes over time. This strategy is perfectly feasible with the WIID Companion, using the cross-country dataset and defining what time bin will be used. However, with the dataset we also provide an alternative solution based on using the information from the same country in other years to impute the missing country–year income distributions.

The solution adopted here is to linearly interpolate the income percentile shares for each country between two available survey years (which implies assuming uniform income changes between both years if the distribution has changed). Income distribution before the earliest survey observation and after the latest are extrapolated (kept constant). That is, the dataset will be a balanced panel, with information for all countries and years. In some cases, the distributive information comes from a survey conducted during the same year; in the majority of cases it is inferred from the nearest surveys.

In the few cases in which we have information for per capita income and population, but the country lacks entirely any information on the within-country distribution, this cannot be inter- or extrapolated.[8] As an alternative, I assigned the population-weighted average per capita income distribution (percentile income shares) in the same geographical region and income group (using the World Bank's classifications). Therefore, the country will still contribute to the between-

[7] Note that about 10 per cent of country–year observations in the WIID Companion cross-country dataset refer only to the Gini index and not to the synthetic distributions. This is due to the lack of information on income shares at the decile or quintile level. Therefore, these observations are not used in the construction of the global dataset. This includes all the observations of Réunion, Saint Kitts and Nevis, Vincent and the Grenadines, San Marino, Saudi Arabia, Turks and Caicos Islands.

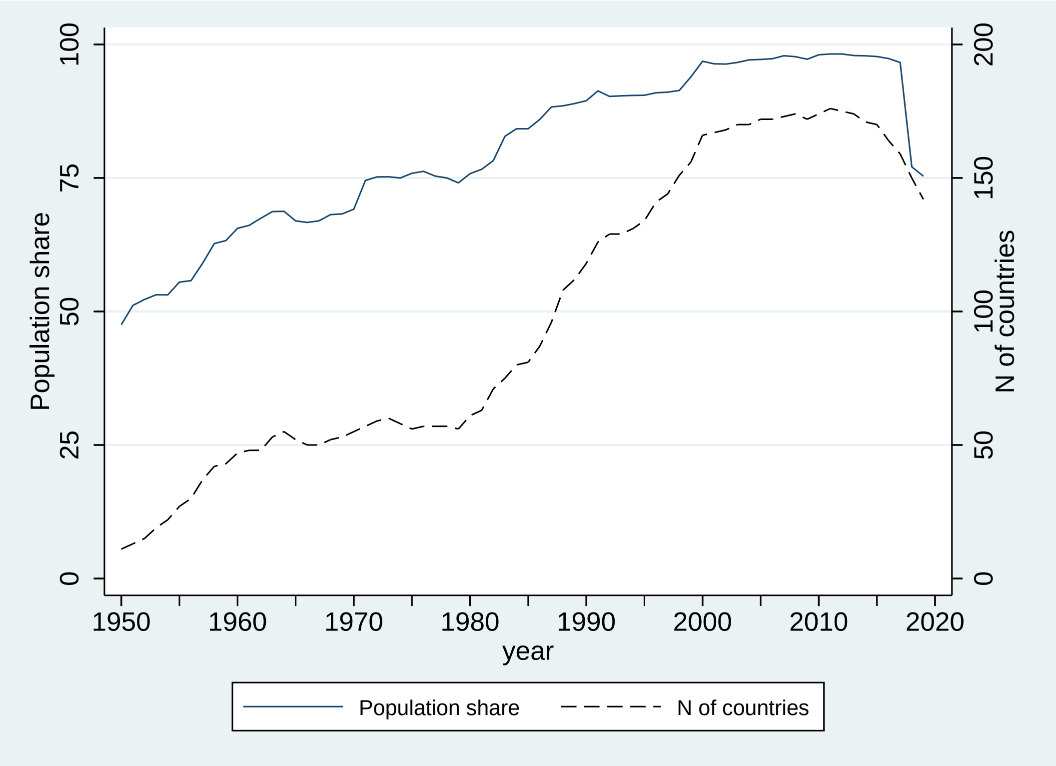
[8] These are countries with no information on income shares in the WIID Companion in any year (some have information only for the Gini index, as previously explained, others are not part of the country dataset). They include Bahrain, Libya, Qatar, and Saudi Arabia in the Middle East and North Africa (MENA) region, as well as, Macao and the Marshall Islands in East Asia and Pacific, San Marino in Europe, and several Caribbean islands (Bermuda, Anguilla, etc.). See Table A1 in Appendix A.

country inequality level and trend (since I use its own mean income and population), but will reflect the region/income group within-country inequality component.[9]

A variable (*interpolate*) in the corresponding file indicates whether the estimates come from a survey year, are interpolated, extrapolated, or imputed. This variable can be used to end up with the dataset with information for only survey years, as in the cross-country dataset.

To give a sense of the extent to which the distributive information for any year comes from a recent survey year, Figure 2 displays for each year the number of countries and the percentage of the contemporary world's population with a survey observation falling within a bandwidth of ±5 years (e.g. number of countries and population share in 2000 with survey observations falling in the time span between 1995 and 2005).[10] Figure 3 displays, for observations falling in those time bins, the population-weighted average gap between each year and the closest survey year (e.g. the gap will be 2 if the closest observation around 2000 is in either 1998 or 2002).

Figure 2: Number of countries and world's population share with distributive information around each year using a ±5-year bandwidth



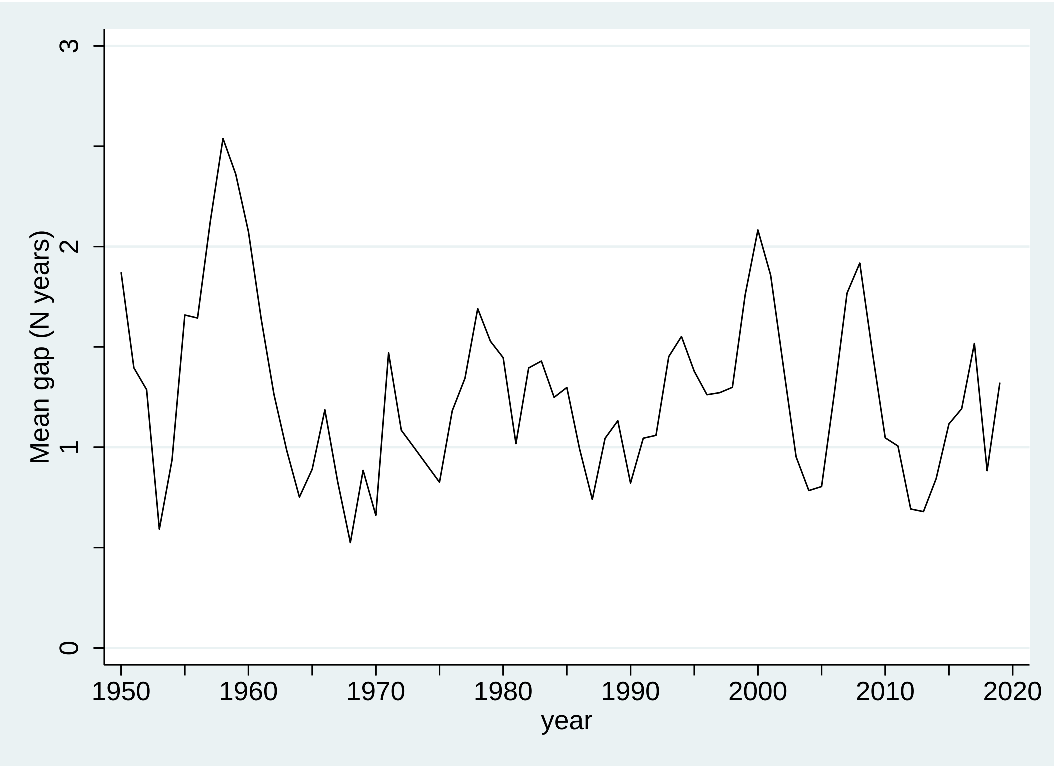
Note: number of countries and share of the world population with information on income shares, falling within a bandwidth of five years below and above the target year.

Source: author's construction based on the WIID Companion.

[9] After this step, the only country in Table A1 that remains with information for mean income and population, but no distributive information, is North Korea. In this case I assigned the same mean income to all percentiles, so the country still contributes to inequality between countries.

[10] However, for the reasons discussed above, the database uses the nearest information of a country, without using a time threshold.

Figure 3: Population-weighted mean gap years between target year and closest survey year falling in a ±5-year bandwidth



Note: the population-weighted average gap in number of years between each target year and the closest survey income shares within a bandwidth of five years below or above the target year.

Source: author's construction based on the WIID.

It turns out that the initial number of countries with observations around 1950 is relatively small (11), but the world's population covered (or, equivalently, the population-weighted number of countries) is already 48 per cent, reflecting the fact that the most populous countries already have some survey information before 1955. Before 1950, the WIID Companion has information for percentile income shares only for the United States (since 1947) and Italy (1948). The next countries that followed between 1950 and 1955 are Mexico (1950); India (1951); Barbados (1952); and Argentina, China, Cuba, Puerto Rico, Sri Lanka, and Taiwan (1953). Distributive information for the first sub-Saharan African (SSA) countries starts being added in 1958–65, with Chad (1958), Benin, Zambia, Côte d'Ivoire, and Nigeria (1959), Niger and Madagascar (1960), Senegal (1961), and Tanzania (1964).

The number of countries and the population coverage within the ±5-year bandwidth progressively rises, reaching 75 per cent of the population (59 countries) in 1972 and 98 per cent (174–76 countries, with the omitted being mainly small countries) around 2010–12. More countries have information in recent decades, particularly after 1990, indicating that also the quality of the within-country inequality estimates and trends is higher in the most recent decades. Examples of key late incorporations are Germany (1974), Indonesia and Brazil (1976), Iran (1986), Russia and Ukraine (1988), Vietnam and South Africa (1993), Ethiopia (1996), and Democratic Republic of the Congo (2005). In most recent years, however, the coverage of countries and population declines again due to the lack of updated information for many countries, included some populated ones like India.

The population-weighted average distance between a survey year and the target year falling within the corresponding ±5-year bin mentioned above oscillates between 0.5 and 2 years over most of the analysed period.

This heterogeneity in the availability of contemporaneous within-country distributions over time indicates that we need to keep in mind that although the database uses information for almost all countries in every year, a big part of the within-country distribution in the initial decades will be extrapolated from the earliest available information for each country. Inequality within those countries lacking a survey in the earliest years will remain constant until the first survey year. Thus, this approach underestimates the changes in inequality within countries over the first decades, particularly before 1980. However, it still reflects changes in between-country inequality for all countries, as explained earlier, given that mean income and population are updated annually. Alternatively, one could impute the trend in inequality for countries before the first survey is conducted based on information of similar countries or other predictors, or just focus one's attention on the most recent decades, when the representativity of the global population with contemporaneous distributive information is the highest.

5 Aggregation and computation of summary measures

Combining the information from income shares by percentiles with country population and mean income, I estimated the global distribution for every year since 1950. Each country is represented by 100 income points, indicating the per capita income in 2017 PPP of each within-country percentile group (1 per cent of the contemporary country population). These observations are weighted by the population they represent to construct the global distribution.

The entire global distribution (mean income of country-level percentiles) is then used to estimate the mean income and income share (of total global income) for different global fractiles (i.e. percentiles and vintiles).[11] These are further aggregated to report information for relevant groups, such as the bottom 40 per cent or the top 10 per cent. The global distribution is also used to estimate a battery of relative and absolute inequality indices (see Appendix B).[12] In this initial version, relative measures such as the Gini index, generalized entropy, and the Atkinson families are included, along with the Palma ratio and the S80S20 ratio. These are complemented with information for the absolute Gini index and the standard deviation as summary measures of absolute inequality. Other distributive indicators may be included in future versions, but can be calculated by users based on the reported information.

The dataset also reports various inequality estimates to assess the importance of inequalities within and between countries.

First, the dataset estimates each summary inequality measure in two counterfactual distributions:

---

[11] Note that each global percentile roughly represents around 1 per cent of the world's population. I assigned each country percentile (all of its population) to one global percentile and, as a consequence, some global percentiles may make up slightly less or more than 1 per cent of the global population.

[12] That is, the indices are estimated using the entire country-level distributions (weighted by their corresponding populations). Inequality measures (or any other distributional measures) can also be directly estimated using the global percentile distribution, even if this will slightly underestimate the level of global inequality.

1 *Between-country inequality*. A first counterfactual distribution indicates the global distribution resulting after all within-country inequalities were removed by replacing the income of all percentiles in every country by the country mean. It therefore reflects only inequality between countries. This is equivalent to estimating inequality measures on country mean incomes, with countries weighted by their populations. In the case of the Atkinson family, the corresponding 'equally distributed equivalent income' is used instead of the mean.

2 *Within-country inequality*. The second counterfactual distribution indicates the global distribution after all inequality between countries has been removed—that is, all country percentiles are rescaled to have the same representative income. The latter is the global mean income in general, but the equally distributed equivalent incomes in the case of the Atkinson family. Therefore, the inequality measures only reflect inequalities within-country.[13] In the case of members of the generalized entropy family, inequality within countries, as defined here, is equivalent to estimating the population-weighted average of country inequality indices. Since this is not the case of the Gini index and the Atkinson family, these population-weighted averages are also provided separately, as they can also be used to assess the importance of inequality within countries.

It is well-known that inequality indices have different decomposability properties. Only in one case, the mean log deviation (GE_0), overall global inequality is exactly the sum of the 'pure' between-country and within-country components, estimated as described above. In the case of other members of the same GE family, the difference between overall inequality and between-country inequality is often also referred to as the within-country component, but note that this component is different from the one defined above—that is, it is a weighted sum of country inequality, where the weights depend on country population but also on country mean incomes, and therefore are partially reflecting between-country inequalities. These are not reported directly in the dataset, but can easily be obtained by subtracting between-country inequality from overall inequality.

For that reason, as an alternative to comparing the importance of between-country inequalities across indices more consistently, I also estimated the percentage of inequality between countries to overall inequality using the Shapley approach, which guarantees that the within and between components add up to overall inequality in all indices (as described in Appendix C).

6 Final considerations

This technical note has described the necessary steps undertaken to construct the WIID Companion global dataset for the study of global inequalities from the within-country dataset. The approach followed here has been used to construct an almost balanced panel of countries and years with annual information since 1950, estimating the income distribution for each year by interpolating between the two closest survey years, or by keeping the distribution constant before the first and after the last survey year.

[13] To keep within-country inequality constant, I multiply each income by the ratio between the global and country means (or 'equally distributed equivalent income' in the case of Atkinson) in the case of relative measures, while adding the differential between the global and country means in the case of absolute measures.

Like the corresponding country income distribution, the global series attempts to primarily reflect the income distribution that would emerge from typical survey information, and has not been corrected for possible outliers or potential mis-estimation of incomes at the extremes of the distribution.

The approach followed here is to report a global dataset that is linked to detailed country-level information from which it was obtained, facilitating tracking any distributional pattern observed at the global level to the primary source at the detailed country level. The approach has also prioritized estimating the entire distribution. This has the advantage of not relying on a particular index (reflecting specific distributive views) to determine the nature of the observed distributional changes. Gradín (2021d) describes the main distributive trends using this database. As that evidence suggests, these distributional changes can be complex, and very often imply at the same time a combination of equalizing and disequalizing movements over time, involving different parts of the distribution. Thus, assessments over whether global inequality declines or increases will often depend on our views on inequality (absolute/relative, with more concern about the improvement of the poor versus about the concentration among the rich, etc.).

References

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

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/0.1257/jel.39.3.771

Chakravarty, S.R. (2009). *Inequality, Polarization and Poverty: Advances in Distributional Analysis*. Springer: New York.

Chantreuil, F., and A. Trannoy (2013). 'Inequality Decomposition Values: The Trade-Off between Marginality and Efficiency'. *Journal of Economic Inequality*, 11(1): 83–98. https://doi.org/10.1007/s10888-011-9207-y

Davies, J., and A. Shorrocks (2021). 'Comparing Global Inequality of Income and Wealth'. In C. Gradín, M. Leibbrandt, and F. Tarp (eds), *Inequality in the Developing World.* Oxford: Oxford University Press.

Feenstra, R.C., R. Inklaar, and M.P. Timmer (2015). 'The Next Generation of the Penn World Table'. *American Economic Review*, 105(10): 3150–82. https://doi.org/10.1257/aer.20130954

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): Global Distribution'. WIDER Technical Note 2021/6. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-6

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. (2021b). 'WIID Companion (March 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. (2021c). '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

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

Maddison Project Database (2020). Version November 2020. Based on J. Bolt and J.L. van Zanden (2020), 'Maddison Style Estimates of the Evolution of the World Economy: A New 2020 Update'. Maddison-Project Working Paper WP-15. Groningen: Groningen Growth and Development Centre, University of Groeningen.

Shorrocks, A. (2013). 'Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value'. *Journal of Economic Inequality*, 11(1): 99–126. https://doi.org/10.1007/s10888-011-9214-z

UNDESA (2019). *World Population Prospects 2019*, online edition rev. 1. New York: United Nations, Department of Economic and Social Affairs, Population Division.

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

World Bank (2021). 'World Development Indicators'. Available at: https://databank.worldbank.org/source/world-development-indicators (accessed 15 January 2021).

Appendix A: Available within-country distributions

Table A1: Initial and final years, and number of survey-year observations (*N*) with information for within-country distribution used to construct the global dataset

Region and country/territory	Distribution		
	Initial year	Final year	*N*
North America			
Bermuda	–	–	–
Canada	1966	2017	20
United States	1950	2019	56
Latin America and the Caribbean			
Anguilla	–	–	–
Antigua and Barbuda	–	–	–
Argentina	1953	2019	40
Aruba	–	–	–
Bahamas	1970	2013	3
Barbados	1952	2010	3
Belize	1993	1999	6
Bolivia	1968	2019	24
Brazil	1976	2019	31
British Virgin Islands	–	–	–
Cayman Islands	–	–	–
Chile	1968	2017	15
Colombia	1964	2019	15
Costa Rica	1969	2019	33
Cuba	1953	1953	1
Curacao	–	–	–
Dominica	2008	2008	1
Dominican Republic	1969	2019	27
Ecuador	1968	2019	25
El Salvador	1965	2019	26
Grenada	2008	2008	1
Guatemala	1979	2014	10
Guyana	1993	1998	2
Haiti	2001	2012	2
Honduras	1968	2019	20
Jamaica	1958	2015	29
Mexico	1950	2018	21
Montserrat	–	–	–
Nicaragua	1993	2014	6
Panama	1962	2019	25

Paraguay	1983	2019	15
Peru	1972	2019	15
Puerto Rico	1953	2003	7
Saint Kitts and Nevis	–	–	–
Saint Lucia	1995	2016	2
Saint Vincent and the Grenadines	–	–	–
Sint Maarten (Dutch part)	–	–	–
Suriname	1962	1999	2
Trinidad and Tobago	1958	1992	6
Turks and Caicos Islands	–	–	–
Uruguay	1961	2019	25
Venezuela	1981	2014	23
Europe and Central Asia			
Albania	1996	2017	9
Andorra	2003	2016	2
Armenia	1996	2018	20
Austria	1987	2019	11
Azerbaijan	1995	2018	8
Belarus	1988	2018	23
Belgium	1979	2019	23
Bosnia and Herzegovina	2001	2011	4
Bulgaria	1963	2019	46
Croatia	1988	2019	18
Cyprus	2005	2019	15
Czechia	1992	2019	10
Denmark	1976	2019	13
Estonia	1992	2019	10
Finland	1962	2019	16
France	1962	2019	17
Georgia	1996	2016	17
Germany	1973	2019	30
Greece	1957	2019	11
Greenland	2002	2018	17
Hungary	1962	2019	20
Iceland	2004	2018	11
Ireland	1973	2019	24
Italy*	1948	2019	33
Kazakhstan	1993	2017	19
Kosovo	2003	2018	12
Kyrgyzstan	1993	2018	23

Latvia	1993	2019	23
Liechtenstein	–	–	–
Lithuania	1993	2019	17
Luxembourg	1985	2019	14
Moldova	1993	2018	23
Monaco	–	–	–
Montenegro	2005	2019	15
Netherlands	1962	2019	17
North Macedonia	1994	2019	24
Norway	1963	2019	17
Poland	1986	2019	12
Portugal	1973	2019	28
Romania	1989	2019	18
Russia	1988	2018	16
San Marino	–	–	–
Serbia	2006	2019	7
Slovakia	1988	2019	12
Slovenia	1987	2019	13
Spain	1965	2019	14
Sweden	1963	2019	23
Switzerland	1982	2019	18
Tajikistan	1999	2015	6
Turkey	1968	2019	22
Turkmenistan	1993	1998	2
Ukraine	1988	2018	23
United Kingdom	1960	2019	32
Uzbekistan	1989	2003	6
Middle East and North Africa			
Algeria	1988	2012	3
Bahrain	–	–	–
Djibouti	1996	2017	5
Egypt	1965	2018	11
Iran	1986	2017	12
Iraq	1956	2013	4
Israel	1980	2018	22
Jordan	1987	2014	8
Kuwait	1973	2000	3
Lebanon	1960	2012	2
Libya	–	–	–
Malta	2005	2019	15

Morocco	1965	2014	8
Oman	2000	2011	2
Qatar	–	–	–
Saudi Arabia	–	–	–
Syria	1997	2007	3
Tunisia	1961	2016	9
United Arab Emirates	2015	2015	1
West Bank and Gaza	1996	2017	11
Yemen	1992	2014	4
Sub-Saharan Africa			
Angola	2001	2019	3
Benin	1959	2015	4
Botswana	1986	2016	5
Burkina Faso	1995	2014	5
Burundi	1992	2014	4
Cameroon	1996	2014	4
Cape Verde	2002	2015	3
Central African Republic	1993	2008	3
Chad	1958	2011	3
Comoros	2004	2014	2
Congo, Democratic Republic of the	2005	2013	2
Congo, Republic of the	2005	2011	2
Côte d'Ivoire	1959	2015	11
Equatorial Guinea	2006	2006	1
Eritrea	1997	1997	1
Eswatini	1995	2017	4
Ethiopia	1996	2016	5
Gabon	1975	2017	4
Gambia, The	1992	2016	7
Ghana	1988	2017	7
Guinea	1991	2012	5
Guinea-Bissau	1991	2010	4
Kenya	1977	2016	6
Lesotho	1987	2018	5
Liberia	2007	2016	3
Madagascar	1960	2012	9
Malawi	1969	2017	8
Mali	1989	2010	5
Mauritania	1987	2014	7
Mauritius	2007	2017	3

Mozambique	1997	2015	4
Namibia	1994	2016	4
Niger	1960	2014	7
Nigeria	1959	2019	8
Rwanda	1985	2017	6
Sao Tome and Principe	2001	2017	3
Senegal	1961	2012	6
Seychelles	2000	2013	3
Sierra Leone	1969	2018	5
Somalia	2002	2016	2
South Africa	1993	2017	10
South Sudan	2009	2009	1
Sudan	1969	2014	3
Tanzania	1964	2018	8
Togo	2006	2015	3
Uganda	1989	2017	9
Zambia	1959	2015	11
Zimbabwe	1995	2017	3
South Asia			
Afghanistan	2008	2017	3
Bangladesh	1963	2016	11
Bhutan	2003	2017	4
India	1951	2012	33
Maldives	2003	2016	3
Nepal	1977	2011	4
Pakistan	1963	2016	22
Sri Lanka	1953	2016	13
East Asia and the Pacific			
Australia	1969	2014	14
Brunei	2005	2016	3
Cambodia	1994	2012	10
China	1953	2019	19
Fiji	1968	2014	6
Hong Kong	1963	2016	13
Indonesia	1976	2018	30
Japan	1956	2014	26
Kiribati	2006	2006	1
Korea, North	–	–	–
Korea, South	1992	2016	11
Laos	1993	2013	5

Macao	–	–	–
Malaysia	1960	2016	17
Marshall Islands	–	–	–
Micronesia, Federated States of	2005	2013	2
Mongolia	1995	2018	10
Myanmar	1958	2017	3
Nauru	2013	2013	1
New Zealand	1973	2018	28
Palau	2014	2014	1
Papua New Guinea	1996	2010	2
Philippines	1957	2015	15
Samoa	2002	2014	3
Singapore	2003	2012	6
Solomon Islands	2005	2013	2
Taiwan	1953	2016	24
Thailand	1962	2018	27
Timor-Leste	2001	2014	3
Tonga	2001	2015	3
Tuvalu	2010	2010	1
Vanuatu	2010	2010	1
Vietnam	1993	2018	12
Total	**1950**	**2019**	**2,147**

Note: in countries with no income distribution information, percentile mean incomes are imputed annually using the average in their region and income group, except for North Korea, in which case the country mean income is imputed. Distributive information for non-survey years is either interpolated between the two closest or extrapolated before the first one or after the last one. (*) Italy 1948 is used for estimating Italy 1950, but not included in the dataset.

Source: author’s construction based on the WIID Companion.

Appendix B: inequality summary measures

Let $y = (y_1, y_2, \ldots, y_n)$ represent the global income distribution with n within-country income percentiles (100 equal-sized groups per country), each one representing a fraction p_i of the world's population, and $\bar{y} = \sum_{i=1}^{n} y_i p_i$ is the global mean income.

All inequality summary measures described below are continuous and meet the usual axioms of anonymity (inequality is invariant to permutations of individuals or country percentile groups), population principle (invariance to replicating the population several times), and the Pigou–Dalton principle of transfers (a mean-preserving transfer from a person—or percentile group—to another one with lower mean income reduces inequality if the ranks among those involved are also preserved).

I estimate the following inequality measures, which are either relative (scale invariant—inequality remains constant if all incomes are multiplied by the same common factor) or absolute (translation invariant—inequality remains constant after adding a common factor to all incomes).

Relative measures (scale invariant)

Gini index

$$Gini(y) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{y_i - y_j}{\bar{y}} \right) p_i p_j$$

General entropy

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

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, and SD is the standard deviation:

$$CV(y) = SD(y) / \bar{y}$$

Atkinson

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

Absolute measures (translation invariant)

Absolute Gini index

$$AbsGini(y)=\bar{y}Gini(y)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i-y_j\right)p_i p_j$$

Standard deviation

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

Appendix C: between- and within-country components of inequality

The global income distribution dataset provides measures of the overall distribution, as well as key information to understand the contribution to inequality of inequalities that take place within and between countries.[14]

Let $y=\left(y^1,\ldots,y^K\right)$ denote the global income distribution made up of K countries, where $y^k=\left(y_1^k,\ldots y_{n^k}^k\right)$ indicates the country distribution of country k with population n^k, total population is then $n=\sum_{k=1}^K n^k$. Furthermore, $I\left(y\right)$ denotes any global inequality measure computed on incomes y .

Now, let us consider the distribution in two counterfactual situations.

The first counterfactual distribution is given by $y_b=\left(y_b^1,\ldots,y_b^K\right)$, where in the distribution of each country k, $y_b^k=\left(\overline{y}^k,\ldots,\overline{y}^k\right)$, the income of every person has been replaced by the country's mean income $\overline{y}^k$. That is, this is the between-country global income distribution, in which all existing inequality within countries has been removed—that is, $I\left(y_b^k\right)=0$ for all countries.

A second counterfactual distribution is given by $y_w=\left(y_w^1,\ldots,y_w^K\right)$, where in the distribution of each country, $y_w^k=y_w^k\frac{\overline{y}}{\overline{y}^k}=\left(y_1^k\frac{\overline{y}}{\overline{y}^k},\ldots,y_{n^k}^k\frac{\overline{y}}{\overline{y}^k}\right)$, the income of every person (or percentile) has been rescaled by the same factor $\frac{\overline{y}}{\overline{y}^k}$, so that the new country mean is the global mean income $\overline{y}$ in all countries. This is the within-country global income distribution, in which all existing inequality between countries has been removed (all countries have the same mean and therefore global inequality across countries using those means is zero).

In the case of absolute measures of inequality, to estimate the within-country distribution, instead of rescaling incomes, they are added to the differential between global and country means to keep within-country inequality unchanged (the measures are translation invariant): $y_w^k=y_w^k+\left(\overline{y}-\overline{y}^k\right)$.

In the case of the Atkinson family, as usual, to construct the two counterfactuals I employ the concept of the 'equally distributed equivalent income' instead of the mean income—that is, inequality-adjusted welfare: $e^k=\overline{y}^k\left(1-A_\varepsilon\left(y^k\right)\right)$.

Measures of inequality $I\left(y_b\right)$ have been widely used as a measure of between-country inequality, while the level of inequality that is gone after equalizing within-country incomes can be interpreted also as a measure of within-country inequality: $I\left(y\right)-I\left(y_b\right)$.

[14] For a discussion of the underlying theory of inequality decompositions, see, for example, the discussion and related literature in Chakravarty (2009).

Similarly, $I(y_w)$ can be understood as the 'true' measure of inequality within countries, while the inequality that is gone after equalizing average incomes across countries can be seen as an alternative measure of the level of between-country inequality: $I(y)-I(y_w)$.

The global dataset reports both $I(y_b)$ and $I(y_w)$, along with overall inequality $I(y)$. The corresponding alternative measures, $I(y)-I(y_b)$ or $I(y)-I(y_w)$, can be easily recovered.

It is a known fact that the only inequality index in which inequality is the sum of the true between- and within-country inequality as defined above is the mean log deviation (GE_0): $GE_0(y)=GE_0(y_b)+GE_0(y_w)$. That is, this index is additively decomposable, and the magnitude of each term is the same in both alternatives (path independence). Other indices have other well-known decomposability properties, but only this one guarantees that both terms are pure, in the sense that the within-country term is not contaminated with between-country inequalities and vice versa.

In the case of other members of the GE_α family, which verify additive decomposability, what is usually interpreted as the 'within' component is $GE_\alpha(y)-GE_\alpha(y_b)=\sum_{k=1}^{K}\frac{n^k}{n}\left(\frac{\overline{y}^k}{\overline{y}}\right)^\alpha GE_\alpha(y^k)$, which is a weighted sum of country inequality, with weights being a function of country means (except when $\alpha=0$, i.e. mean log deviation). These terms, therefore, are not true within-country in the sense that they reflect prevailing inequality across country means too.[15] In the case of the Gini index, the decomposability is more complex, since it also depends on the level of overlap among country income distributions along the income space.

In the case of the Atkinson family, an index that is multiplicatively decomposable, and using the equally distributed equivalent income instead of the mean as the representative income of each country or globally, we get that:

$$1-A_\varepsilon(y)=(1-A_\varepsilon(y_b))\left(1-\sum_{k=1}^{K}\frac{n^k}{n}\frac{\overline{y}^k}{\overline{y}}A_\varepsilon(y^k)\right)$$

Note also that for all members of the *GE* family, the true within-country term (after the mean income has been equalized across countries) is just the population-weighted sum of country inequality: $GE_\alpha(y_w)=\sum_{k=1}^{K}\frac{n^k}{n}GE_\alpha(y^k)$.

To cope with this heterogeneity in decomposability properties, the WIID Companion also reports an additional estimate based on the Shapley decomposition, in line with Davies and Shorrocks (2021). The Shapley decomposition (Chantreuil and Trannoy 2013; Shorrocks 2013) is a simple method that allows us to obtain a consistent decomposition for all indices, with both terms adding up to overall inequality, regardless of their decomposability properties. It means, in this context, to just compute the average between the two possible estimates for each component:

---

[15] It also raises some normative issues, since inequality in rich countries has a higher contribution to overall within-country inequality than inequality in poor countries.

$$I(y)=S_b(y)+S_w(y)$$

with $S_b(y)=\frac{1}{2}\left(I(y_b)+I(y)-I(y_w)\right)$ and $S_w(y)=\frac{1}{2}\left(I(y_w)+I(y)-I(y_b)\right)$.[16]

Only in the case of the mean log deviation (GE_0) it happens that $S_b(y)=I(y_b)$ and $S_w(y)=I(y_w)$.

The importance of each component is then estimated as the percentage of total inequality:

$$s_b=100S_b(y)/I(y);\ s_w=100S_w(y)/I(y);\ s_b+s_w=100.$$

Note that $s_b>s_w$ if and only if $I(y_b)>I(y_w)$.

Summing up, the WIID Companion global dataset reports annual series for the indices described in Appendix B as follows:

1. $I(y)$, overall global inequality;
2. $I(y_b)$ and $I(y_w)$, 'true' raw inequality between and within countries (only add up to $I(y)$ for GE_0);
3. $\sum_{k=1}^{K}\frac{n^k}{n}\frac{\bar{y}^k}{\bar{y}}I(y^k)$, population-weighted sum of country inequality (the same as $I(y_w)$ for all GE_α; only equal to $I(y)-I_b(y)$ for GE_0); and
4. s_b, Shapley between-country inequality percentage (= $\frac{I(y_b)}{I(y)}100$ only for GE_0).

From the above, it is straightforward to recover other key terms, such as $I(y)-I(y_b)$, $I(y)-I(y_w)$, $s_w=100-s_b$, $S_b(y)=s_bI(y)/100$, and $S_w(y)=s_wI(y)/100$.

[16] In the case of the income share of the q% of the population, $q(y)$, we need to account for the fact that inequality is given by $I(y)=|q(y)-q|$. The expression in the Shapley decomposition then becomes: $S_b(y)=\frac{1}{2}(|q(y_b)-q|+|q(y)-q|-|q(y_w)-q|)$.