WID.world WORKING PAPER N° 2020/03

# Factor Shares in the Long Run

Matthew Fisher-Post

March 2020

World Inequality Lab

# Factor Shares in the Long Run

Matthew Fisher-Post[*][†]

*March 2020*

**Abstract**

The purpose of this study is to construct and then analyze a new dataset that systematically documents the labor share of national income for more than 200 countries over the past 70 years. Using new archival data on national accounts, we measure the long-run evolution of national income between factor shares (labor and capital). In addition to its implicit importance in the study of inequality, the data naturally lends itself to novel empirical analysis of international patterns in tax progressivity, trade integration, technology, and labor force composition.

[*]Center for International Development (Harvard Kennedy School)
*79 John F Kennedy St, Cambridge, MA 02138*
matthew_fisher-post@hks.harvard.edu
& World Inequality Lab (Paris School of Economics)
*48 Boulevard Jourdan, Paris 75014 France*
matthew.fisher-post@psemail.eu

[†]The author gratefully acknowledges research support and instructive guidance from Anders Jensen and Gabriel Zucman, and helpful comments and discussion from Facundo Alvaredo, Thomas Blanchet, François Bourguignon, Ignacio Flores, Jérémie Gignoux, Douglas Gollin, Loukas Karabarbounis, Brent Neiman, and Rafael Proença. I would be remiss not to acknowledge the kind support and historical data provided by Hermanus Smith and the National Accounts Section of the United Nations Statistics Division, as well as archival researchers at Harvard University's Lamont Library Government Documents collection. Errors are mine.

# Contents

# 1 Introduction

Recent studies have suggested an international rise in the capital share of national income along with increased concentration of both income and wealth (e.g., Piketty and Zucman, 2014; World Inequality Report, 2018). Such striking results raise further questions on the global extent of capital's rise, and on the fiscal policy tools that either mitigate or amplify inequality and growth. The first question, then, is how national incomes (and their growth) have been distributed between capital and labor, over the long run, and especially in the less-studied developing countries.

To answer this question, we build a novel and comprehensive database that harmonizes and systematically integrates 70 years of national accounts data from more than 200 countries across all levels of development.

Before studying these trends, the definitions will be important: Labor and capital shares can be defined as the proportion of national income that accrues to labor and capital inputs, the factors of production. We will return to their precise measurement below, but economists since the time of Solow (1956) have generally understood the labor share to be a static proportion of national income (between 60 and 80 percent of GDP, according to early estimates).

However, more recent empirical studies have called that assumption into question. While these studies vary in their methods, they generally find that the compensation of labor (relative to the returns to capital) has declined in a majority of countries observed. Although Karabarbounis and Neiman (2014) limited their measure to the corporate sector labor share (more on this in Section 2 below), they found that the relative compensation of labor (relative to the returns to capital) had declined in a majority of countries, and attributed these results to a declining relative price of capital investment. By contrast, Cette, Koehl and Philippon (2019) have used detailed French data to argue that any decline in the observed labor share may be an artifact of property values, and could be overstated if one does not adequately account for self-employment. Nevertheless, Piketty, Saez and Zucman (2018) documented that overall income for 90 percent of the US population (largely labor income) has not increased since 1980, despite skyrocketing growth at the top of the distribution (largely capital income). The first World Inequality Report (Alvaredo et al 2018) documented similar trends worldwide.

Simply put, scholars disagree on the basic facts of the labor share—its level and its trend—let alone the mechanisms driving those trends.

In this study, we clarify and reconcile several methodological differences among measures of the labor share, and extend the temporal and geographic coverage of these estimates—with a single benchmark measure over 200 countries and 70 years.

To construct factor shares, then, we begin with a pioneering dataset of national accounts variables, including archival materials not available elsewhere. Where mixed income is missing in some early country-years, we use information about its later values and other national accounts variables to retrieve and impute its earlier values. Our statistical model reflects persistence in each country's time series and captures constant country characteristics. After finally estimating mixed income from the rich coverage of long-run national accounts data that we do have, we are in position to estimate the complete long-run returns to labor and capital.

Our study of factor shares eventually aims, then:

i. to clarify and reconcile methodological distinctions in measuring the labor share;
ii. to expand coverage to include 200 countries over 70 years; and
iii. to analyze global trends in the labor share of national income.

The database should naturally lend itself to analysis of inequality and the determinants of labor's relative compensation, along dimensions of public finance and taxation, trade and labor regulations, and fiscal policy controls.

## 2 Conceptual Framework

After building our original comprehensive database of harmonized, long-run national accounts data, we can begin to calculate the labor share of GDP at factor cost:

$$\frac{Y_L}{Y} = \frac{CE + \alpha \cdot OS_{PUE}}{CE + OS_{CORP} + OS_{PUE}} \tag{1}$$

where $\frac{Y_L}{Y}$ is the labor share of GDP at factor cost (subtracting indirect taxes net of subsidies, which are neither labor nor capital income); $CE$ is labor income, the compensation of employees in formal wages and salaries; $OS_{CORP}$ is capital income, the operating surplus of corporations in the corporate sector (net of employee compensation and other costs); $OS_{PUE}$ is "mixed income," the return to self-employment and informal/unincorporated enterprises; and $\alpha$ represents the (usually unobserved) labor component of mixed income, usually represented as 0.7 (with 0.3 the capital share of mixed income).

Our preferred specification will in fact be in terms of *net* domestic product (i.e., net of depreciation) rather than gross domestic product, in which consumption of fixed capital is subtracted from the operating surplus of both corporations and unincorporated enterprises, in both the numerator and the denominator of equation (4). We will return to this calculation below in Section 4 and equation (15).

In a second adjustment, we prefer to discuss these factor shares in terms of net national income, rather than net domestic product (cf. Blanchet and Chancel 2016). The difference between net domestic product and net national income is the inclusion of net foreign income in the latter (which can be decomposed into the net foreign income accruing to labor and that accruing to capital). In other words:

$$NNI = GDP - CFC + NFI \tag{2}$$

and

$$NFI = NFI_L + NFI_K \tag{3}$$

where $NNI$ is net national income; $CFC$ is depreciation (consumption of fixed capital); and $NFI$ is net foreign income, attributable to labor ($NFI_L$) and capital ($NFI_K$).

Taking these adjustments into account, our benchmark specification represents the labor share of factor-price net national income:

$$\frac{Y_L}{Y} = \frac{CE + \alpha \cdot OS_{PUE} + NFI_L}{CE + OS_{CORP} + OS_{PUE} + NFI} \tag{4}$$

Before moving on, we should note an important assumption we are making in this estimation. That is, we currently do not have a better measure of $\alpha$ (the labor share of mixed income) than to follow the literature that estimates $\alpha = 70\%$.[1] It would be reasonable to speculate that $\alpha$ may differ across countries and over time, for both differences in measurement of what counts as self-employment and unincorporated enterprise (across countries) and for differences in the economic fundamentals within corporations vs. unincorporated enterprises (which could differ both across countries, with different mixes of industries and self-employment by industry, and over time). However, we do not currently have enough information to pass judgment on the likely ranges of $\alpha$ by place and time, so for now we leave it as it has been canonically represented in the literature. We will return to this point in future research.

Now, in considering the variety of ways to measure the labor share, two earlier efforts are particularly worthy of note. Gollin's seminal study (2002) underlined the pitfalls of estimating labor shares without taking self-employment into account, then set forth what has since become the most established method(s) for doing so.[2] While Gollin's data and method covered relatively few countries and years, Karabarbounis and Neiman (2014, henceforth KN) tracked the labor share for almost 60 middle- and high-income countries over more than 15 years.

We extend this coverage to developing countries, over the past 70 years: See Table 1 for a comparison of our data coverage to these earlier sources. We also compare studies from Feenstra et al (Penn World Table [PWT] 2015) and Guerriero (2019).

Because of the difficulty of measuring $\alpha$ in (4) above, and the potential controversy or model sensitivity from any assumption, KN (2014) simply omit mixed income $OS_{PUE}$ from their calculation and discussion of the labor share. As such, their labor share represents only the corporate sector of the economy, and does not capture self-employment and unincorporated enterprises. Gollin's (2002) benchmark specification does the same, in assuming that the labor share in the corporate sector is the same as in the economy as a whole.

[1] For further discussion of this point, refer to the World Inequality Lab's Distributional National Accounts guidelines (Alvaredo et al 2020).

[2] First, Gollin (2002) adjusted for mixed income by assuming that it follows the same proportion of labor income as in the corporate sector. Second, he proposed to improve on this unsatisfactory measure by imputing the same wage/salary to informal and self-employed workers as is found in the formally employed corporate sector. The first technique is perhaps oversimplistic, while data for the second technique is hard to come by, and even then it may not be fully satisfactory. Refer to Guerriero (2019), ILO (2019) and Cette, Koehl and Philippon (2019) for further discussion, as well as Table 2 below.

In contrast to that approach, we explicitly make use of national accounts information on mixed income—with expanded data coverage from archival records—and a simple, transparent statistical imputation procedure for the remaining missing data on mixed income. Table 2 presents a snapshot of existing methodological choices in the related literature, including Bengtsson and Waldenstrom (2015, henceforth BW), Cette, Koehl and Philippon (2019, henceforth CKP) and ILO (2019), in addition to the studies cited in the preceding Table 1.

Nonetheless, it could be argued that a limitation of the conceptual framework discussed here is the extent to which we must assume that $\alpha$ is a stable parameter and accurately represented at 70%. If the labor share in the corporate sector is significantly greater than (or less than) 70%, perhaps there would be less reason to believe that $\alpha = 70\%$ accurately represents the unincorporated enterprises' sector—and one might imagine that the bias could go either way. If corporate labor share is low, perhaps that is because the unincorporated sector labor share is high and self-employed workers are unwilling to enter a poorly remunerated sector. Or perhaps the reverse is the case, and the self-employed sector simply reflects the labor share of the corporate sector proportionally (or worse). These unknowns would strike at the robustness of the assumption in our benchmark specification of $\alpha$ above; however, we would like to show that the results are not sensitive to the range of that parameter, and that there is perhaps little reason to be skeptical about self-employment labor income in the first place. We will return to further discussion and alternative specifications below.

With this conceptual framework in mind, we can move to a presentation of our data sources and then a more detailed discussion of the method.

# 3 Data

In this section we describe the raw sources to create a unique, harmonized panel of national accounts data: from the World Inequality Database, UN System of National Accounts, Harvard Library archives, and complementary sources.

From the World Inequality Database (WID), we draw an extensive historical and geographic coverage of aggregate variables on gross domestic product (GDP) and net national income (NNI), including depreciation (consumption of fixed capital, CFC) and net foreign income (NFI),[3] as well as variables on population and purchasing power parity (PPP) and local currency unit (LCU, current and constant) price indices. We transform all WID variables into percentages of GDP, for comparison purposes.

From the United Nations System of National Accounts (UN SNA), we retrieve online national accounts statistics that cover approximately 4,000 country-years (across approximately 180 countries) for aggregate variables from the production and income accounts of their "Main

[3]For data on the labor share of net foreign income, we referred to IMF Balance of Payments data. See Blanchet-Chancel (2016). We take from current accounts "primary income, compensation of employees abroad, net (*credit - debit*)" as the labor share of net foreign income (and subtract this $NFI_{total}$ to reach the capital share), under the working hypothesis that self-employment income accounts for a negligible proportion of income from abroad, particularly in developing countries.

Aggregates and Detailed Tables" (UN 2019, henceforth MADT). These variables include gross value added; compensation of employees; gross operating surplus of corporations and quasi-corporations; that of private unincorporated enterprises; consumption of fixed capital; and indirect taxes net of subsidies—as well as some further-disaggregated components of these, all of which may be occasionally expressed net of CFC instead of gross.

Beyond what is available online, the System of National Accounts office (at the UN Statistics Division in New York, henceforth UNSD) has provided us with electronic archival material on the components of GDP, with more than 2000 country-year observations from the 1960s and 1970s. These observations are framed in the SNA1968 format (UN SNA 1968). To this we have added manually digitized data records from the Harvard Library archive, where national accounts coverage dates even to the 1950s and the format of the UN SNA data predates even the SNA1968 framework.

Further extensions are also possible, including to retrieve disaggregated labor and industry variables from additional UN sources and survey datasets. However, for the time being we only complement the WID and UN SNA data with some additional information on indirect taxes (when information on these are missing from those sources).

For missing data on indirect taxes, we draw from a parallel construction in the same project, a harmonized set of tax revenue variables from sources including OECD, IMF, ICTD, and national statistical offices' historical records both online and in the Harvard University government documents archive (manually digitized, as above).

From these several complementary but nonetheless distinct reference materials, some harmonization across statistical yearbooks and sources is therefore required.

# 4 Method

In this section we will discuss the steps we follow to achive a factor shares estimate for every country-year for which we ever have any data on the cost components of GDP in the national accounts. In order to get there, we need to: (a) harmonize data across sources; then (b) retrieve occasional missing values that are in fact implicitly available; and finally (c) compensate for remaining missing values, which can be imputed from other values in the same countries' long-run series.

To harmonize across sources we identify and smooth over any artificial statistical outliers; pay attention to database footnotes; draw from external data; and prioritize conflicting data.

To recover missing values that are in fact implicitly available, we leverage national accounting identities that help us extend coverage beyond the data that is available on the surface level.

Last, to compensate for any remaining missing values in our critical variables of interest, we apply a simple and transparent imputation procedure.

First, to harmonize series and units across sources, we need a common frame of reference. We first put all variables in terms of GDP. To abstract away from raw levels particularly helps in cases where currency units or price indices varied across sources. This is particularly the case for comparing WID (constant LCU or PPP) with UN SNA (current LCU) variables, as well as across country-years of SNA data where the currency changes.

In order to select among the several available sources of data discussed above—and in some cases, several available series within a single country-year—we prioritize as follows: First, we prefer online data. Where this is unavailable, we turn to archival data provided by the UNSD. Where this is unavailable, we refer to the manually digitized material from the Harvard Library archives.[4] In some cases, there is more than one 'series' available within a given country-year, even from the same 'source.'

That is, the UN SNA has revised its framework national accounting guidelines several times since the initial accounts. National statistical agencies have followed suit with parallel time series that map one era with one methodological 'series', and then map another with a usually slightly revised methodological 'series.' To the extent that these series can differ during the year(s) of their overlap, it is important to 're-base' them across the transition year(s), in order to avoid sharp jumps in the time series that are only artifacts of national accounting procedure and not of the actual economies (and data-generating processes) they represent.

We use a reverse smoothing procedure to dampen (if not entirely remove) the effect of any artificial shocks during these 'series change' years. Rather than using the well-known Hodrick-Prescott filter (to remove stationary business cycle shocks), nor a constant arithmetic term imputed backwards, we refer to Hamilton (2018) to implement a simple and intuitive technique. That is, following Hamilton's benchmark specification, but in reverse to suit this context,[5] we regress earlier values (from the earlier series) on later ones (from the later series):

$$y_{t-h} = \beta_0 + \beta_1 y_{t+1} + \beta_2 y_{t+2} + \cdots + \beta_h y_{t+h} + v_{t-h} \tag{5}$$

where $t$ is a five-year window of years prior to the series change; $h = 5$, per Hamilton (2018); and $y$ represents each of the national accounting variables in our analysis.

In all cases of multiple series within a country-year-source, we prefer the later series to the earlier series, for the likely measurement accuracy improvements that revised series reflect.

However, when in some cases the earlier series include variables that the later series do not include, we incorporate this information in a way that is faithful to the aggregate levels of the later series, by way of equation (5). This is especially true for the long-run measurements of mixed income (operating surplus of private unincorporated enterprises, $OS_{PUE}$), whose importance we will return to below.

Online and archival data from UN SNA includes a raft of footnotes—many of them reflecting necessary adjustments to the raw data. For instance: In some cases we needed to remove FISIM

[4] Where no data is available, we impute the (earlier) missing values based on the (later) observed values, within each country. More on this in Section 4 below.

[5] We use lead variables instead of lag variables.

(financial intermediary services indirectly measured) from gross value added, as they are a type of intermediate consumption. In other cases, statistics for many country-years were recorded as gross of depreciation while the footnote explained they were in fact net. Or the value for gross domestic product was sometimes actually the gross value added, per footnotes. Sometimes the value of taxes was included in operating surplus or other inappropriate locations. Other footnotes simply noted an irreconcilable discrepancy in the national accounting, for which perhaps little could be done, but it is nonetheless important to flag these errors (more on this in the next section). Indeed the SNA footnotes alone represent a minefield for accurate measurement from national accounts data, but after careful adjustment we can be more confident in the resulting series.

In general, in the case of conflicting accounts within a country-year-series on GDP or its components, we prefer data from the 'production account' (MADT Table 2 series) for gross value added and taxes on production, but data from the 'income account' (MADT Table 4 series) for statistics on compensation of employees, operating surplus, and mixed income. The same disclaimers apply within both online and archival national accounts datasets.[6]

When government operating surplus is listed as a non-zero value, we split it proportionally between (quasi-)corporate and unincorporated surpluses ($OS_{CORP}$ and $OS_{PUE}$, respectively), in order to match the convention of its absence (or *zero* value) in the accounts of most countries. In practice this distinction does not make much difference, as the value of $OS_{GOV}$ is generally negligible even when it does exist.

Where data on NIT (net indirect taxes, i.e., indirect taxes net of subsidies) is missing, we make several adjustments. First, we refer to additional tables in the national accounts online, parallel to the value added and GDP tables but which do not include their components. Then, if data on subsidies is still missing, we reconstruct the aggregate amount from disaggregates (i.e., from individual types of subsidies listed in the national accounts). If disaggregates are also missing, we assume subsidies to be zero. Where data on total indirect taxes is missing, we similarly reconstruct from disaggregated statistics on types of indirect taxes in the national accounts. Where these also are missing, we import indirect tax revenues as a proportion of GDP from complementary outside sources, including the OECD Revenue Statistics (2019), IMF Government Finance Statistics (2019), and the ICTD Government Revenue Dataset (Prichard et al 2014)—as well as our own internal compilations from archival materials (referenced above).

Where data on CFC (by which to calculate net national income and net $OS_{CORP}$ and $OS_{PUE}$; more on this below) is missing from UN SNA records, we retrieve these values from the World Inequality Database (cf. Blanchet-Chancel 2016). For its disaggregates, when these are unavail-

[6]This has most of all to do with the purpose of the accounts in question: While the MADT Table 2 series does in principle have some information on compensation of employees and operating surplus in the formal sector, these are often deductively arrived at from the household surveys that inform the MADT Table 4 series—especially in developing countries. That is, to put it in shorthand, where the formal economy represents a less significant part of overall production, household survey data on income generation in the informal economy becomes correspondingly more important, and the total economy production estimates become more of a derivative estimate working backward from the GDP estimate in the 'income account.' See Lequiller Blades (2014) for discussion and specific country examples in MADT 2018 for examples. In principle MADT series from Tables 2 and 4 (SNA 'production' and 'income' accounts) should always agree with one another, but in practice they do not always do so; in those cases we prefer the estimates from the Table 4 series. Meanwhile the Table 4 series do not contain any information on value added (this is strictly recorded within the 'production account' in the Table 2 series.

able, we assign CFC from corporations vs. CFC from unincorporated enterprises to be equivalent to the ratio of operating surplus in these two sectors. It should be noted that this a preliminary assumption, as one may also suspect that the use (and depreciation) of fixed capital could differ across corporate and unincorporated enterprises. In practice, the sensitivity or propensity of CFC to differ by formal vs. informal industry will not affect our results in the calculation of factor shares.

While UN SNA benchmark values are most easily (and up to here) treated in terms of GDP, our benchmark measure is to prefer all values in terms of national income. This simple transformation is to subract depreciation from GDP and to add back net foreign income (also retrieved from WID).

Now that we have harmonized data across sources, we can turn to the problem of missing values within the multi-source harmonized dataset. This is the second step. For these remaining missing values, we can often make use of accounting identities to (implicitly) retrieve SNA values, or we can impute these values based on neighboring values within a country time series.

So, in our second step to achieve broad database coverage, here we discuss the use of accounting identities. In order to retrieve missing values that are in fact implicitly available, we make use of several standard accounting identities in the national accounts.

For example, if the national accounts raw data contains information on GDP, and on compensation of employees, and on taxes, but does not contain information on overall operating surplus (of both corporate and unincorporated enterprises), then we can retrieve the latter ($MIOS$) from the following national accounting identity:

$$GDP = CE + MIOS + NIT \tag{6}$$

where $MIOS$ is the sum of 'mixed income' ($OS_{PUE}$, operating surplus of the unincorporated sector) and operating surplus of the corporate sector ($OS_{CORP}$):

$$MIOS = OS_{PUE} + OS_{CORP} \tag{7}$$

In the same way, we can make many similar calculations to retrieve implicit values. To start, $NIT$ can be decomposed into net indirect taxes on products (i.e., net of subsidies on products), plus other net indirect taxes on production (i.e., net of other subsidies on production):

$$NIT = NIT_{products} + NIT_{other} \tag{8}$$

which helps us move between $GDP$ and gross value added ($VA$):

$$GDP = VA + NIT_{products} \tag{9}$$

such that

$$VA = CE + MIOS + NIT_{other} \tag{10}$$

So far, all terms are *gross* of depreciation, but their *net* values can be used (or retrieved) as follows:

$$MIOS_{gross} = MIOS_{net} + CFC \tag{11}$$

$$CFC = CFC_{PUE} + CFC_{CORP} \tag{12}$$

$$OS^{gross}_{PUE} = OS^{net}_{PUE} + CFC_{PUE} \tag{13}$$

$$OS^{gross}_{CORP} = OS^{net}_{CORP} + CFC_{CORP} \tag{14}$$

$CFC_{PUE}$ is the fraction of overall $CFC$ attributable to private unincorporated enterprises—and also can be retrieved as follows:

$$\frac{CFC_{PUE}}{CFC} = \frac{(1-\alpha) \cdot OS^{gross}_{PUE}}{MIOS_{gross} - (\alpha \cdot OS^{gross}_{PUE})} \tag{15}$$

As discussed above, we are tentatively assuming that consumption of fixed capital in (self-employed) unincorporated enterprises is similar to consumption of fixed capital in the corporate sector. We therefore assign unincorporated enterprises' share in total capital depreciation according to unincorporated enterprises' share in overall capital income. Capital income within operating surplus of unincorporated enterprises is represented as $(1-\alpha) \cdot OS_{PUE}$ while overall capital income is represented as $MIOS - (\alpha \cdot OS_{PUE})$.

Gross capital income could equally be denoted as $OS_{CORP} + (1-\alpha) \cdot OS_{PUE}$ by way of equation (7).

With these accounting identities we can therefore move from approximately 4,000 (raw, online) to approximately 5,500 (implicitly filled, and including archival data) complete country-year observations in which no more than one value of our important accounting identities is missing.

In addition to our use of national accounting identities to retrieve (implicitly non-)missing values, we also use accounting identities to replace some obviously erroneous values. Both before and after implicitly filling in the observations that had only one missing value (via accounting identities, above), we winsorize the distribution of each variable at $p < 0.01$ in each tail, and replace with the value retrieved via accounting identity. True values return the same as before; extreme and implausible values return with a corrected value to fit within the identity. For any remaining identities in which the sum of the parts is not 100 percent of the whole (now always within five percent), for precision we re-scale the parts to 100 percent of the whole,[7] proceeding in a hierarchical manner (e.g., re-scaling (6) before (7) and (8)).

However, even after filling in the gaps in national accounts with implicitly available information, we still often lack some information on mixed income. This brings us to our third step.

That is, in the final step, for missing values that are still missing after our work with accounting identities above, we follow the imputation procedure from Blanchet and Chancel (2016).

[7] Since we are working in percentages of GDP and not in levels, by construction it becomes immaterial whether we re-scale GDP to match its components, or re-scale the components to match GDP; and so forth at every subsequent level of national accounting components' decomposition.

They put forward a straightforward, simple and transparent method for imputing consumption of fixed capital (depreciation) in the World Inequality Database (WID) series on GDP (without decomposing GDP into factor shares). In the same way as they observed three stylized facts about $CFC$ to impute its missing values, so we also used a similar set of stylized facts about mixed income $OS_{PUE}$ in order to statistically model its missing values. Our stylized facts for $OS_{PUE}$ are as follows:

1. Mixed income ($OS_{PUE}$) tends to represent a lower share of $GDP$ (and of $MIOS$) in more developed countries (with a higher $GDP$ per capita at purchasing power parity). This is an artifact of the outsized role of the informal economy and household enterprise in developing countries, where formalization and modernization and industrialization and/or corporatization are highly correlated with each other and with income.

2. Some countries have structurally high (or low) levels of mixed income (as a share of $GDP$, or as a share of $MIOS$)—whether due to long-term economic trends (e.g., the informal economy, as above, or the role of certain industries that that are themselves characterized by informality or household enterprise) or even due to a certain long-term tendency or practice in the national statistical office (e.g., in some countries many household enterprises are closely followed by tax authorities and therefore mapped by statisticians into the corporate sector; or vice-versa for countries where some quasi-corporations are poorly accounted for).

3. Mixed income as a share of $GDP$ (and of $MIOS$) is persistent, such that the value in year $t$ will be closely correlated to the value in year $t+1$. Exogenous shocks that affect $OS_{PUE}$ without similarly affecting $GDP$ per capita are rare, so its decrease is a slow process as $GDP$ per capita increases.

Therefore, we can model $OS_{PUE}$ as a function of both log $GDP$ per capita (at PPP) and $MIOS$ (as a percentage of GDP), with a random effect to capture constant country characteristics:

$$OS_{PUE_{it}} = \beta_0 + \beta_1 GDPpc_{it} + \beta_2 MIOS_{it} + u_i + \varepsilon_{it} \tag{16}$$

where there is a random effect term $u$ for each country $i$, and $\varepsilon$ is the error term for each country-year $it$.

To account for persistence in $OS_{PUE}$ we model the error term $\varepsilon_{it}$ in (16) as an AR(1) process:

$$\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \eta_{it} \tag{17}$$

where $\eta_{it}$ is i.i.d. white noise.

As in Blanchet-Chancel (2016), when we know part of the $OS_{PUE}$ series for a given country (observing it in later years), we can estimate the country's random effect $u_i$ and use that in the imputation. When no later value of $OS_{PUE}$ is observed, we assume $u_i = 0$. $OS_{PUE}$ returns to its expected long-run value at a rate of $\rho^t$.

We run the same imputation procedure for $OS_{CORP}$ as in equations (16) and (17) for $OS_{PUE}$, and then scale imputed $OS_{PUE}$ and imputed $OS_{CORP}$ to match 100 percent of $MIOS$ via national accounting identity (7). In this way, we can impute a value for mixed income in the country-years where it is missing.

We generate an estimate of factor shares, as in (4), for every country-year in which we have the complete $GDP$ components $CE$, $MIOS$ and $NIT$, as in (6). In fact, we insert a first-stage estimate for the missing observations of $CE$, $MIOS$ and $NIT$, using a similar procedure as in equations (16) and (17):

$$Z_{it} = \beta_0 + \beta_1 GDPpc_{it} + u_i + \varepsilon_{it} \tag{18}$$

where the vector of first-stage $Z_{it}$ variables represent those cost components of $GDP$.

After re-scaling the imputed cost components to match 100 percent of GDP per (4), the imputed value of $MIOS$ can be inserted into (16) to expand our factor share coverage. It is worth noting that there is a slight loss of precision in this two-stage imputation procedure. However, we test the procedure on observed values and note that the AR(1) persistence model effectively predicts missing values to a first-order approximation for both $Z_{it}$ and $OS_{PUE}$. In all cases, we keep the 80 percent prediction interval to show an estimation range of plausible upper and lower bounds for these imputed values.

In sum, we achieve a factor share estimate for all countries in which we ever have any data on the cost components of $GDP$, and for all country-years in which we have an estimate of $GDP$ from WID.

The benchmark factor shares series, then, is completed according to the definition in equation (4) discussed in Section 2 above. From here we can analyze the series for a long-run decomposition of global income growth and its distribution.

For now, we leave that task for further research.

## References


Alvaredo F, et al. 2020. National Accounts (DINA) guidelines: Concepts and methods used in WID.world. World Inequality Lab Working Paper. (Forthcoming.)

Alvaredo F, et al. 2018. World Inequality Report. Harvard University Press.

Bengtsson E and D Waldenstrom. 2018. Capital shares and income inequality: Evidence from the long run. Journal of Economic History 78(3): 712-743.

Blanchet T and L Chancel. 2016. National Accounts Series Methodology. World Inequality Lab Working Paper.

Cette G, L Koehl and T Philippon. 2019. Labor shares in some advanced economies. NBER Working Paper No. 26136.

Feenstra RC, R Inklaar and MP Timmer. The next generation of the Penn World Table. American Economic Review 105(10): 3150-82.

Guerriero M. 2019. The labor share of income around the world: Evidence from a panel dataset. Asian Development Bank Institute (ADBI) Working Paper No. 920.

Guerriero M and K Sen. 2012. What determines the share of labour in national income? A cross-country analysis. IZA Discussion Paper No. 6643.

Hamilton JD. 2018. Why you should never use the Hodrick-Prescott filter. Review of Economics and Statistics 100(5): 831-843.

International Labour Organization (ILO). 2019. The Global Labour Income Share and Distribution. Methodological Description.

International Monetary Fund (IMF). 2019. Government Finance Statistics.

Karabarbounis L and B Neiman. 2014. The global decline of the labor share. Quarterly Journal of Economics 129(1): 61-103.

Organization for Economic Cooperation and Development (OECD). 2019. Government Revenue Statistics.

Piketty T, E Saez and G Zucman. 2018. Distributional national accounts: methods and estimates for the United States. Quarterly Journal of Economics 133(2): 553-609.

Piketty T and G Zucman. 2014. Capital is back: Wealth-income ratios in rich countries 1700â2010. Quarterly Journal of Economics 129(3): 1255-1310.

Prichard W, A Cobham and A Goodall. The ICTD Government Revenue Dataset.

United Nations Statistics Division. 2019. National Accounts Statistics: Analysis of Main Aggregates and Detailed Tables

Solow R. A contribution to the theory of economic growth. Quarterly Journal of Economics 70(1): 65-94.

Zucman G. 2014. Taxing across borders: Tracking personal wealth and corporate profits. Journal of Economic Perspectives 28(4): 121-48.

# Tables

## List of Tables

Table 1: Countries covered *(N)*, by World Bank (2019) income classification

| | **WID(2020)** | **Gollin(2002)** | **KN(2014)** | **PWT(2015)** | **Guerriero(2019)** |
|---|---|---|---|---|---|
| High Income (81) | **68** | 17 | 35 | 57 | 58 |
| Upper-Middle Income (56) | **55** | 5 | 15 | 36 | 34 |
| Lower-Middle Income (47) | **46** | 7 | 7 | 28 | 16 |
| Low Income (34) | **33** | 1 | 1 | 15 | 5 |
| **Total (218)** | **202** | **30** | **58** | **136** | **113** |

Table 2: Methodological comparison across related literature

| **Authors** | ***N*** | **years** | **Benchmark approach to $OS_{PUE}$** |
|---|---|---|---|
| **WID** (2020) | *202* | 1950- | See equation (4) above: While the labor share in the corporate sector is simply $\frac{CE}{CE+OS_{CORP}}$, the labor share of unincorporated enterprises (mixed income, $OS_{PUE}$) is assumed to be $\alpha = 0.7$. $OS_{PUE}$ is imputed where missing based on within-country persistence and time trends of observed values. We use net national income in lieu of GDP: Net indirect taxes and depreciation are excluded, while net foreign income (and its labor share) are included. |
| **Gollin** (2002) | *30* | 1970- | Adjusted series assumes labor share of mixed income is same as in the corporate sector. In other words, $\alpha = \frac{CE}{CE+OS_{CORP}}$. Preferred specification would multiply $CE$ by total number of workers, but data deemed lacking. |
| **KN** (2014) | *58* | 1975- | Benchmark series only tracks corporate sector labor share: $\frac{CE}{CE+OS_{CORP}}$. |
| **PWT** (2015) | *136* | 1950- | Definition differs across countries and over time. For some countries, adds agricultural value added to $CE$ (double-counting agricultural *employee* income). Often extrapolates a constant share for >25 years. |
| **Guerriero** (2019) | *113* | 1960- | Preferred specification follows Gollin's (2002, above), but subtracts self-employed (non-corporate sector) *employers* from the total number of workers. |
| **BW** (2015) | *21* | 1900- | Gollin's (2002) preferred "labor method" ($Y_L = \frac{CE}{employees} \cdot workers_{total}$) where available, otherwise our "proportional method" ($Y_L = CE + \alpha \cdot OS_{PUE}$). |
| **ILO** (2019) | *95* | 2004- | Uses ILO Harmonized Microdata to estimate total labor income, including mixed income—which is predicted as a self-employment pseudo-'wage' by imputing on matched observable characteristics of employees. Separate predictions after disaggregating self-employment into: own-account workers; contributing family workers; and employers. |
| **CKP** (2019) | *10* | 1949- | Imputes the average formal-sector (i.e., *employee*) wages ($CE$) to *all* workers—à la Gollin (2002, above) best practice—but indexed by industry. Calculated variously for total economy, corporate sector only, and corporate sector without real estate sector. |