



IARIW 2025

IARIW 2025

Thursday, October 2 & Friday, October 3

Measuring Inequality in a Formalizing Economy

David Garcés Urzainqui
(University of Copenhagen)
Jane Kanina
(Kenya Revenue Authority)
Josephine Mugure
(Kenya Revenue Authority)

Paper prepared for the IARIW–World Bank–UEB/VNU Conference on “Improving Well-being Measurement in Data-challenged Environments in Developing Countries for Better Evidence-based Policies” October 2-3, 2025

Session 3A Poverty and Inequality
Time Slot: Thursday, October 2, 3:30- 5:30 PM

Measuring Inequality in a Formalizing Economy

David Garcés Urzainqui^{*}, Jane Kanina[†] and Josephine Mugure[†]

September 2025

PRELIMINARY VERSION - PLEASE DO NOT CITE WITHOUT PERMISSION

Abstract

Inequality estimates that address the undercoverage of top income earners in household surveys are rare in Sub-Saharan Africa. We measure interpersonal inequality in employment income in Kenya between 2016 and 2022 using household survey and tax data. Since our tax data mainly covers individuals in formal employment, we investigate variations of data combination approaches centered around correcting the income distribution by type of employment. Both these approaches and more standard methods reveal very high levels of inequality. We further evaluate the impact of personal taxes on inequality using combined tax and survey data.

^{*}Corresponding author. Department of Economics, University of Copenhagen (d.garces.urza@gmail.com).

[†]Kenya Revenue Authority (KRA).

Acknowledgements: This paper was developed as part of Royal Danish Embassy in Nairobi, RDE-supported collaborative research project entitled “Economic Research and Policy Making in Kenya (ERPDK)”, supported by the African Economic Research Council (AERC).

1 Introduction

The incorporation of tax data to the study of income inequality over the last few decades has revolutionized our understanding of this phenomenon (Alvaredo et al., 2013; Atkinson and Piketty, 2007). It is by now well established that household surveys, the backbone of conventional inequality analysis, fail to capture the top of the income distribution. While some researchers have also incorporated National Accounts (Garbinti, Goupille-Lebret, and Piketty, 2018; Piketty, Saez, and Zucman, 2017; Zwiijnenburg, 2019) and other data sources, the combination of tax and survey data has emerged as the canonical approach for addressing the shortcomings of survey data, with researchers employing a wide variety of methods, reviewed in Lustig and Vigorito (2025) to leverage the strengths of each of these data sources.

The implementation of these approaches have mainly focused on higher-income countries. However, the underlying issues leading to the ‘missing rich’ problem (Lustig, 2019), such as difficulties to obtain responses from the rich or income misreporting, are unlikely to be less serious in developing countries. In addition, capacity and budget constraints often prevent these countries from incorporating sophisticated but costly sampling techniques to better capture the upper tail of the income distribution; or even from collecting high-quality income data at a sufficient frequency. Existing applications to e.g. Latin America (De Rosa, Flores, and Morgan, 2024), the Middle East (Assouad, 2023) or India (Chancel and Piketty, 2019) often reveal very large income disparities, sometimes at odds with estimates of levels and trends exclusively based on household surveys.

Unfortunately, such income inequality estimates combining household survey and administrative tax data are still relatively uncommon in the context of Sub-Saharan Africa (Chatterjee, Czajka, and Gethin, 2023; Czajka, 2017) due to data quality and availability issues concerning both household surveys and administrative tax data. The discourse on inequality in Sub-Saharan Africa is thus dominated by consumption inequality estimates, since the underlying data are more reliable, as stronger emphasis has traditionally been placed in region on the precise quantification of consumption expenditure to measure poverty, and

they are believed to better capture differences in living standards (Ravallion, 2016). However, income inequality estimates, which may widely differ from their consumption counterparts, convey valuable information on disparities in command over resources that have important economic, social, and political implications. Projections based on the few countries for which data combination has been possible suggest that income inequality estimates adequately capturing top incomes may lead to substantially different conclusions on the extent of inequality in the region (Chancel et al., 2023).

We aim to contribute to our understanding of disparities in Sub-Saharan Africa by studying the case of Kenya. For that purpose, we draw on detailed administrative tax data provided by the Kenya Revenue Authority (KRA) and household surveys conducted by the Kenya National Bureau of Statistics (KNBS) between 2016 and 2022. Our main source of administrative tax data is the universe of Pay-As-You-Earn (PAYE) returns on yearly employment income, which are filed by employers and, as third-party reported data (Kleven et al., 2011), generally perceived as accurate. Since most of the surveys in this period also limit themselves to provide information on employment income, our estimates should be essentially understood as measuring interpersonal employment earnings inequality between individuals that are between 18 and 65 years old. The neglect of sources of income relevant for either tail of the distribution (social and informal transfers at the bottom, capital incomes at the top) is thus a limitation of our exercise.

The nature of our tax data is thus different from most applications in higher income countries, where one of the main advantages of administrative tax data with respect to well-established household surveys concerns the measurement of capital income. Even though we complement PAYE data with personal income tax returns providing some information on non-employment income, we can expect our tax data to mainly enhance our coverage of formal worker incomes, since only a very small share of informal and own-account workers accurately report their income, and most capital income is not captured by these returns. When incomes in the informal economy largely overlap with those in the formal sector, as

often in Sub-Saharan Africa (Günther and Launov, 2012) and also in Kenya, the standard method simply replacing survey with tax data beyond a certain income level (Bartels and Metzing, 2018; Blanchet, Flores, and Morgan, 2022; Jenkins, 2016) has the drawback of missing informal employees in the higher-income segment measured with tax data, and may also (unless the survey is very accurate) misrepresent the distribution of formal employees in the lower-income range.

We thus propose a simple alternative data combination procedure to obtain improved inequality estimates for developing countries with a large informal economy where incomes largely overlap with those in the formal sector. We rely on administrative data to (by definition) accurately measure, fully or partially, the formal sector, while information about the informal economy is obtained from survey data. The link to combine survey and tax data is given by a survey variable capturing whether income tax is deducted from the worker’s earnings. Inequality indicators can then be readily computed from combined tax and survey data after proportionally correcting sampling weights to account for the underrepresentation of formal workers. As will be discussed below, whether this procedure is more appropriate than the standard approach dealing with the overall income distribution depends on the reasons that explain the discrepancies between the income distribution for formal workers in tax and survey data.

This approach requires a nationally representative household survey providing information on formality status and the income distribution, and administrative tax data on personal income covering high-income earners in formal employment better than the survey. Following the classification in Lustig and Vigorito (2025), it can be seen as combining replacement and reweighting elements. We believe that it might prove useful in the context of ongoing digitalization of tax collection and administration, and hence increasing availability of and improved access to tax data across Sub-Saharan Africa, where high quality administrative tax data may often disproportionately cover income derived from formal employment.

An additional challenge in our context is to estimate the distribution of income as mea-

sured by the household survey across all the paid employees and self-employed. Unfortunately, income data are reliable only for a non-representative subset of them, skewed towards those in more stable jobs. We use that set of reliable observations to predict income on the basis of consumption, allowing for heterogeneous saving rates, and several relevant individual characteristics, estimating separate models for employees and the self-employed. We then use those models to impute income for observations without reliable income information.

Our data combination exercise allows us to provide estimates of interpersonal inequality in labor earnings that address the undercoverage of top incomes in household surveys. We find high levels of inequality, with pre-tax Gini coefficients ranging between 0.60 and 0.65, and earners in the top 1% capturing slightly above 20% of aggregate income. In all cases, data combination yields higher inequality estimates than each of the data sources regarded in isolation. Surveys particularly struggle to precisely measure top income shares. The approach suggested in this paper tends to lead to somewhat higher inequality levels than the standard approach.

We also study the distributional impact of personal income taxes. Our data combination approach is well suited for that purpose because it preserves the formal-informal distinction when combining the data, as required to credibly ascertain the effects of personal income tax structure on inequality in a context characterized by high informality. In spite of the well-known difficulties of household surveys to capture the top of the income distribution, existing literature has typically evaluated the progressivity of tax systems or reforms in developing countries using household surveys (Lastunen et al., 2024; Inchauste and Lustig, 2017). We show how combining both data sources can substantially alter the conclusions on the impact of personal income taxes on inequality.

2 Data

2.1 Household Survey Data

This study draws on several rounds of nationally representative household survey data collected by the KNBS. We use the 2015/16 Kenya Integrated Budget Survey (KIHBS). The KIHBS is a large survey that, over 12 months, collects information on a wide array of demographic, welfare and economic indicators, such as education, health, employment, income, consumption expenditure, housing, agricultural production, etc. The survey informs various official statistics (CPI, National Accounts). It is designed to be representative for urban and rural areas at the county level. The 2015/16 round covers 92,846 individuals in 21,773 households.

The KIHBS is collected once every ten years. Therefore, the KNBS introduced the Kenyan Continuous Household Survey (KCHS) in order to obtain timely information on some socio-economic indicators. The KCHS uses a nationally representative sample of households, selected through a multi-stage stratified sampling process, and employs a rolling sample design, where data are collected continuously throughout the year. While this survey collects substantially less extensive and detailed information than the KIHBS, it includes an employment module that is used to compile official labor force statistics aligned with ILO standards. The questionnaire in this module is very similar to that in the KIHBS. We use three KCHS rounds corresponding to 2019, 2021 and 2022. Their sample sizes range between 17,042 and 20,691 households including 68,277 to 86,647 individuals.¹

Income data in the KIHBS are substantially more comprehensive, as the survey includes data on transfers from various sources (social networks, NGOs, public, remittances), as well as pension, rental, interest, and investment income. However, both surveys are broadly comparable in their coverage of income derived from employment, to which we will therefore

¹Sampling for the 2019 round was based on the 2009 Population Census and used a rotating panel design. We use the data provided as annual cross-section. Sampling for 2021 and 2022 was purely cross-sectional and based on the 2019 Census.

restrict our attention in the current version of the paper. Both surveys use the same reference period of 7 days for economic activity, and collect income data that allow deriving monthly earnings on the basis of a similar questionnaire. Income data are collected on the primary and secondary economic activity of the individual. The KCHS also collects information on income from any other jobs, but amounts are comparatively quite small.

Income data collection is similar, but not identical, across surveys. The KCHS uses different questions for individuals, depending on whether their primary economic activity is self-employment (own-account workers and working employers) or paid employment outside the household. The KIHBS, on the other hand, collects that information with a single question common to all economically active respondents. In addition, it includes a question in all allowances received. This information is also collected by the KCHS for employees through separate questions on medical, housing, transportation, and other allowances. Both surveys take a different approach for casual workers, who provide information on average daily rate and number of days worked in the last month. All surveys collect income data for paid employees and the self-employed, including those in agriculture, but coverage of contributing family workers and other smaller categories varies, as the KCHS 2021 and 2022 do not elicit information on their income. We therefore restrict our attention to paid employees and the self-employed in our main analysis, who make up 65% to 75% per cent of the economically active population.

Data preparation mainly involves dealing with two issues. The first of them is a certain lack of clarity regarding the frequency of recorded payments. We deal with this issue following a conservative approach, where recorded duration is respected except in flagrant cases. Second, there are some missing values. For individuals who provide a range of income instead of an exact value, we assign them an income randomly drawn from the empirical distribution within that range. We deal with the remaining missing and unreliable values by imputing incomes for employees and the self-employed, as explained in further detail in Section 3.1. For this purpose, we use data on consumption expenditure, household demographics, education,

and housing characteristics available from these surveys.

2.2 Administrative tax data

We combine household survey data with detailed administrative tax data provided by the KRA. We use the universe of PAYE returns and personal income tax returns filed by individuals (the so-called P9 form) between 2016 and 2022. PAYE is a wage withholding tax system in which companies deduct a portion of employees' pay and send that tax to the KRA. It accounts for the majority of personal income tax revenue (over 95% on average between FY 2017/18 to FY 2021/22). The two main sources of PAYE tax revenues are the public sector and large firms, respectively contributing 39% and 35% over that period. We observe employment income at a yearly frequency, which consists of basic pay and wage, taxable allowances in cash or in kind (housing, transport, etc.), and bonuses, commissions and overtime pay.

Using anonymized taxpayer numbers, we combine PAYE data with data from personal income tax returns, which include data on employment and non-employment income. Individuals are expected to file these returns annually to reconcile the tax withheld with their overall tax liability. The number of observations in the combined dataset grows from around 2.36M observations in 2016 to around 2.97M observations in 2022. Individuals in the PAYE system are often missing from P9 data (25% of cases), probably because they simply do not file their individual tax returns. However, for 71% of observations, taxpayer numbers can be linked across P9 and P10, and employment incomes exactly match across sources in most (88 %) of those cases. When a discrepancy exists, we privilege PAYE data on employment income, since those returns are filed by employers, and third-party reporting that has been shown to foster truthful reporting (Kleven et al., 2011), which is in line with the perceptions of Kenyan tax officials of PAYE returns as a more reliable data source. Finally, the small share of observations (4%) for which only personal income returns are available partly consists of people fully deriving their income from other sources than employment, particularly

self-employment.

One conceptually relevant question to consider is what type of income should be considered to preserve construct validity when matching survey and tax data. On the one hand, self-reported personal income is not restricted to income from third-party employment, and includes income derived from a variety of sources including investment, business and professional income. The data at hand do not allow us to disentangle these sources, and some of these income types (essentially those corresponding to capital incomes) are clearly not part of the income concept measured in the KCHS. However, there are two arguments in favor of incorporating non-employment income to our income concept. First, business and professional income is the prevailing source of income by orders of magnitude, remaining dominant even at the top of the distribution.² Second, these income sources not conceptually covered by survey data are very heavily concentrated at the top of the distribution, which we will infer from tax data. It is thus safe to assume that the part of the distribution captured by household surveys would not look very different if households were also asked to report capital income. We thus keep non-employment income as part of our personal income variable.

3 Methodology

3.1 Income distribution in household surveys

In order to combine household survey and administrative tax data, the first challenge is to obtain an estimate of the interpersonal income distribution across survey respondents. Unfortunately, KCHS 2021 only provides reliable data on a subset of them. We use an imputation model for income based on consumption and several relevant individual characteristics:

²This insight is extracted from ancillary data, not linked to PAYE in this study. One of the reasons for this fact is that much of capital income (dividends, rental income, capital gains) is subject to final withholding tax or not included in the personal income tax base and therefore typically not recorded in the personal income return. It should be kept in mind that capital income is likely to be severely underestimated throughout this paper.

$$\log(y_i) = \gamma_0^T + \rho_k^T D_{ik} \log(c_i) + \gamma^T X_i + \epsilon_i \quad (1)$$

$\log(y_i)$ is log income, c_i consumption per working age adult,³ D_{ik} (with $k = 1, \dots, 9$) a set of dummies for deciles in the consumption distribution, which allow us to capture the different saving rates at different levels of the distribution through ρ_k^T . Independent variables X_i include sex, age, education, number of other working adults in the household, area of residence (rural or urban) and information on whether the individual works in the agricultural sector. Separate models are estimated for each type T of workers: employees and the self-employed. We estimate these income models with ordinary least squares on the set of observations R with reliable income data and use them to predict incomes for U , the set including the rest of own-account workers and paid employees outside the household for which information on income is unreliable or not directly available.⁴ Individuals with unreliable income data represent between 6% and 18% of the paid employees and self-employed, being predominantly found among the latter.

Assuming log-normality, we further take a draw from $\mathcal{N}(0, \sigma_\epsilon^{R,T})$, where σ_ϵ^R is the variance estimated for income residuals in R for workers of type T . This last step adds random noise that is uninformative for each household, but it is necessary to account for the variance not explained by observable characteristics and thereby adequately reflect the dispersion of the income distribution.

The validity of this procedure depends on three assumptions:

1. $\mathbb{E}_{R,T}[\log(y_i)|X_i, c_i] = \mathbb{E}_{U,T}[\log(y_i)|X_i, c_i]$
2. $Var_{R,T}(\epsilon_i) = Var_{U,T}(\epsilon_i)$

³Since household income is mainly generated by working adults, consumption per working age adult seems intuitively more strongly linked to income than consumption per capita or per equivalent adult.

⁴We consider this to be the case when income information is missing or monthly income is implausibly low, i.e. below Ksh. 1500, the earnings of a casual agricultural worker for 15 days of work at the lowest daily rate regularly observed in the data. We do not venture prediction of incomes for contributing family workers, apprentices, etc. because zero or very small incomes are more plausible for them. In addition, KCHS 2021 and 2022 do not aim to collect income information from them, and it is likely that the relation to income of the variables used for prediction is different to that for paid employees or the self-employed.

$$3. \epsilon_i^{U,T} \sim \mathcal{N}(0, \sigma_\epsilon^{R,T})$$

Essentially, these assumptions require that the set of individuals for which income is reliably captured in KCHS is representative of that type of workers overall both in terms of the conditional correlation between predictors and income, as well as in the dispersion of the unexplained component of income. These assumptions are still relatively strong and might not be fully satisfied, as workers with a reliable income measure differ in their observable characteristics. They are, for illustration, more likely to be paid monthly rather than daily. It is conceivable that e.g. the association of education with income, or the extent of wage dispersion not explained by observables might differ across these types of workers, and hence between R and U.

Notwithstanding its limitations, this procedure is likely to provide a more reasonable approximation than the most straightforward alternatives, such as direct replacement by per capita consumption. It is well known that consumption distributions are more compressed than income distributions for various reasons, so that this procedure would notably underestimate income inequality and, in particular, underestimate the share of workers with incomes above the tax-exemption threshold. Another simple alternative, directly extrapolating the distribution of income within R to the whole workforce, does not seem convincing given the easily verifiable differences in observables between R and U mentioned above, reflected in clearly visible differences in the distribution of predicted and observed incomes within a given employment category, which our method is able to absorb.⁵

We do not have consumption expenditure data for 2019. Instead, we incorporate housing characteristics (dwelling type, number of rooms, roof, floor and wall materials, source of water and lighting, cooking fuel, etc.) to the prediction model, which results in only slightly lower predictive power. In the case of 2022, using these characteristics instead of consumption does not lead to markedly different income distributions. Sensitivity checks show that variations

⁵Assumption 2 could be eventually relaxed to a conditional version by allowing for and modelling heteroskedasticity. Similarly, the normality assumption is taken for simplicity, and other distributions (including the empirical distribution of residuals) could be easily employed. We consider this procedure sufficiently good as a first-order approximation.

in defining the set of reliable observations, such as imputing very small income observations instead of treating them as daily incomes, or in the imputation model employed, such as having a joint model for employed and self-employed with dummies for employment type, have a rather limited impact on the resulting income distribution.

3.2 Merging household surveys and administrative tax records

We consider several options for integrating household and survey data. The key distinction is whether, following canonical approaches, we only use tax data to correct the top of the distribution, or instead we use it to fully or partly replace the distribution of the formal sector in the household survey by the distribution observed in tax data.

Since we define formality as registration with tax authorities, the distribution of formal incomes provided by administrative data is correct by definition, or at least to the extent that reported incomes correspond to actual annual incomes. It is thus intuitively appealing to rely on administrative data to capture the formal sector, while survey data are considered informative about the informal economy. To combine both data sources, we characterize formal workers in the household survey as those individuals that report paid employment outside the household as their primary income and respond affirmatively to the question whether their employer deducts income tax from their earnings. This implies that this approach is not feasible for KIHBS 2015/16 because information on informality is only available in KCHS. In practice, this option can be implemented simply by merging tax data for formal employees with survey data (rescaled to annual incomes) for informal employees and the self-employed, using the appropriate sampling weights.⁶

It is unclear to what extent this survey question provides a valid link between tax and survey data for lower-income workers whose earnings are below the tax-exemption threshold implicitly defined by the Kenyan personal income tax structure.⁷ Since these workers are not

⁶In general, when we discuss frequencies or number of observations of a given type in one of the surveys, the use of appropriate sampling weights is implied.

⁷All Kenyan taxpayers are entitled to personal relief, a fixed amount that is deducted from their personal

supposed to pay income tax, we might expect to respond negatively to the formality survey question, but to still feature in PAYE data, as companies are expected to report incomes for all workers. In fact, even if there is a non-negligible number of survey formal workers below the threshold,⁸ a heavier left tail is visible in tax data for all survey years (see Figure 1). Other possible reasons behind this tail that would recommend disregarding tax data for these income levels are observations that do not correspond to yearly incomes (i.e. short spells of formal sector employment) and underreporting to tax authorities. Therefore, we consider Method 1B, a variation of the approach above that only relies on tax data to capture formal worker incomes above the implicit tax-exemption threshold.⁹ Household surveys thus provide information on both informal workers and formal workers earning incomes below that threshold.

When implementing these adjustments, we exclude observations from individual tax returns that cannot be matched to PAYE, as they do not seem to correspond to formal workers. In principle, it would be possible to add the upper tail of these observations to the distribution of workers other than formal employees. We do not implement this as a separate option given the very small number of such observations in our data.

We also implement an adjustment more in line with standard methodologies where only the upper tail of the survey is replaced by tax data. The usual assumption is that each of the two data sources represents the income distribution accurately on either side of a merging point (Blanchet, Flores, and Morgan, 2022). Rather than looking at the crossing of densities, we choose a merging point beyond which the *number* of taxpayers (see Figure 2) exceeds (or tends to exceed) the number of individuals reporting those incomes in the survey, as this seems more indicative of the survey failing to include some high-income earners.

tax liability. In practice, this relief defines thus an income level until which no income tax is due, provided that personal relief is applied. In this paper, we refer to this level as (implicit) tax-exemption threshold

⁸This could be explained by lack of information or a different interpretation of the question. We also observe that income tax is deducted in a few cases in PAYE data, even if that would not be necessary if employers correctly applied personal relief.

⁹For the sake of comparability, we take the highest of these thresholds over the study period (KSh. 288,000, corresponding to years 2021 and 2022) for all years.

In all cases, a reweighting step is necessary to account for the fact that by underestimating the mass of formal workers or top income earners, surveys assign excessive weight to (some) other observations. We reweigh survey observations by

$$\kappa = 1 - \frac{\phi_T - \phi_S}{\pi_S - \phi_S}, \quad (2)$$

where the numerator $\phi_T - \phi_S$ reflects the number of workers added from tax data (in Method 1, the difference between the number of formal workers in tax and survey data in the relevant range of the income distribution; in Method 2, additional workers at the top in tax data) and π_S represents the survey population of interest. For Method 2, this is exactly the reweighting step suggested by [Blanchet, Flores, and Morgan \(2022\)](#); and for Method 1, it follows the same logic.

Table 1: Example of a wide-title table with yearly values

	2016	2019	2021	2022
Formal workers: Tax	2,361,301	2,597,664	2,654,056	2,879,937
Formal workers: Survey	-	2,276,237	1,693,932	1,686,232
Taxpayers*: Tax	1,312,237	1,582,129	1,643,770	1,760,521
Taxpayers*: Survey	-	1,277,448	1,213,409	1,158,549
Top Incomes: Tax	227,144	619,934	270,555	297,070
Top Incomes: Survey	178,206	416,978	237,024	217,678
Informal Employees & Self-employed	-	10,714,184	11,162,625	9,385,633

Notes: This Table indicates the number of individuals in various categories and data sources. Formality means tax formality and is self-reported in the survey. “Taxpayers” refers to formal workers with incomes above the 2021/22 tax-exemption threshold, and “Top Incomes” to those above the merging points depicted in Figure 2. Survey totals computed using sampling weights.

The implicit assumption behind all these methods is that formal sector incomes are not accurately measured in the survey, and particularly so in upper segments of the distribution.¹⁰ A strong indication of this fact for the case of Kenya is that the household survey fails to capture the significant increase in the number of taxfilers and taxpayers observed during the study period (see Table 1). The left panel in Figure 1 shows that the distribution of

¹⁰Method 2 incorporates top incomes from sources other than wage employment. In practice, an overwhelming majority of top incomes is found in PAYE.

formal worker income differs across data sources. The right panel of Figure 1 shows that in 2019 we only find more formal workers in tax data from a given point of the income distribution, while for 2021 and 2022 tax data includes more formal workers over the whole income distribution, due to the severe underestimation of the number of formal workers in the survey.

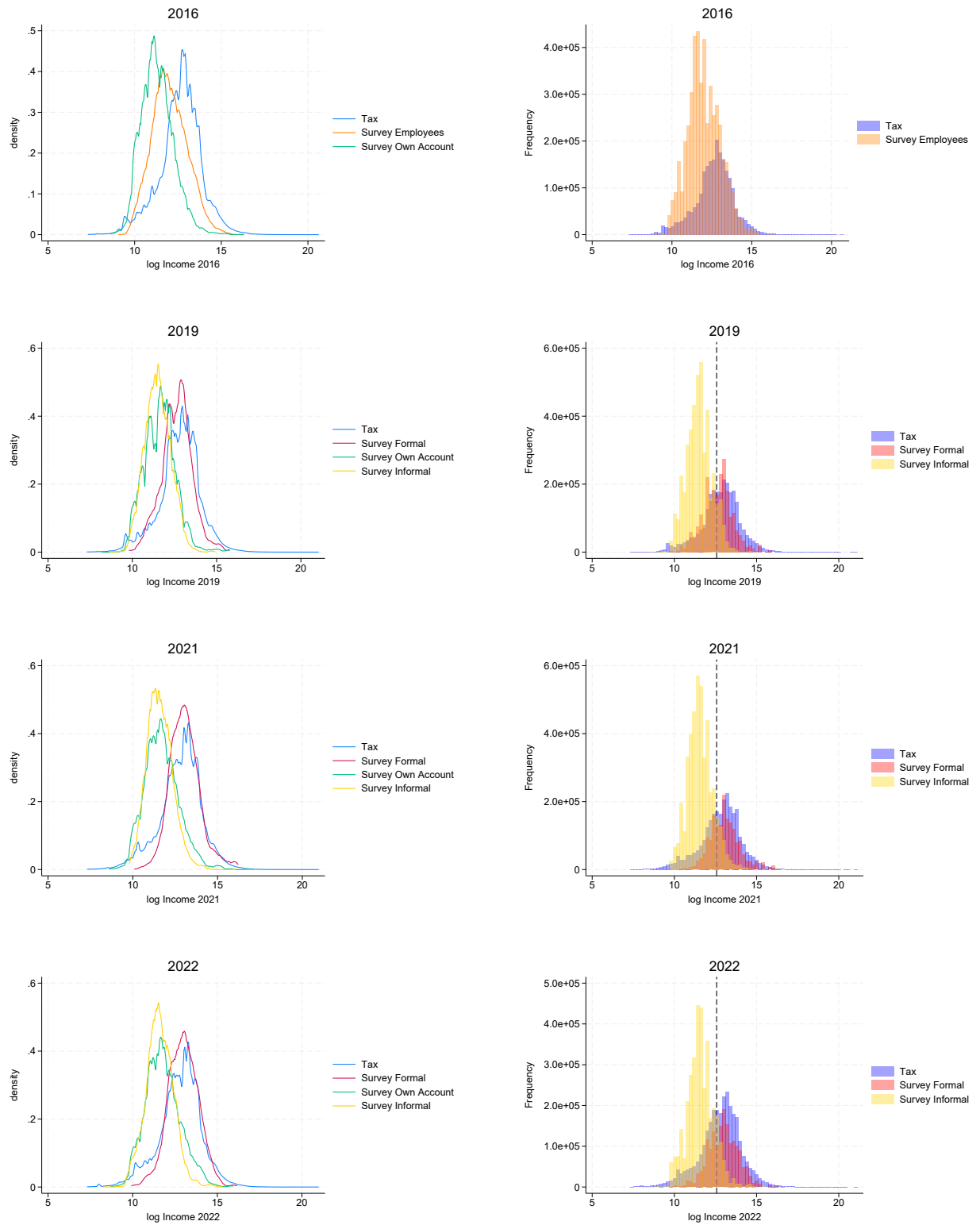
The question of which method is more appropriate can then be posed in terms of why this discrepancy between tax and survey data arises, or more precisely, how it is related to the survey income distribution for informal and own-account workers. Method 1 assumes that missing incomes in the formality variable are unrelated to the income distribution among the self-employed and informal worker. This distribution is thus assumed to be correctly measured in the household survey, even if the share of these workers in the population is overestimated.¹¹ This could be the case if the survey fails to capture these formal incomes altogether, due to e.g. differential non-response not adequately corrected for.

Alternatively, one may suspect that the reason behind ‘missing’ formal employees is less related to obtaining responses from those formal employees than to their classification into that category in the household survey. If this is the case, adjustments as in Method 1 might spuriously increase the density of the distribution in segments where this mislabeling is particularly pronounced. This can be avoided through an adjustment following Method 2 that only relies on tax data at income levels where it includes more individuals than implied by the survey for all employment categories. Of course, this has the drawback of omitting any individuals that are indeed not formal workers in the higher-income segment measured with tax data (see Figure 2). It also leads to potentially misrepresenting (depending on survey accuracy) the distribution of formality in the income range covered by survey data, which is however not necessary problematic as long as researchers are only interested in aggregate indicators for the whole population.

It is not obvious which of these two forms of mismeasurement prevails. On the one hand,

¹¹This is the assumption behind proportional reweighting. We also assume that this excess mass is restricted to the informal and self-employed and does not affect e.g. contributing family workers.

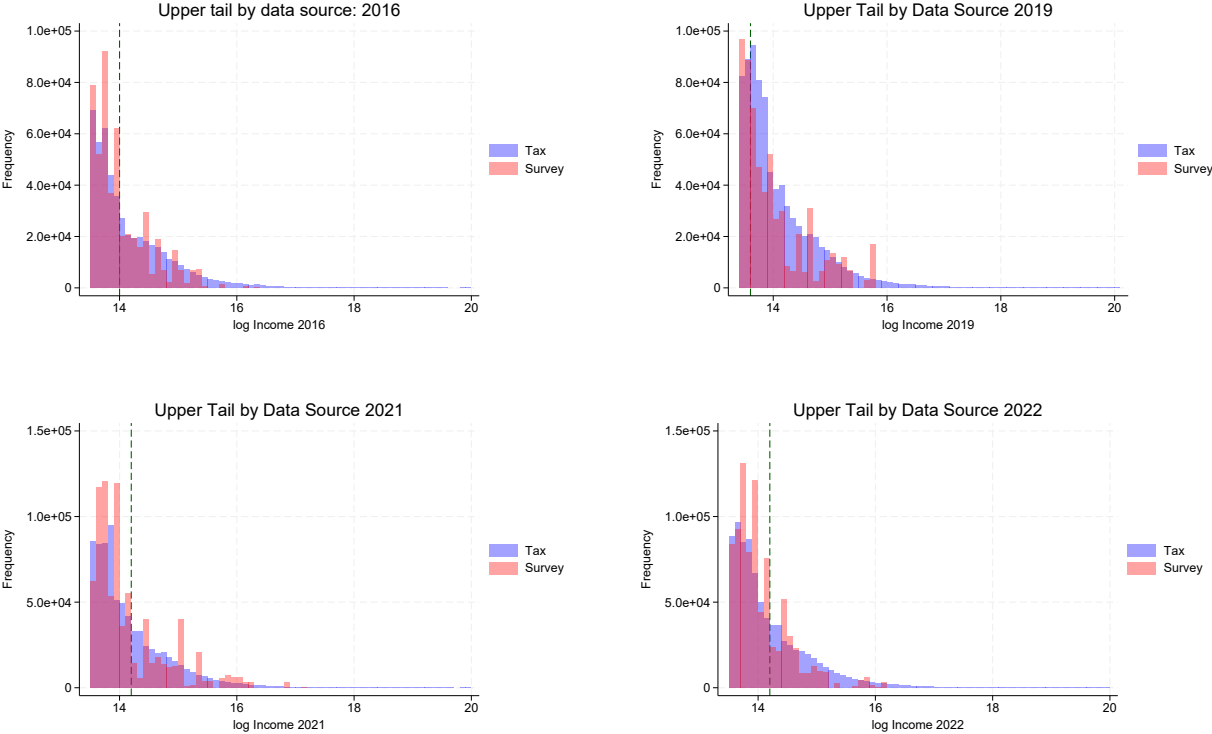
Figure 1: Densities and frequencies by employment and data source



Note: All survey densities and frequencies computed using sampling weights. The dotted line on the right panel represents the threshold used in Method 1B.

the substantial gaps between the number of formal employees found in survey and tax data on the right panel of Figure 1 dissuade us from excessively trusting the formality indicator in the survey. On the other hand, Figure 1 also shows that most of the excess mass of formal workers beyond the tax-exemption threshold cannot be simply attributed to the formality question and, if resulting from classification errors, would arise from inaccurate measurement of the primary economic activity, which is perhaps less prone to errors as it is more salient to respondents and central to the survey.

Figure 2: Income frequency in the right tail by data source



Note: The figure shows the upper tail of the Kenyan income distribution in survey and tax data. Bars indicate the number of observations in the respective data sources at a given income level. The green dotted line denotes the income level selected to implement Method 2.

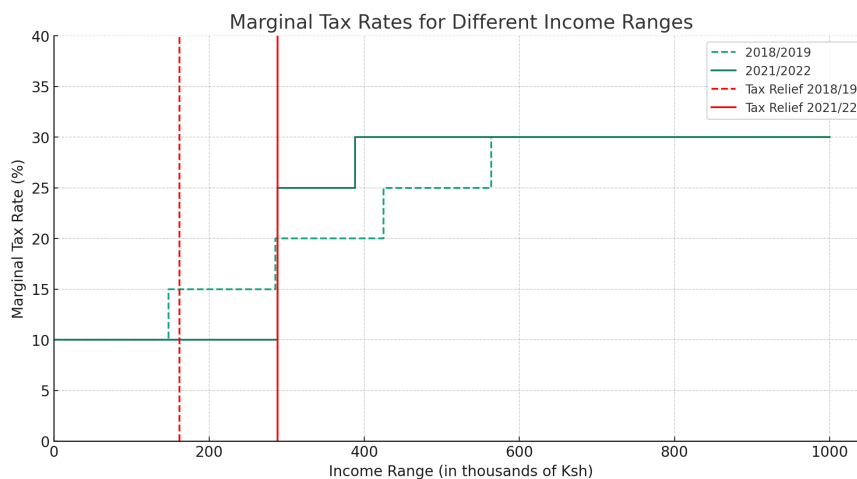
3.3 Personal income taxes and inequality

An obvious advantage of our simple data combination method is that it preserves the distinction between formal and informal workers, allowing for a realistic and straightforward

analysis of the impact of personal income taxes on inequality. We conduct this analysis on the basis of income distributions analogous to those constructed with Method 1B above.¹²

Once surveys and tax data have been combined into a single distribution where the formally employed can be identified, we can assess the impact of taxes on the income distribution by simply applying statutory tax rates on households in formal employment.¹³

Figure 3: Personal Income Tax Rates in Kenya



Note: The figure shows the progression of marginal tax rates along nominal income in 2019 and 2021/2. The vertical red lines represent the tax-exemption thresholds implicitly defined by personal relief.

The Kenyan personal income tax structure (see Figure 3) is based on graduated tax bands with marginal tax rates mildly increasing in income. An additional progressive characteristic is the exemption of low-income earners from paying personal income tax in practice through personal relief. In our simulations, we apply full personal relief to all taxpayers. After a long period of stability where only inflation adjustments were implemented, there were two subsequent reforms between 2019 and 2021.¹⁴

¹²This time, we use the relevant personal relief for 2019 as a cut-off for using tax data. This implies that we use tax data for a somewhat larger portion of the formal worker distribution. We reweigh household survey observations accordingly.

¹³For simplicity and consistency with section 4.1, we directly apply tax rates to the personal income variable analyzed above. This abstracts from tax benefits such as those linked to mortgage payments, insurance and pension contributions. Our analysis also abstracts from social contributions, such as those to the National Hospital Insurance Fund (NHIF) and National Social Security Fund (NSSF).

¹⁴In 2020, policymakers increased personal relief by over 70% and generally lowered marginal tax rates at all income levels in order to shelter taxpayers from economic distress. Tax rates were then increased in 2021 to compensate for decreasing revenues.

4 Results

4.1 Inequality levels and trends: pre-tax income

Table 2: Pre-tax income inequality

Year	Survey	Tax	Method 1A	Method 1B	Method 2
Panel A: Gini Coefficient					
2016	0.576	0.611	-	-	0.623
2019	0.555	0.601	0.629	0.626	0.605
2021	0.614	0.606	0.645	0.643	0.620
2022	0.581	0.613	0.643	0.640	0.636
Panel B: Top 1% Share					
2016	0.135	0.193	-	-	0.214
2019	0.156	0.182	0.211	0.210	0.207
2021	0.197	0.178	0.212	0.210	0.204
2022	0.126	0.185	0.200	0.198	0.211
Panel C: Top 5% Share					
2016	0.324	0.383	-	-	0.398
2019	0.327	0.367	0.411	0.410	0.391
2021	0.387	0.365	0.413	0.411	0.393
2022	0.320	0.375	0.398	0.385	0.402

Note: Authors' calculations on the basis of KCHS and administrative tax data from KRA on personal incomes. See Section 3.2 for a description of the estimation methods. Method 1B uses KSh. 288,000 (the implicit tax-exemption threshold in 2021 and 2022) as a starting point for tax data in all years.

Table 2 displays the implications of data combination for the measurement of income inequality in Kenya. Gini coefficients are already remarkably high in each of the two data sources when considered separately, but still underestimated with respect to combined data, which yields values between 0.60 and 0.65. Estimates of top income shares from survey data are very noisy, probably due to sparsity of the right tail, and severely underestimated. Tax data on their own show higher inequality than survey data, but still underestimate top income shares by 1.5 to 3.5 percentage points. Using data combination methods, we estimate the 1% and 5% income share at levels around 20% and 40%, respectively.

Regarding differences across data combination methods, their levels are (as expected)

most similar for measures focused at the very top. The variations of Method 1 lead to generally higher Top 5% shares and Gini coefficients. Relying on tax data to represent formal workers also below the tax-exemption threshold has no perceptible effects. Methods taking the formal sector distribution from tax data show an inverse U-shaped pattern between 2019 and 2022, with all inequality indicators peaking in 2021 and decreasing thereafter.¹⁵ In contrast, when only the upper tail is corrected with tax data, income seems to have become increasingly concentrated also after 2021.

4.2 Inequality levels and trends: post-tax income

Table 3 shows our results on the impact of personal income taxes on the income distribution, reporting post-tax interpersonal inequality and the difference to pre-tax income inequality. Combined data indicate that the personal income tax reduces the Gini coefficient by 3 to 4 points - and hence by between 5% and 6% of its pre-tax levels. The relative impact on top income shares is much larger, as they are reduced by over 10%. It makes sense that top-sensitive measures react more strongly to personal income taxes when considering that the upper segments of the income distribution generate the greater part of tax revenues.

Table 3 also shows that combining tax and survey data can substantially alter the conclusions on the overall impact of personal income tax on inequality. Survey data downplay the reduction of the Gini coefficient caused by taxes, as they underestimate high-income taxpayers, and hence also the average tax rates they face. In contrast, tax data overestimate this effect by ignoring the large share of informal income earners that are unaffected by tax rate changes. Due to their difficulties in capturing the top of the distribution, surveys lead to noisy and too small estimates of the impact of taxes on top 1% shares. In tax data, this impact is more stable and larger, but still below the effects measured through data combination - in this case, probably because tax data overestimate average incomes.

¹⁵For top income shares, this decline might be related to the smaller share of economically active individuals captured by KCHS 2022 as compared to previous rounds, which also results into less informal and self-employed individuals (see Table 1). Since the formal sector is taken from tax data and hence independent from surveys, this results into higher average income.

Table 3: Post-tax income inequality

Year	Survey		Tax		Combined	
	Net	Δ	Net	Δ	Net	Δ
Panel A: Gini coefficient						
2019	0.532	-0.023	0.555	-0.046	0.589	-0.037
2021	0.588	-0.025	0.557	-0.049	0.610	-0.033
2022	0.558	-0.023	0.564	-0.049	0.604	-0.037
Panel B: Top 1% Share						
2019	0.149	-0.007	0.162	-0.019	0.185	-0.025
2021	0.181	-0.016	0.158	-0.020	0.189	-0.022
2022	0.118	-0.009	0.165	-0.020	0.168	-0.030
Panel C: Top 5% Share						
2019	0.284	-0.043	0.332	-0.034	0.365	-0.045
2021	0.357	-0.030	0.330	-0.036	0.373	-0.038
2022	0.299	-0.021	0.338	-0.037	0.353	-0.032

Notes: “Net” refers to post-tax values after applying the corresponding personal income tax structure to formal workers’ pre-tax income. Δ indicates the change relative to pre-tax indicators. Combined data is based on Method 1B above, but using the contemporary tax-exemption threshold for 2019.

5 Conclusion

This paper combines household survey data and administrative tax data with the aim of shedding light on the levels and evolution of income inequality in Kenya between 2016 and 2022. While certain caveats may exist around the scope and quality of the household survey income data and the nature of the tax data we use to correct these surveys, our data combination exercise reveals very high levels of interpersonal employment income inequality.

Our focus on the interpersonal distribution of employment income entails several limitations, mostly imposed by data availability and comparability considerations. One of these limitations is ignoring the household structure, the unit across which resources are typically pooled. A further limitation is the exclusion of contributing family workers, which might understate the degree of inequality in the economically active population and may also lead

to undesirable properties for inequality measures.¹⁶ Finally, ignoring sources of income that tend to be concentrated at the extremes of the distribution and are thus important to inequality, such as transfers and capital income, may result in misleading estimates.

Some of these limitations might be best addressed simultaneously. It should be possible to adjust the methods explored in this paper to perform a rank-preserving replacement of incomes for the upper tail of (formal) workers in the survey with those observed in tax data, allowing for household-level inequality estimates that incorporate tax data. The impact of uncertainty around the incomes of contributing household workers might be diluted when adding those to the incomes of other household members. In addition, KIHBS 2015/16 provides household-level information on other relevant income sources that may enable us to extend this exercise to a broader income concept. Since capital income is typically poorly captured by household surveys, this should be ideally reinforced by additional tax data on capital income.

Considered from an alternative perspective, it is remarkable that we find substantial effects of data combination on earnings inequality estimates, and in particular top income shares, even when essentially relying on labor income. We barely capture any capital income, which has centered much of the recent discussion about inequality and has been singled out as a key source of discrepancy between survey-based inequality estimates and those based on other sources (Burdín et al., 2022; De Rosa, Flores, and Morgan, 2024; Flores, 2021). It may seem that, for Kenyan household surveys, the ‘missing rich’ also include wage employees. It might be that differential non-response extends well into the first centiles of the income distribution. Another possible explanation is given by the fact that a substantial share of Kenyan top-earners still derive an important share of their income as income from employment.¹⁷

¹⁶For instance, the incorporation of a contributing family worker to a low-paid job would be likely to increase inequality with our current measures. However, the impact of this exclusion might be limited. In 2019, when data for most contributing household workers is collected, the Gini coefficient in the survey only increases by one point when incorporating them to the calculations.

¹⁷This is based on personal communication with KRA officials specialized in the taxation of High Net Worth Individuals

We hope to have illustrated that data combination methods can be helpful in addressing some of the limitations suffered by individual sources in data-scarce environments. However, the compilation of reliable income inequality estimates requires an intensification of efforts to collect high-quality data on a broad definition of income with increased frequency. The very high levels of income inequality found in this study underscore that this is not an issue that can be deemed irrelevant for developing countries.

References

- Alvaredo, Facundo, Anthony B Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. “The Top 1 Percent in International and Historical Perspective.” *Journal of Economic Perspectives* 27 (3):3–20.
- Assouad, Lydia. 2023. “Rethinking the Lebanese economic miracle: The extreme concentration of income and wealth in Lebanon, 2005–2014.” *Journal of Development Economics* 161:103003.
- Atkinson, Anthony B. and Thomas Piketty. 2007. *Top incomes over the twentieth century: A contrast between continental European and English-speaking countries*. Oxford University Press.
- Bartels, Charlotte and Maria Metzger. 2018. “An integrated approach for a top-corrected income distribution.” *The Journal of Economic Inequality* 17 (2):125–143.
- Blanchet, Thomas, Ignacio Flores, and Marc Morgan. 2022. “The weight of the rich: improving surveys using tax data.” *The Journal of Economic Inequality* 20 (1):119–150.
- Burdín, Gabriel, Mauricio De Rosa, Andrea Vigorito, and Joan Vilá. 2022. “Falling inequality and the growing capital income share: Reconciling divergent trends in survey and tax data.” *World Development* 152:105783.
- Chancel, Lucas, Denis Cogneau, Amory Gethin, Alix Myczkowski, and Anne-Sophie Robilliard. 2023. “Income inequality in Africa, 1990–2019: Measurement, patterns, determinants.” *World Development* 163:106162.
- Chancel, Lucas and Thomas Piketty. 2019. “Indian Income Inequality, 1922–2015: From British Raj to Billionaire Raj?” *Review of Income and Wealth* 65 (S1).
- Chatterjee, Aroop, L’eo Czajka, and Amory Gethin. 2023. “Redistribution without Inclusion? Inequality in South Africa Since the End of Apartheid.” URL <https://leo-czajka.eu/wp-content/uploads/2023/10/ChatterjeeCzajkaGethin2023.pdf>.
- Czajka, Léo. 2017. “Income Inequality in Côte d’Ivoire.” , WID World Working Paper Series, 2017/8.
- De Rosa, Mauricio, Ignacio Flores, and Marc Morgan. 2024. “More unequal or not as rich? Revisiting the Latin American exception.” *World Development* 184:106737.
- Flores, Ignacio. 2021. “The capital share and income inequality: Increasing gaps between micro and macro-data.” *The Journal of Economic Inequality* 19 (4):685–706.
- Garbinti, Bertrand, Jonathan Goupille-Lebret, and Thomas Piketty. 2018. “Income inequality in France, 1900–2014: Evidence from Distributional National Accounts (DINA).” *Journal of Public Economics* 162:63–77.

- Günther, Isabel and Andrey Launov. 2012. “Informal employment in developing countries.” *Journal of Development Economics* 97 (1):88–98.
- Inchauste, G. and N. Lustig, editors. 2017. *The Distributional Impact of Taxes and Transfers: Evidence from Eight Low and Middle-Income Countries*. Washington DC: The World Bank.
- Jenkins, Stephen P. 2016. “Pareto Models, Top Incomes and Recent Trends in UK Income Inequality.” *Economica* 84 (334):261–289.
- Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, S. Pedersen, and E. Saez. 2011. “Unwilling or Unable to Cheat? Evidence From a Tax Audit Experiment in Denmark.” *Econometrica* 79 (3):651–692.
- Lastunen, Jesse, Antoine de Mahieu, Katrin Gasior, H. Xavier Jara, and Jukka Pirttilä. 2024. *Microsimulation of tax-benefit systems in the Global South: a comparative assessment*.
- Lustig, Nora. 2019. “The Missing Rich in Household Surveys: Causes and Correction Approaches.” , Commitment To Equity Working Paper 75.
- Lustig, Nora and Andrea Vigorito. 2025. “Including the Rich in Income Inequality Measures: An Assessment of Correction Approaches.”
- Piketty, Thomas, Emmanuel Saez, and Gabriel Zucman. 2017. “Distributional National Accounts: Methods and Estimates for the United States*.” *The Quarterly Journal of Economics* 133 (2):553–609.
- Ravallion, Martin. 2016. *The Economics of Poverty. History, Measurement, and Policy*. Oxford University Press.
- Zwijnenburg, Jorrit. 2019. “Unequal Distributions: EG DNA versus DINA Approach.” *AEA Papers and Proceedings* 109:296–301.