Income Inequality in India, 2000-2020

Anmol Somanchi [*]

June 2, 2023

Abstract

Tracking the dynamics of income inequality in India is fraught with serious challenges related to the quality and availability of data. Combining national accounts aggregates, tabulated tax data and household surveys, Chancel and Piketty (2019) develop a measurement framework suited to India and present harmonized long-run inequality estimates. The key roadblock to extend the series to recent years is the lack of any reliable household survey on consumption or income after 2011-12. In this paper, using alternate survey sources and tax data, I begin by developing a measurement approach which allows us to study income dynamics over the 2000-2020 period. As per benchmark estimates, I find that the share of national income going to the top 10% has risen from 39.9% in 2000-01 to 59.5% in 2019-20 while the share of the top 1% has grown from 15.1% to 25.1% over the same period. Based on two alternate approaches that relax certain measurement assumptions, I verify that the these estimates are largely robust to different ways of estimating bottom incomes from surveys. These results would imply that nearly two-thirds of the real-income growth between 2000-2020 was captured by the top 10% and nearly a third by the top 1% alone. Lastly, given the declining quality and availability of data in recent years, I highlight limitations with the current estimation approaches and how these can be improved upon by incorporating additional data sources.

JEL codes: D31, N35, O15

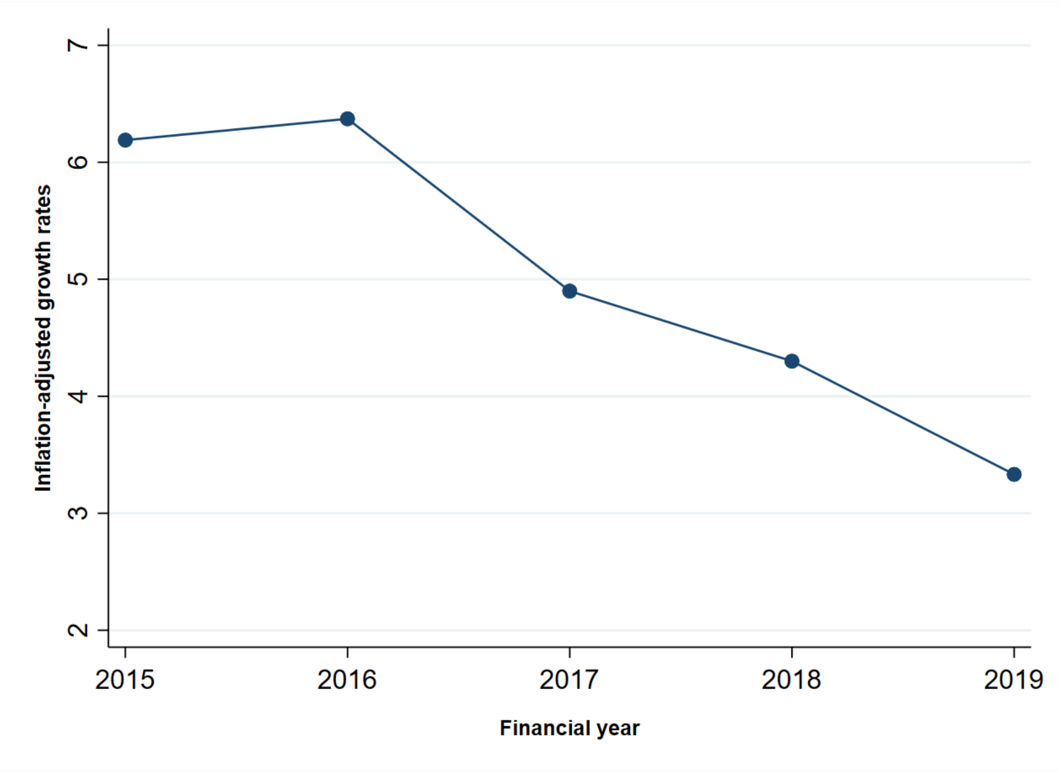
Key words: Inequality; India; Top Incomes

[*] Paris School of Economics and EHESS (anmol.smnch@protonmail.com). I am very grateful to Lucas Chancel and Thomas Piketty for their motivation, patient guidance and invaluable suggestions throughout the study as well as for making their data files accessible to me. I am also very grateful to David Margolis for agreeing to referee this work. Lastly, I would like to thank Nitin Bharti, Thomas Blanchet, and Abhijit Tagade for helpful discussions and clarifications, and Francois Fontaine and Thomas Breda for thoughtful comments and suggestions at the Public & Labour Research Seminar at PSE.

1 Introduction

Given its large geographical size and population - now the highest in the world - the distribution of economic growth in India has significant consequences on global inequality dynamics, which in turn is crucial to our understanding and evaluation of global economic arrangements. This makes the careful measurement of income inequality in India an important exercise. However, tracking the dynamics of income inequality in India is fraught with serious challenges related to the quality and availability of data. These issues have intensified in recent years because various key data sources have become unavailable in the most recent decade. At the same time, recent years have seen important political and economic changes in India. In 2014, the right-wing Bharatiya Janata Party (BJP) came to power at the centre with a sweeping majority after 60-years of nearly un-interrputed rule by the left-of-centre Indian National Congress (INC). While the BJP came to power on a developmental agenda with a mandate for economic reforms, many observers believe that the move to an authoritarian government with centralization of decision-making power coupled with the prevailing nexus between big-business and government (Banaji, 2022) has led to a situation that one observer has termed "conglomorate capitalism" (Damodaran, 2020) and another a "conclave economy" (Bardhan, 2022). At the same time, the available evidence suggests that economic performance has been relatively sluggish, especially post-2016 with growth rates declining (see Figure 1), investment falling, unemployment rising, and real wages stagnating (Ghatak and Mukherjee, 2019; Nagaraj, 2020a; Drèze, 2023). These indicators paint a grim picture regarding the state of economic affairs for a majority of India's poor. The BJP government has certainly invested in expanding the coverage of various infrastructural benefits like housing, toilets, electricity, banking, what some have termed the 'new welfarism of India's right' (Subramanian et al., 2020), but it is unclear if these investments have led to an improvement in incomes and purchasing power on the market. At the same time, certain policy choices like the harsh "demonetization" shock delivered to the economy in November 2016 - when nearly 86% of currency in circulation was banned overnight in a move to challenge corruption - is believed to have disproportionately hurt the poor and the small-medium businesses in particular with credible estimates suggesting that short-term GDP fell by 2 percentage points (Chodorow-Reich et al., 2019). Understanding the distribution of economic growth in recent years is crucial for a thorough evaluation of these key political and economic changes in India in recent years. This paper is an attempt at filling that gap for the most recent years, 2015-2020. This further allows me to study the distributional dynamics over the 2000-2020 period.

Figure 1: **Growth rate of per-adult net national income**



Source: *Economic Survey 2022-23, Ministry of Finance, Government of India and UN adult population estimates. Nominal values deflated to real-terms using the GDP Deflator sourced from the World Bank's World Development Indicators.*

Building on the pioneering work by Piketty (2014), the Distributional National Accounts (DINA) project at the World Inequality Lab (WIL) has aimed to rigorously estimate the income distribution in various countries using a range of data sources and cutting-edge statistical methods (World Inequality Lab, 2021). For developed economies like the United States (US) and France, quasi-exhaustive tax microdata covering majority of the population forms the core basis for estimating the income distribution. Other sources like households surveys are used to fill-in pieces of information like returns on assets classes, ownership of certain assets etc. which are missing on tax returns (Piketty et al., 2017; Garbinti et al., 2018). In contrast, inequality measurement in a developing country like India is fraught with various challenges relating to data coverage, quality, and availability. To begin with, the large-scale relevance of informal employment, relatively low-income levels, and relatively high thresholds for non-taxable incomes means that tax data (in whatever form it is available) covers only a tiny fraction of the adult population - less than 10% as recently as 2017. The limited coverage means tax data can at best shed light on top incomes. Consequently, nationally-representative household surveys conducted periodically by the National Sample Survey Organization (NSSO) formed the basis for studying issues like poverty and inequality (Deaton and Dreze, 2002). However, there are two issues with using NSSO surveys for income inequality measurement. First, NSSO steered clear from measuring incomes (on the grounds that agricultural incomes - which are highly seasonal - are hard to decipher) and instead focused on measuring consumption expenditure. While certainly a good proxy for incomes, consumption dynamics significantly differ from income dynamics and in particular given that the rich only consumer a fraction of their incomes, consumption inequality tends to understate income inequality. Second, household surveys in general (not just in India) are known to be fraught with concerns of under-reporting and differential non-response by income (Bourguignon, 2018; Korinek et al., 2006). This has meant that surveys alone are not sufficient for shedding light on inequality where the right-tail mattes a lot.

Banerjee and Piketty (2005) were the first to mobilize annual tax tables provided by the Income Tax authorities in combination with national accounts to shed light on inequality dynamics over the long run (1922-2000). However, their analysis was restricted to estimating only very top income shares (1%, 0.1%, 0.01%). Building on this, the most comprehensive analysis on income inequality in India is found in Chancel and Piketty (2019). They combine national accounts aggregates, household surveys on consumption and income and tax tabulations to present harmonized long-run estimates of income inequality between 1922 and 2014. Their comprehensive measurement framework allowed them to distribute the national income not just to the very top but across the full income distribution. Their key results are as follows. Inequality levels were high under Colonial rule with top 1% shares in excess of 20% in the late 1930s. Post-independence, between 1950 and 1980, top 1% and top 10% shares fell owing to broadly socialist policies adopted by the government. Since the 1980s however, inequality began to creep back up and has sky-rocketed especially since the early 2000s. By 2014, the top 1% share was higher than its Colonial-era peak.

Given that the Chancel and Piketty (2019) measurement framework is in place, one would think extending it to years beyond 2014 would be a straightforward task. Except, it is not. This is mainly because many of the key data sources that served as inputs to the measurement framework have been suppressed or discontinued by the government. Most crucial of these perhaps is the fact that the most recent nationally representative survey on consumption in India (2011-12) is now more than a decade old. A consumption survey round was conducted by the NSSO in 2017-18, but the microdata and a preliminary report drafted by NSSO were suppressed on rather unqualified grounds of data quality (Press Information Bureau, 2020). On the other hand, reports leaked to the media suggest that the preliminary findings suggested that mean consumption expenditure in rural areas declined between 2011-12 and 2017-18. Given that the NSSO consumption surveys were the primary source for understanding the economic progress for a majority of Indian households, the suppression of the 2017-18 round has made understanding the dynamics of incomes quite difficult.

Two new survey sources have become available in recent years collecting some information on consumption and incomes, but they come with issues for inequality measurement. The first of these, the Periodic Labour Force Survey (PLFS), a nationally representative labour force survey collects information on labour incomes but not non-labour incomes. Ignoring the latter can have serious consequences on income inequality measures given that non-labour incomes tends to be more concentrated. PLFS also collects basic information of 'usual' consumption expenditure but due to differences in coverage of questions and the survey instrument, the consumption reported in PLFS is not comparable to earlier CES surveys. The other source, the Consumer Pyramids Household Survey (CPHS), is a large-scale privately-executed all-India survey which collects data on both labour and non-labour incomes as well as consumption expenditures. However, the CPHS sample is not a representative sample and in particular it under-represents women, young children and the poor (Drèze and Somanchi, 2021; Somanchi, 2021). At the same time, the coverage of commodities for which expenditures are recorded again does not match the NSSO CES survey coverage. For these reasons, neither PLFS nor CPHS is independently sufficient for measuring income inequality.

In this paper, I present three different approaches for combining information in these new household surveys to reliably reconstruct an income distribution. In the first approach, I am able to construct NSSO CES-comparable consumption estimates from those reported in PLFS by exploiting the fact that both CES and PLFS rounds coincided in 2017-18 - this allows me to construct a mapping between the two distinct consumption concepts which I can apply to subsequent PLFS rounds as well. This approach is my benchmark estimates as it allows me to essentially replicate the methodology in Chancel and Piketty (2019) allowing for clearer inter-temporal comparisons. In the second approach, I start with the representative PLFS sample which contains information on labour incomes and use statistical matching methods - coarsened exact matching and multivariate distance matching - to import non-labour incomes from the (biased) CPHS sample by matching every PLFS households to a corresponding CPHS household. Finally, in the third approach, I make use of various socio-demographic information in PLFS as well as other nationally-representative household surveys to calibrate the sampling weights of the CPHS sample to correct for various observable biases. Approaches two and three serve as alternate strategies for estimating survey-based incomes as robustness checks for the recent years. As per my benchmark estimates, the share of national income going to the top 10% has increased from 56.1% to 59.6% between 2014 and 2019. Estimates from the alternate strategies suggest similar estimates. These results would cement India's place as one of the most unequal countries in the world (Assouad et al., 2018).

Before moving onto present the data sources and methodology, I briefly present some of the recent attempts to measure income inequality in India. In the absence of reliable survey data, the literature has been sparse, nonetheless, some attempts have been made using data sources from recent years to track movement in incomes. The 'State of Inequality in India Report' (Kapoor and Duggal, 2022), commissioned by the Economic Advisory Council to the Prime Minister, provides a broad-based review of inequality in India along various dimensions including income, health, education, and household assets. On income inequality, the report acknowledges the extreme inequality reported in Chancel and Piketty (2019) but goes on to present its own analysis for using data from the Periodic Labour Force Surveys (PLFS). They find that the top 1% earn just over 6-7% of the total incomes, much lower than the figures in Chancel and Piketty (2019). The analysis in the report is however fraught with several issues - it only uses data on labour incomes, restricted to those employed at the time of survey (non-zero incomes), and does not make use of tax tabulations. Nonetheless, the report draws attention to the fact that "*the benefits of growth have been concentrated and has marginalised the poor further*".

In another recent contribution, Ghatak et al. (2022) use different survey sources to study inequality in India. Based on data from the privately-executed Consumer Pyramids Household Survey (CPHS; 2014-2020) and the National Sample Survey Organization's Consumption Expenditure Survey (CES; 2017-18), the authors argue that the slowdown in macroeconomic growth starting 2017 may have led to a reduction in income inequality as well. However, they too do not make use of tax data. Moreover, the CPHS dataset is not a representative sample as it is found to under-represent the poor with the bias *growing over time* (Drèze and Somanchi, 2021; Somanchi, 2021). This poses a serious challenge when trying to use CPHS to study trends over time. The same critique applies to the analysis of Gupta et al. (2022) who use the (biased) CPHS sample to argue that income (and consumption) inequality declined during Covid in India.

On the other hand, other evidence points to serious economic inequalities in India. Anand and Thampi (2021) document significant wage gaps for women, scheduled castes and persons in rural areas based on labour force surveys as well as significant shortfalls in ownership to basic household amenities. Khera and Yadav (2020) report starkly unequal median-to-top pay ratios among NIFTY50 companies based on legally-mandated disclosures. Pai and Vats (2023) note that there is a "*brutal power law in most Indian consumer transactions*" - based on various data sources for consumer spending, they find that 1% of Indians take 45% of flights, only 2.6% of Indians invest in mutual funds, 6.5% of users are responsible for 44% of digital transactions on the Unified Payment Interface (UPI), and 5% of users account for a third of the orders placed on Zomato (most prominent food-delivery application). At the same time, as per the annual Forbes rich lists, the net worth of 100 richest Indians has grown by over 350% cumulatively between 2014 and 2022 in real terms. During the same period, the total net national income grew by just over 35%.

The divergence between survey sources and other evidence is striking. However, it may not be very surprising given that surveys are known to notoriously miss the right-tail of income distributions and it may be possible that the issue of differential non-response among the rich may have gotten worse overtime with the proliferation of gated communities in recent years (Verghese, 2021) combined with possibly greater concentration at the very top. Given this divergence, reliance on surveys alone is unlikely to provide a complete picture on income inequality dynamics. To overcome this challenge, I attempt to estimate the income distribution in India in recent years using a range of sources combined in a consistent manner based on the framework and guidelines laid out in World Inequality Lab (2021). However, it is worth noting at the outset that despite to access to new data sources, numerous measurement challenges remain. Moreover, the quality of data has declined further in recent years. Therefore, the results must be interpreted with caution and are best read as a first-step towards a more comprehensive understanding of inequality in India.

2 Data

In line with the framework developed in Chancel and Piketty (2019), I combine data from various sources, but most importantly national accounts, household surveys and tabulated data on income tax returns. Table 1 summarizes the full set of data used in this paper.

Population aggregates: Adult (20+) population figures are sourced from United Nations World Population Prospects (UN-WPP). The age cut-off for defining adults as well as the data source are chosen in line with DINA guidelines (World Inequality Lab, 2021) to allow for consistent cross-country comparisons.

National accounts aggregates: Total net national income (NNI) is obtained from the Statistical Appendix of the Economic Survey 2022-23 published by the Ministry of Finance (MoF), Government of India (GoI). I divide the total NNI by the total adults to arrive at per-adult NNI. Then I assume that a set of deductions apply (eg. to account retained earnings of corporates) to go from NNI to fiscal income (Atkinson, 2007). In practice, I assume that roughly 70% of net national income accounts for fiscal income.

Price indices: I source annual series for the Gross Domestic Product (GDP) deflator from the World Bank World Development Indicators and for the Consumer Price Index (CPI; Rural + Urban combined) from the Reserve Bank of India (RBI). I prefer to use the GDP deflator to convert nominal to real values (discussed further below) but I also test the robustness of our results to using the CPI instead.

Table 1: **Summary of data and sources for recent years**

Data	**Source**	**Years**	**Data type**
Adult (20+) Population	UN-WPP	2011-2020	Aggregates
Consumer Price Index (Combined)	RBI-HSIE	2011-2020	Aggregates
Gross Domestic Product Deflator	WB-WDI	2011-2020	Aggregates
National Income Accounts (NIA)	MoF, GoI	2011-2020	Aggregates
Income Tax Time Series Data (ITSD)	MoF, GoI	2011-2017	Aggregates
Income Tax Returns Statistics (ITRS)	MoF, GoI	2011-2017	Tabulated
Income Tax e-filing Statistics (ITeS)	MoF, GoI	2012-2020	Tabulated
Consumption Expenditure Survey (CES)	NSSO	2017	Tabulated
Periodic Labour Force Survey (PLFS)	NSSO	2017-2020	Microdata
Consumer Pyramid Household Survey (CPHS)	CMIE	2014-2020	Microdata
National Family Health Survey (NFHS)	IIPS	2015 & 2019	Microdata
India Human Development Survey (IHDS)	NCAER	2005 & 2011	Microdata

Notes: *(1) UN-WPP - United Nations World Population Prospects; MoF - Ministry of Finance; GoI - Government of India; RBI-HSIE - Reserve Bank of India Handbook of Statistics on Indian Economy; WB-WDI - World Bank World Development Indicators; NSSO - National Sample Survey Organization; CMIE - Centre for Monitoring Indian Economy Pvt. Ltd.; IIPS - Indian Institute of Population Sciences; NCAER - National Council of Applied Economic Research. (2) Adult population projections are sourced from the UN World Population Projections (2019 release). (3) National accounts aggregates (net national income) sourced from the Statistical Appendix of the Economic Survey 2021-22. (4) All datasets more-or-less correspond to financial year (so 2011 here means FY2011-12), with the exception of UN-WPP and WB-WDI which are available by calendar year and CPHS which is available for every 4-month "wave" starting January 2014. (5) The distinction between 'aggregates' and 'tabulated' data is that the latter contains some distributional information (eg. fractile thresholds and/or averages).*

Tabulated tax data: Aggregate data on 'total effective tax payers', defined as total tax return filers plus non-filers who paid tax at source, is sourced from Income Tax 'Time Series Data' released by the Income Tax department (IT Dept.), MoF. The IT Dept. also released data on total number of tax filers and total income assessed dis-aggregated by 25 income brackets in the publication 'Income Tax Return Statistics' (ITRS).[1] More recently, some information on total 'e-returns' are made available on the dashboard on the e-filing portal of the IT Dept.[2] These sources allow us to construct a distribution of top income earners. For the rest of the distribution, I rely on survey sources.

Household surveys: While the microdata of the Consumption Expenditure Survey (CES) 2017-18 round was suppressed by the government, tabulated data was available to the public by Subramanian (2019c,b), which I make use of to extract the full consumption distribution for 2017-18 (described below). Besides CES, I also rely on two new surveys with information on consumption and incomes which have become available in recent years - the Periodic Labour Force Survey (PLFS) conducted by NSSO for the years 2017-2020 and the Consumer Pyramids Household Survey (CPHS) conducted by CMIE for the years 2014-2020. I make use of data on socio-economic indicators, demographic trends and ownership of assets from fourth and fifth rounds of the National Family Health Survey (NFHS). Lastly, I use data on consumption and income in the India Human Development Survey (IHDS) conducted in 2005 and 2011-12 by the National Council for Applied Economics Research (NCAER). The precise role each of these datasets play is detailed in the next section.

3 Methodology

Very broadly speaking, the idea is to take national account aggregates of total income and combine it with estimates of income at various points of the income distribution. As noted above, in some Western economies like France and the USA, quasi-exhaustive tax data forms the main basis for estimating the income distribution and is supplemented by survey-based data to fill in specific pieces of information missing in tax returns like ownership of certain types of assets, market rate of returns, etc. (Piketty et al., 2017; Garbinti et al., 2018). On the other hand, only $\sim$ 10% of the population in India paid income taxes in recent years. While household surveys provide much wider coverage, they are known to be prone to non-response bias and under-reporting of incomes among the rich, both of which are likely to be a serious concern for inequality measurement. The idea then is to consistently combine national accounts aggregates with top incomes from tax data and data on middle and low incomes from survey data.

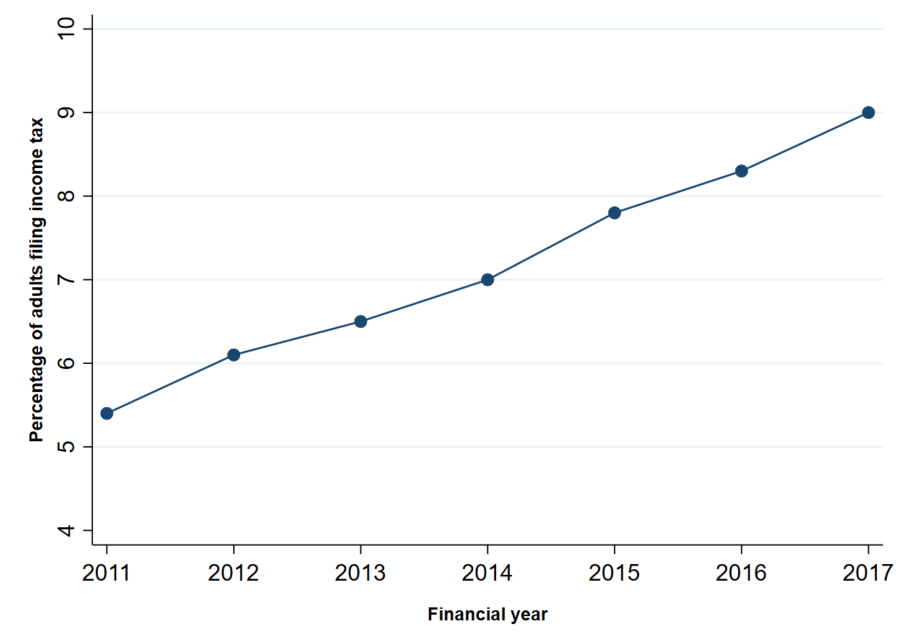
[1] Both the ITRS and the Time Series Data are available here:
`https://incometaxindia.gov.in/Pages/Direct-Taxes-Data.aspx.`

[2] Available here: `https://www.incometax.gov.in/iec/foportal/statistics-data.`

3.1 Top Fiscal Incomes: Tax Tabulations

The Income Tax Department (IT Dept.) of the Ministry of Finance has been publishing annual tabulated data on income tax returns. The annual publication of the "Income Tax Return Statistics" (ITRS) starting 2011-12 provides *tabulated data* on the total number of income tax returns filed and the total income assessed disaggregated for 25 income brackets and by type of tax filer (individuals, corporation, etc.). I consider individuals and Hindu Undivided Families (HUFs) as the tax units relevant for our analysis and combine their data at the income tax bracket level for our analysis. In a separate annual publication titled "Income Tax Time Series Data" (ITSD), the IT Dept. provides data on *total effective tax payers* defined as the sum of units that filed income tax returns and those that paid tax at source but did not file a return. By comparing the total filers in the ITRS data with the total effective tax payers in the ITSD, I get an estimate of the *non-filers* which are missing from the ITRS data. I scale up the total returns in the ITRS data to account for these.[3] Combining the estimates of total effective tax filers with the adult (20+) population, I find that the share of tax-filers has considerably increased, from 5.4% in 2011 to 9% in 2017.

Figure 2: **Growth of income-tax filers, 2011-2017**



Source: *Author's estimates based on ITRS and ITSD tax data.*

[3] Since ITSD does not provide data on effective tax payers disaggregated by tax bracket, I apply a constant scaling ratio across all brackets for a given year. Chancel and Piketty (2019) test different ways of distributing the non-filers, for instance, assuming that they all fall in the lowest brackets, and find they do not make much difference.

While aggregates are useful, from this tabulated tax data I would like to extract a distribution of top fiscal incomes. To do so, I rely on generalized Pareto interpolation (GPinter) methods developed and described in Blanchet et al. (2022b). Going back to Pareto's empirical observation that top incomes are often well approximated by a *Power law*, various interpolation methods rely on the assumption that top incomes are well approximated by a distribution with a probability density function (PDF) and complementary cumulative distribution function (CCDF) of the form:

$$\text{PDF: } \mathbb{P}(Y \in [y, y+dy]) = f_Y(y) = \left(\frac{\alpha}{y_0}\right)\left(\frac{y_0}{y}\right)^{\alpha+1}; \text{ CCDF: } \mathbb{P}(Y > y) = 1 - F_Y(y) = \left(\frac{y_0}{y}\right)^{\alpha} \quad (1)$$

where Y is the continuously distributed random variable denoting incomes, y a realization of it, $y_0 > 0$ some minimum income level from which the distributional form applies, and $\alpha > 0$ the tail parameter which describes the thickness of the tail of the distribution. Lower values of α imply the probability of observing large realizations of income decays to 0 slower, leading to fatter tails.[4] The property of power laws that proves most useful for the purpose of interpolation is that of *scale invariance*. In particular, it is easily shown that the ratio of the average income above a threshold divided by the threshold is given by:

$$\frac{\mathbb{E}(y \mid y > y^*)}{y^*} = \frac{\alpha}{\alpha - 1} = \beta \quad (2)$$

That this ratio does not depend on the threshold itself (hence scale invariant) means once β is identified (from the tax data), average incomes above any threshold $y^* > y_0$ can be interpolated as $\mathbb{E}(y \mid y > y^*) = \beta \times y^*$. In this context, β has a more intuitive interpretation than the tail parameter α and has come to be known as the inverted Pareto coefficient in the literature (Atkinson et al., 2011). Higher values of β imply more concentration of incomes, i.e. fatter tails. Building on this framework, "Standard" Pareto interpolation techniques relied on the strict Paretian assumption that an exact Power Law with a constant tail parameter β holds within each income bracket to interpolate the distribution between two bracket thresholds from tabulated tax data (Pareto, 1896; Kuznets, 1953; Banerjee and Piketty, 2005; Piketty et al., 2017). However, standard interpolation methods do not make use of all the information in the tax data and they do not always work well, even for describing top incomes. Relaxing the strict Paretian assumption that β remains constant across the distribution, instead allowing it to vary by rank, Blanchet et al. (2022b) develop *generalized* Pareto curves. Letting $p \in [0, 1]$ denote rank in the distribution and $Q(p)$ its associated quantile function,

[4] With $\alpha > 2$, both the mean and variance are finite; with $1 < \alpha \leq 2$, the mean is finite but not the variance; with $\alpha \leq 1$, both the mean and variance are infinite.

the generalized Pareto coefficient is given by:

$$\beta(p) = \frac{\mathbb{E}\big(y \mid y > Q(p)\big)}{Q(p)} = \frac{1}{(1-p)Q(p)} \times \int_p^1 Q(s)\, ds \tag{3}$$

This yields generalized Pareto curves (typically U-shaped) summarizing the concentration of income across the distribution. Applying quintic spline interpolation to the income thresholds and averages in the tabulated tax data, generalized Pareto interpolation (GPinter) is able to extract a continuous and smooth Pareto curve which is used to interpolate a full distribution from the tabulated data in a manner that outperforms other interpolation methods (Blanchet et al., 2022b). I use the adult (20+) population to define the fractiles (ranks) corresponding to the respective income thresholds in the tax data based on the number of returns in each bracket. Then using GPinter, I am able to extract a smooth distribution of top incomes from the tax data.

3.2 Bottom Incomes: Household Surveys

Given the limited coverage of tax data, it can be considered reliable only for top incomes, covering at most about 10%-12% of the population in recent years. To estimate incomes for the rest of the population, I must rely on household surveys. Historically, NSSO has steered clear of measuring incomes owing to the difficulty of accurately assessing incomes in the highly seasonal agricultural sector. Instead, they have focused on collecting data on consumption expenditure. While providing a relevant proxy for incomes, consumption expenditure is likely to under-state incomes of the rich, and consequently inequality too, since they consume only a fraction of their incomes. To work around this issue, Chancel and Piketty (2019) relied on the India Human Development Survey (IHDS) which collected data on both incomes and consumption, and estimate consumption-to-income scaling ratios at each percentile of the consumption distribution. Letting y denote income, c consumption, and p percentiles in their distributions, the scaling ratios are defined as $\alpha_p = y_p/c_p$. Estimated from the two rounds of IHDS in 2005 and 2011-12, they applied these scaling ratios to the NSSO consumption expenditure surveys (CES), available at regular intervals between 1951 and 2011, to estimate middle and bottom incomes.

The main empirical challenge with extending their series to recent years is the lack of any CES data post 2011-12. A CES round was conducted in 2017-18 but the microdata and report, which showed mean monthly consumption marginally declined between 2011-12 and 2017-18, were suppressed by the government. Without reliable survey data, it is not possible to shed light on dynamics of income inequality in India. I propose an approach that allows constructing a CES-comparable consumption distribution from PLFS data to which I can apply consumption-to-income scaling ratios, allowing me to replicate the Chancel and Piketty (2019) methodology for estimating survey incomes. Given that the PLFS series is currently being released annually by the government, this could provide one of the basis for continually estimating survey-based incomes going forward.

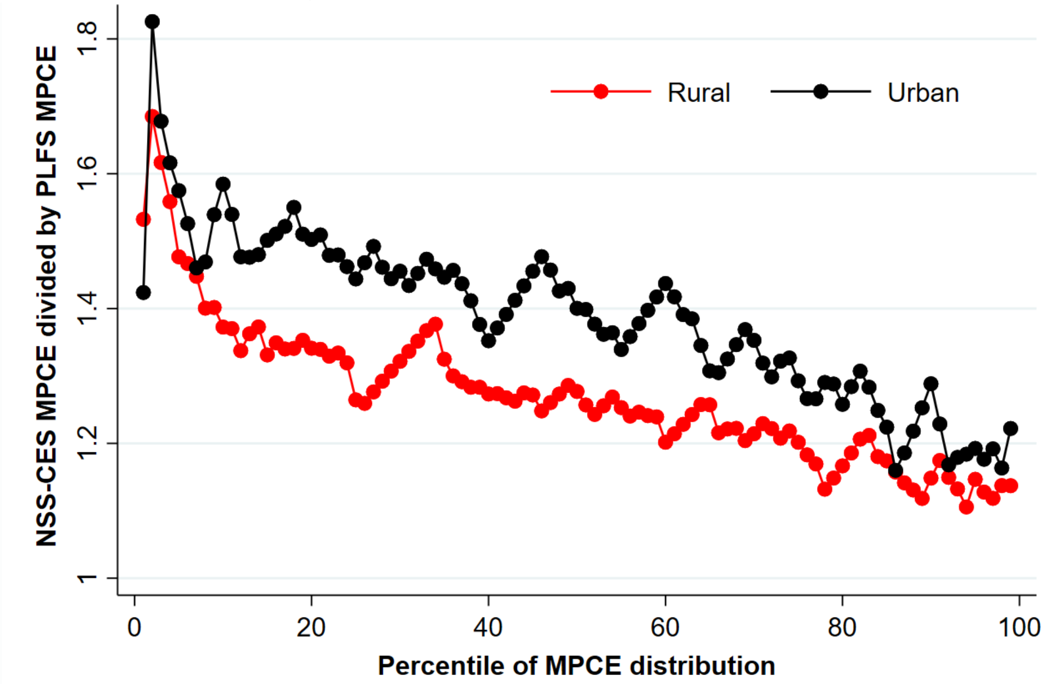
The PLFS series, available annually 2017-18 onwards, collects data on ‘usual’ consumption expenditure of households. However, unlike the traditional approach of CES which collects detailed data on expenditures on individual commodities, this data in PLFS is based on 4 or 5 very broad questions. This difference is likely to render PLFS and CES consumption estimates incomparable (Deaton and Kozel, 2005). The approach I propose here relies on the fact that the suppressed CES 2017-18 round roughly coincided with the PLFS 2017-18 round. Further, given that both surveys were conducted by the NSSO and their sampling designs were nearly identical, I make the relatively mild assumption that the PLFS and CES samples are approximately independent and identically distributed ($i.i.d.$). Differences in reported consumption across the two surveys can then largely be attributed to differences in instrument. By point-wise comparing the two, I could create a mapping from the PLFS distribution to the CES distribution. However, the microdata from the CES 2017-18 round was suppressed by the government on sparsely justified grounds of data quality. Nonetheless, some tabulated data on the consumption distribution from the 2017-18 CES round was made available to the public based on a leaked report in a series of articles (Subramanian, 2019c,b). I use GPinter to extract a full distribution from this tabulated data for rural and urban areas separately. I also compute the consumption distribution for rural and urban areas using the ‘usual’ consumption in PLFS. Subsequently, by point-wise comparing the distribution at each percentile, I can estimate ‘PLFS-to-CES’ scaling ratios. Formally, for p_g denoting percentiles in the consumption distribution and $g \in \{R, U\}$ denoting rural/urban areas, let c_{p_g} denote the average consumption in a percentile. Then PLFS-to-CES scaling ratios (δ_{p_g}) are estimated as follows:

$$\delta_{p_g} = \frac{c_{p_g}^{CES}}{c_{p_g}^{PLFS}} \;\; ; \;\; \forall \; g \, \in \{R, U\} \tag{4}$$

These scaling ratios by percentile are presented in Figure 3. Three key things to note. First, in both rural and urban areas, PLFS underestimates CES-type consumption - CES/PLFS ratios are significantly greater than 1 for all percentiles. Second, the underestimation in PLFS is higher in urban areas throughout the distribution. Third, and perhaps most importantly, in both rural and areas, the underestimation decreases as I move up the distribution. In fact, the underestimation is most severe in bottom decile. Not accounting for these issues with the consumption reported in PLFS would lead to overestimation of inequality.

This scaling, by construction, *exactly* re-creates the CES distribution for the year 2017-18. The

Figure 3: **PLFS-to-CES scaling ratios**



Source: *Authors estimates based on unit-level PLFS data and CES tabulations in Subramanian (2019c,b). A percentile-wise CES distribution was extracted from tabulations using Gpinter.*

advantage of using the scaled PLFS data instead of the generalized CES distribution is that the latter relates to the *per-capita* distribution (including children) whereas I would like to distribute income only among adults. The PLFS microdata allows us to do so without making ad hoc assumptions about adult-child ratios across the consumption distribution. Further, given that the key difference between the PLFS and CES instruments did not change over time, the δ_g scaling factors maybe considered largely time invariant and hence I apply them to subsequent PLFS rounds of 2018-19 and 2019-20 as well.

Once I have CES-type consumption expenditure from the PLFS data, I proceed to apply the consumption-to-income scaling ratios estimated from IHDS data. The IHDS surveys, conducted in two rounds in 2005 and 2011-12, collected data on both consumption and income for a representative sample of Indian households. Chancel and Piketty (2019) use this data to construct consumption-to-income ratios in a few different ways. To begin with, they estimate mean consumption and mean income for each percentile at all-India level and define consumption-to-income ratios as:

$$\alpha_{1p} = \frac{y_p}{c_p} \tag{5}$$

They then define an alternate scenario where the poor are constrained to have non-negative savings such that:

$$\alpha_{2p} = \begin{cases} 1 & \text{if } \alpha_{1p} \leq 1 \\ \alpha_{1p} & \text{otherwise} \end{cases} \tag{6}$$

Finally, they define a third scenario as the average of the first two scenarios:

$$\alpha_{0p} = \frac{(\alpha_{1p} + \alpha_{2p})}{2} \tag{7}$$

These three scaling ratios can be estimated from both rounds of surveys. Given their objective of estimating incomes over the long-run, Chancel and Piketty (2019) use the third scenario α_{0p} averaged over the two IHDS rounds as their benchmark. In contrast, given that my analysis is focused solely on the most recent period, I prefer to use α_{2p} from the more recent IHDS round as my benchmark. I discuss the estimates from alternate scaling ratios in section 5. As it happens, this also delivers the most conservative estimates. Formally, for an individual in PLFS sample with consumption c_i in percentile p_g of the rural/urban consumption distribution and percentile p of the all-India consumption distribution, income is estimated as:

$$y_{i,p_g,p} = c_i \times \delta_{p_g} \times \alpha_p \quad ; \quad \forall \ g \ \in \{\text{rural} \ , \ \text{urban}\} \tag{8}$$

Lastly, PLFS data is only available starting 2017 while the current series ends in 2014. Consequently, for all percentiles $p \in [0, 1]$ in the survey distribution for the years $t \in \{2015, 2016\}$, I interpolate average consumption as $\widehat{y}_{p,t} = y_{p,t-1} \times g_p$ where $g_p = (y_{p,2014}/y_{p,2017})^{1/3}$ is the annual growth rates of (nominal) incomes at each percentile between 2014 and 2017.

3.3 Merging Tax and Survey Distributions

At this point, I have two income distributions - one generalized from the tax data and the other estimated from household surveys. The former is more reliable for top incomes while the latter for bottom incomes. To put both data sources effectively to use, Chancel and Piketty (2019) deem that for $0 < p_1 < p_2 < 1$ representing percentiles in the income distribution, survey data is reliable till p_1 while tax data is reliable from p_2 to 1. Between p_1 and p_2 they try linear, concave, and convex junction profiles. In their benchmark strategy, they fix $p_1 = 0.9$, let p_2 vary accounting for tax-payers and population growth, and use a convex junction profile between the two. This results in a ‘merged distribution’ combining information in survey and tax data.

In recent work, Blanchet et al. (2022a) develop a data-driven methodology for the purpose of merging tax and survey data that frees us from making choices regarding the merge point and junction profiles. The “BFM correction” algorithm proceeds in three steps. It begins by endogenously identifying a "merging point" between the two distribution. Let f_Y and F_Y denote the PDF and CDF of the *true* income distribution, parts of which are observed in tax data, and f_X and F_X the PDF and CDF of the income distribution observed in surveys. Then the relative density in the two distributions at an income level y can be denoted as:

$$\theta(y) = \frac{f_X(y)}{f_Y(y)} \tag{9}$$

If $f_X(y) = f_Y(y)$, then $\theta(y) = 1$ and the survey is unbiased at income level y. Since $f_Y(.)$ is only observed for top incomes and not over the full support of y, the algorithm proceeds by assuming a constant θ below a certain income level such that:

$$\theta(y) = \begin{cases} \overline{\theta} & \text{if } y \leq \overline{y} \\ \dfrac{f_X(y)}{f_Y(y)} & \text{if } y > \overline{y} \end{cases} \tag{10}$$

Where to set $\overline{y}$, the merging point, is of course the key issue. To do so, the authors introduce the cumulative bias function, defined as:

$$\Theta(y) = \frac{F_X(y)}{F_Y(y)} \tag{11}$$

Re-arranging the terms in equation (11) gives the following relationship:

$$\Theta(y)F_Y(y) = \int_{-\infty}^{y} \theta(t)f_y(t)dt \tag{12}$$

For all $y \leq \overline{y}$, I have $\theta(y) = \overline{\theta}$ from equation (10), implying that $\Theta(y) = \theta(y) = \overline{\theta}$. Consequently, the merging point $\overline{y}$ is chosen as the highest income level y such that $\theta(y) = \Theta(y)$.

Assuming that surveys under-represent top incomes in turn also implies assuming that they over-represent bottom incomes. The second step in the BFM correction algorithm involves down-weighting survey observations by a factor $1/\theta(y)$ to account for the over-representation of bottom incomes in surveys. This ensures consistency with population totals at the end of the merging procedure. Lastly, to correct for sampling bias of top incomes in surveys, the algorithm replaces survey observations at the top of the distribution with observations corresponding to the distribution of top incomes extracted from the tax data. This leaves us with a survey dataset that matches the tax data at the top of the distribution with very little sampling variability.

3.4 Tax data for 2018-19 and 2019-20

The Income Tax Department released tax tabulations for the years 1922-23 to 1998-99 but stopped their release subsequently for unexplained reasons. In light of multiple calls for democratic release of data, the IT Dept. began releasing tax tabulations starting 2011-12. However, much to our disappointment and frustration, they have once again abruptly stopped releasing them after 2017-18.[5] On the other hand, in recent years the IT Dept. has made publicly available an online dashboard from its e-filing portal. In principle, the portal provides some relevant information which could stand-in as substitute for the tax tabulations. Unfortunately, for a variety of reasons detailed in the Appendix A.1, the data on the e-filing portal does not seem reliable enough for our purposes. Most importantly, the data has too few tax brackets and only reports total tax returns filed but not the total incomes assessed. This severely limits the extent to which the data can be used. Instead, for the years 2018-19 and 2019-20, I choose to extrapolate forward from the data available for the 7 years 2011-12 to 2017-18. I proceed in the following manner. To estimate total incomes from the tax data, I need to estimate: (i) the total number of tax filers and (ii) average incomes. The data for the years 2011-12 to 2017-18 is available for fairly detailed 25 tax brackets and as shown in Figure 11 in the Appendix, given that the tax brackets in the data are quite small, the average income in all brackets has remained almost constant across years. This means, within each tax bracket, I just need to impute the total number of tax filers to identify the total income in that bracket. I use the mean annual growth rates in the 2011-12 to 2017-18 to do so - details are in Appendix A.2.

3.5 Fiscal Income to National Income

A long-standing and well-acknowledged issue in the Indian context is the sharp divergence of consumption growth rates as observed in NSSO household surveys and national accounts data since the 1980s. Chancel and Piketty (2019) show that some of this gap can be explained by missing top incomes in surveys. However, a non-trivial fraction of the gap still remains unexplained. To facilitate cross-country comparisons and remain consistent with DINA guidelines, they scale-up the merged distribution to make income aggregates match those in national account data. This step is distributionally-neutral as it involves scaling-up all income threshold and averages by a constant factor but leaves income shares unchanged. I follow the same procedure in order to match the total fiscal incomes from the merged survey plus tax distribution and total national income in national accounts. This in turn implies that the growth rate of average incomes in our series matches the growth rates observed in national accounts data.

[5] Once again, the motivation behind the decision is not known, however, this follows a broader trend of suppression of various kinds of data by the Government of India.

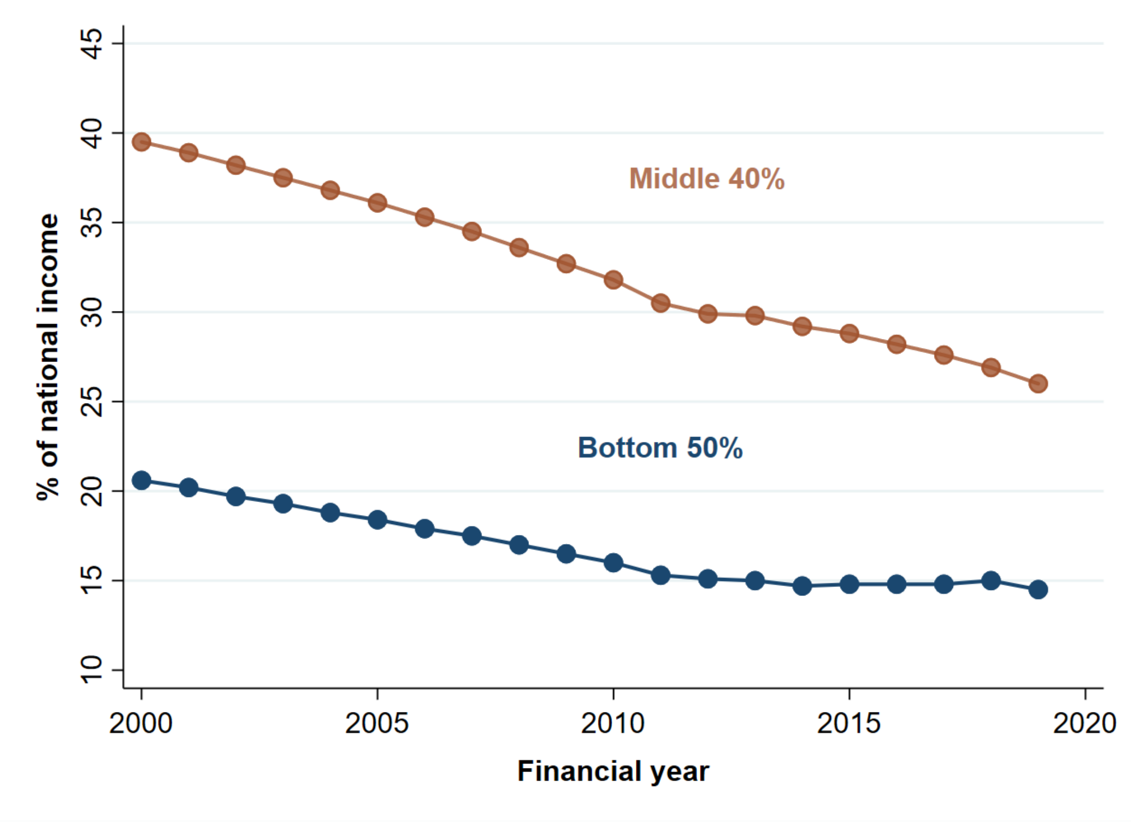
4 Benchmark Results

A few caveats worth noting before I present my results. I focus on the period 2000-01 to 2019-20 for my analysis here. In the appendix, I also present results for the extended period 1980-81 to 2019-20. For the period pre-2014, I rely on the income series from the Chancel and Piketty (2019) series and for post-2014 based on my estimates. Second, for my benchmark series, I assume non-negative savings among the poor when scaling consumption to incomes - this by definition is a conservative choice among the options available. As I discuss in the robustness section, alternate scaling ratios where I allow for negative-savings among the poor lead to lower bottom shares and higher top shares. Lastly, it is worth re-iterating that given that both survey and tax data are patchy during the most recent years as a result of which there naturally are various measurement challenges implying an inherent degree of uncertainty. Consequently, the results must be interpreted with caution and are certainly not nearly a final word on the matter. Instead, we see this as a first-step towards better understanding income dynamics during a period when data sources have become scarce. In subsequent work we intend to incorporate additional data sources that would improve the quality of measurement.

4.1 Bottom 50% and Middle 40% shares

With those caveats in mind Figure 4 presents the share of national income going to the bottom 50% and middle 40%. We see that the shares for both groups have more or less monotonically decreased year-on-year between 2000 and 2020. The share of the bottom 50% fell from just over 20% in 2000 to 15% by 2011 and has remained stagnant there till 2020. These bottom shares are extremely low, perhaps among the lowest in the world (Assouad et al., 2018). This is in line with some of the available evidence that consumption levels as well as wages, especially of the rural poor, have grown very little during the most recent decade (Subramanian, 2019b,c; Das and Usami, 2017; Drèze, 2023). During the same period, the decline in the income shares of the middle 40% was more rapid - it fell by 15 percentage points from 40% in 2000 to 30% by 2011 and to just over 25% by 2020. This would suggest that even the relatively better-off part of the population (percentile 50-90) have lost out on their share of the pie during the last two decades when the economy experienced relatively rapid economic growth. The new estimates for the most recent years (2015-2020) are further evidence that growth of the "middle class" has been rather modest in India compared to observed growth process in other economies (Chancel and Piketty, 2019).

Figure 4: **Bottom 50% and Middle 40% Income Shares**



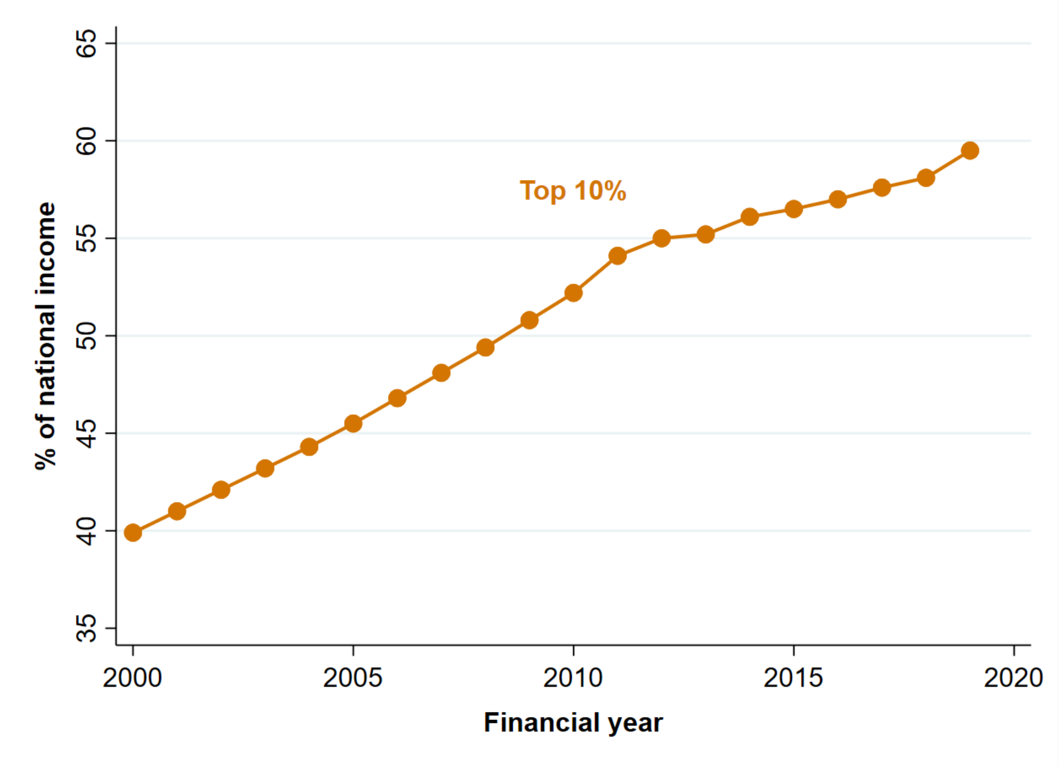
Source: *Author's estimates and Chancel and Piketty (2019).*

4.2 Top 10% shares

The decline in bottom 50% and middle 40% shares in the recent years is compensated by a rise in top 10% which has risen rather sharply over the 20 year period. From 40% in 2000, the top 10% share rose to just under 55% in 2011 and then to nearly 60% by 2020 (Figure 5). For one, these top 10% shares are the highest anywhere in the world. As per recent estimates on the World Inequality Database (WID), the top 10% share was 56% in the Middle East region, 58% in Brazil and 65% in South Africa in recent years. This would imply that India would place behind only South Africa in terms of the levels of income concentration.

Two pieces of evidence lend support to this trend. First, there is evidence of a skill and language premium in India with science education and English language fluency being associated with as large as 22% higher earnings on the labour market (Jain et al., 2022). This is further backed by reports suggesting strong income growth for those employed in skill-intensive employments such as those in the information technology (IT) sector (Sangani, 2021). Second, the labour share in the formal manufacturing sector has decline quite considerably in recent decades and settled at very low-levels - just over 10% (Abraham and Sasikumar, 2017; Jayadev and Narayan, 2018). Third there has been a rise in concentration of wealth inequality - the share of top 10% and top 1% in the total wealth has risen between 2000 and 2018 as evidenced from NSSO wealth surveys as well as information from rich lists like those published by Forbes (Bharti, 2018; Anand and Kumar, 2022).

Figure 5: **Top 10% Income Shares**

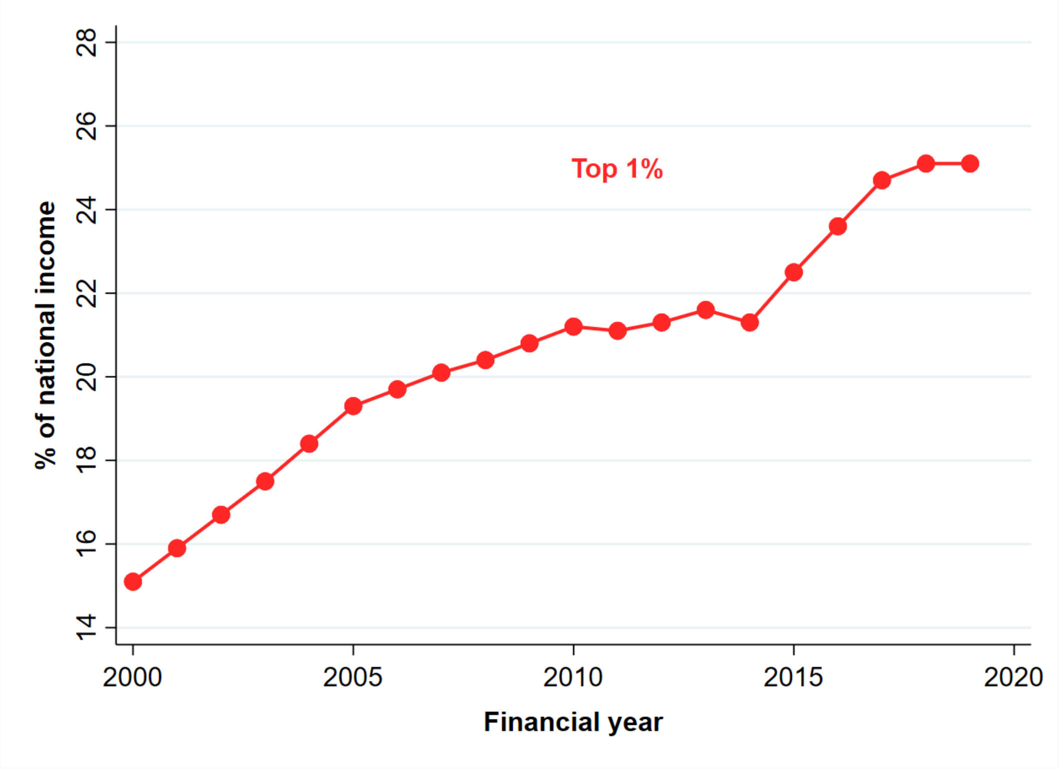


Source: *Author's estimates and Chancel and Piketty (2019).*

4.3 Top 1% shares

Figure 6 presents the top 1% income shares. Much like top 10% shares, top 1% shares have risen quite considerably in the last two decades. Starting at about 15% in 2000, top 1% shares rise to about 22% by 2014 and then further increased to 25% by 2020. The rise is relatively more pronounced in the post-2014 period. This would be in line with evidence of increasing concentration of wealth in recent years as mentioned above. This would also be in line with assessments shedding light on the pro-big business and government nexus (Banaji, 2022) that has developed in recent years leading some observers to characterize the situation under the new BJP government as that of conglomerate capitalism (Damodaran, 2020) and a conclave economy (Bardhan, 2022). On the other hand, given the fragility of the tax data, all the more so in recent years, there is an added degree of uncertainty around these results. Nonetheless, these provide a useful benchmark till better data becomes available to track the right-tail of the income and wealth distributions. I present estimates for recent years from alternate estimation strategies in section 5.

Figure 6: **Top 1% Income Shares**

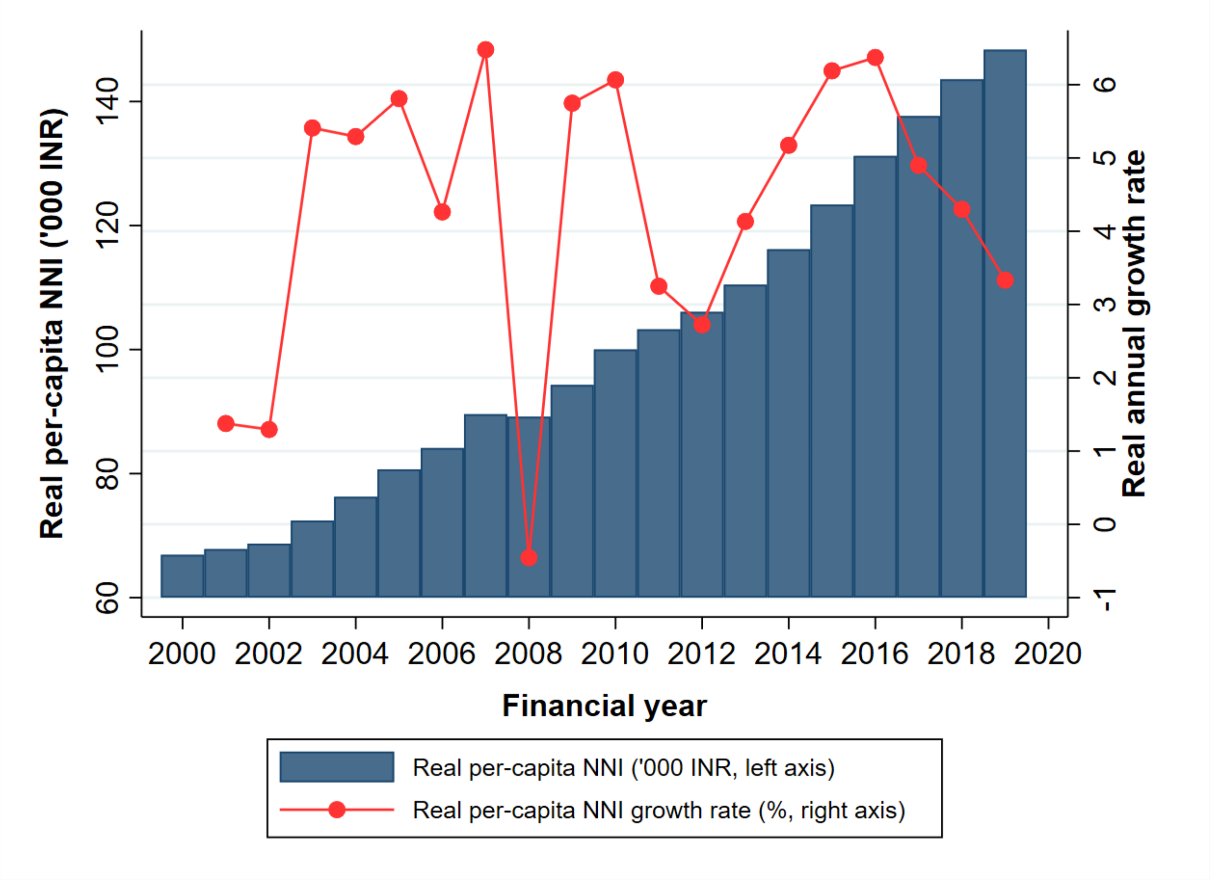


Source: *Author's estimates and Chancel and Piketty (2019).*

4.4 Growth rates, 2000-2020

With a set of estimates in place for the most recent period (2015-2020), we are now in place to better understand growth dynamics during the period 2000-2020. Figure 7 presents trends of average (per-capita) net national income (NNI) along with its respective growth rates based on official statistics released in the Economic Survey 2022-23 released by the Ministry of Finance. Average incomes more than doubled in real-terms, from INR 67,000 to INR 148,000, between 2000 and 2014. Per-capita growth rates in particular picked up post-2003 to between 5%-6% annually and then dipped briefly post-2011. Subsequently, growth rates picked up between 2012 and 2015 only to show a decline once again post-2016. Before moving on to look the distribution of growth across the distribution during this period, a word of caution is due relating to official national account statistics in recent years. The government introduced a new GDP series in 2015 replacing the earlier base year of 2004-5 with an updated base year for 2011-12. While part of relatively usual updates to the series, this particular update was peculiar in that it resulted in a marginal shrinkage of the absolute GDP in 2011-12 as well as produced systematically higher growth in subsequent years (Nagaraj, 2020b). Moreover, economists closely following India's GDP numbers, including an ex Chief Economic Advisor to the Government of India, have argued that the new series likely overestimates GDP levels and trends for a myriad of reasons (Subramanian, 2019a; Morris and Kumari, 2019).

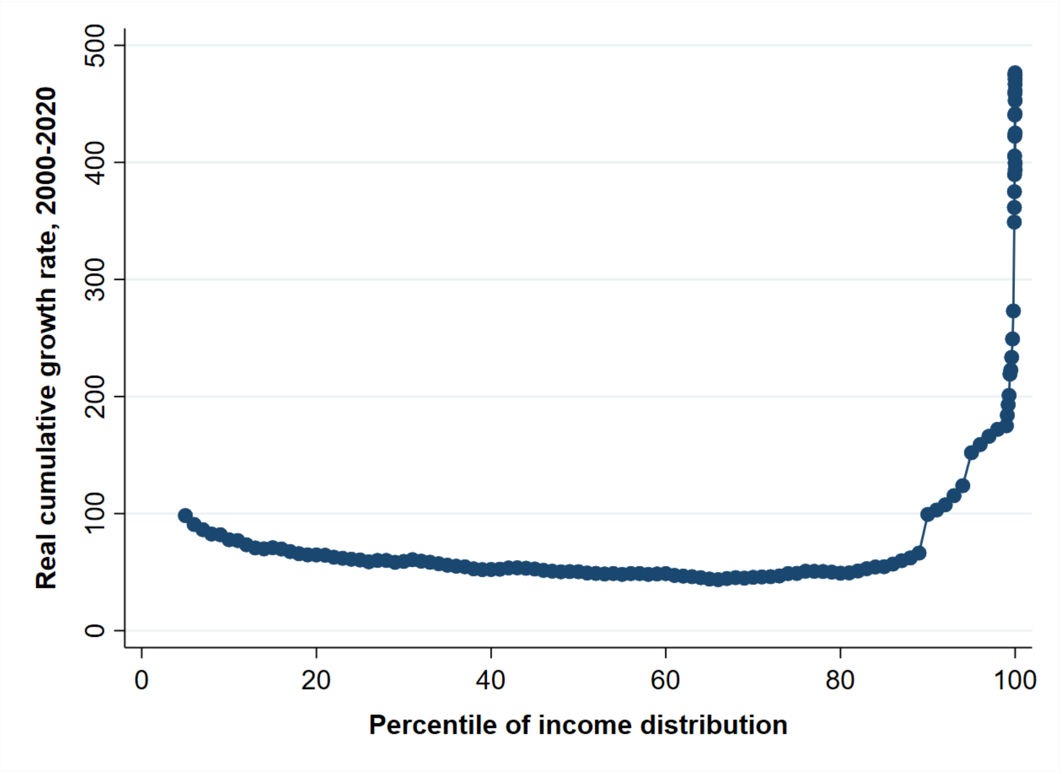
Figure 7: **Growth of average incomes, 2000 to 2020**



Source: *Author's estimates based on Economic Survey 2023, Statistical Appendix for net national income (NNI) and adult-population based on UN World Population Prospects.*

We now look at the distributional dynamics of growth. Figure 8 presents the cumulative growth rate of incomes between 2000 and 2020 along the full distribution. As a benchmark, average incomes grew by 121% during this period. On the other hand, for a majority of the population, right up until the top decile, growth rates were below-average. At the very top, however, growth rates were between 400%-500% within the top 1%. This suggests that growth was favourable to very rich in the top 10% and in top 1%.

Figure 8: **Inflation-adjusted cumulative growth rates of incomes, 2000 and 2020**



Source: *Author's estimates combining the income distribution in 2000-01 based onChancel and Piketty (2019) and 2019 based on authro's estimates. Nominal values adjusted for inflation using the GDP deflator.*

To get a better understanding of the distributional dynamics, it is useful to look at the how the total growth in aggregate net national income in real-terms between 2000 and 2020 was captured by different income groups. This along with the income shares in 2000 and 2020 is presented in Table 2. We see that the bottom 50% was able to secure only 12% of the total growth in this period, while the middle 40% was able to capture 20%. This implies that a majority of India's population - 90% of it more precisely - captured less than a third of the total growth during the last two decades. In contrast, the top 10% was able to amass 68% of the growth with the top 1% itself securing nearly a third of the growth. Which is to say that the top 1% captured as much of the real income growth during 2000 and 2020 as the entire bottom 90% did. While there is some tentativeness in the results given the fragility of the tax data, the evidence suggests that the "Billionaire Raj" that Chancel and Piketty (2019) document has remained un-challenged in recent years and perhaps even intensified.

Table 2: **Share of real growth distributed to percentile groups, 2000-2020**

Income group (percentile)	**Income shares (%)**		**Share of real growth**
	2000-01	**2019-20**	**captured b/w 2000-20 (%)**
Bottom 50%	20.6	14.5	11.9
Middle 40%	39.5	26.0	20.3
Top 10%	39.9	59.5	67.8
incl. Top 1%	15.1	25.1	29.3
incl. Top 0.1%	4.9	9.7	11.7
incl. Top 0.01%	2.3	3.8	4.4

Note: *The second and third columns present the share of total net national income going to different parts of the income distribution in 2000-01 (Chancel and Piketty, 2019) and 2019-20 (author's estimates). The last column presents the share of aggregate real growth captured by the various income groups. Nominal values adjusted for inflation using the GDP deflator.*

5 Robustness: Alternate survey-based approaches

Traditionally NSSO household surveys in India have shied away from measuring incomes, largely owing to the difficulties in assessing incomes in the agricultural sector, instead choosing to measure consumption expenditure as a proxy. This creates the need for approximating an (unobserved) income distribution from the consumption distribution in CES surveys using consumption-to-income scaling ratios. Moreover, in recent years, I additionally also need to scale the PLFS consumption distribution to make it CES-comparable. How robust are our results to the assumptions that these scaling ratios embody? In recent years, two all-India surveys collected data on incomes. Using these surveys, I develop two alternate approaches for estimating bottom incomes which do not rely on scaling consumption.

The first of these surveys is PLFS available from 2017-18 onwards which is an all-India representative survey of over 130,000 households (400,000 individuals) which collects data on wages of casual labourers and regular salaried workers and incomes of the self-employment. However, it does not record non-labour incomes (rents, interest, dividends, etc.). The second survey, CPHS available from 2014-15, is also a large all-India survey covering over 450,000 households which, unlike PLFS, records both labour and non-labour incomes. However, the CPHS sample is not all-India representative as claimed by the private company Centre for Monitoring Indian Economy (CMIE) that executes the survey. Comparing CPHS to various other credible sources on a range socio-economic variables suggests CPHS under-represents women and young children, over represents well educated households, and under-represents the poor (Drèze and Somanchi 2021; Somanchi 2021). Thus, while both surveys contain relevant data on incomes, neither is independently sufficient to reliably estimate the income distribution. Nonetheless, it is possible to combine the information in the two surveys in a meaningful manner.

5.1 Statistical matching

The empirical challenge with solely using the PLFS data is that there is a "missing data" problem in that non-labour incomes are unobserved. On the other hand, CPHS records both labour and non-labour incomes but is a non-representative sample. The first approach I propose is to treat PLFS as the "base" and use statistical matching techniques to match every household in the PLFS sample to the most similar household in CPHS sample. This allows us to retain the representative sample of PLFS and simply import non-labour incomes from CPHS. I match households on a set of observable covariates including, most importantly, labour incomes which are observed in both surveys. This approach is much like how matching is used in the impact evaluation literature; it relies on the "selection on observables" assumption, except selection here being into a survey instead of a treatment.

Coarsened exact matching (CEM): I begin with a simple procedure with only 3 matching co-variates: principal employment group (x_1), rural-urban (x_2), and total labour incomes (x_3). The first two are discrete but the last is a continuous variable. I "coarsen" labour incomes into bins of Rs. 100 (denoted x_{3b}) and then classify all households in PLFS and CPHS into unique $x_1 \times x_2 \times x_{3b}$ cells. For every household h in the PLFS sample, I find all households j in the CPHS sample that match exactly on $x_1 \times x_2 \times x_{3b}$. In most cases, I find multiple matches. So for household h in the PLFS sample I impute non-labour incomes (nl) as the un-weighted average of non-labour incomes of all matched CPHS households:

$$\widehat{nl}^{PLFS}_{h,x_1,x_2,x_{3b}} = \frac{1}{k}\sum_{j=1}^{k} nl^{CPHS}_{j,x_1,x_2,x_{3b}} \tag{13}$$

Multivariate distance matching (MDM): While CEM has the advantage of being simple and transparent, there are some potential drawbacks. First, I would like to incorporate more co-variates to match on. Second, since I corasen labour incomes, all PLFS households in the same $x_1 \times x_2 \times x_{3b}$ cell are assigned the same non-labour incomes, implying there is possibility that I artificially compress the variance of the income distribution. To account for these concerns, I adopt an alternate matching method to match each PLFS household to a single CPHS household. Let $\boldsymbol{x} = (x_1\, x_2\, \ldots\, x_{10})^T$ denote the expanded co-variates vector with the following variables: (i) state (ii) rural-urban (iii) month of survey[6] (iv) household size (v) principal employment group (vi) religion (vii) social group (viii) average age of household members (ix) average years of schooling of household members and (x) total labour income of household. For each PLFS household h, I find the CPHS household j that minimizes the multivariate Mahalanobis distance metric:

$$\min_{\{j\}} \left(\boldsymbol{x}_h - \boldsymbol{x}_j\right)^T \widehat{\Sigma}^{-1} \left(\boldsymbol{x}_h - \boldsymbol{x}_j\right) \;\; \forall\, h \tag{14}$$

where $\boldsymbol{x}_h$ is the vector of matching covariates for household h in the PLFS sample, $\boldsymbol{x}_j$ the covariates of a potential match household j in the CPHS sample and $\widehat{\Sigma}$ an estimate of the sample covariance of the covariates. In some cases, two or more CPHS household minimize the distance for a given PLFS household - I randomly pick one CPHS household as a match in that case.[7]

5.2 Weight calibration

The CPHS dataset includes data on both labour and non-labour incomes, freeing us from having to either impute it from consumption or wages. Moreover, the CPHS dataset is available from 2014 onwards while the PLFS series only begins in 2017-18. So I would like to use the CPHS data but the empirical challenge is that I know that the data does not come from a representative sample. Consequently, in the second approach, I treat CPHS as the base and use credible auxilliary data from PLFS and other surveys to correct some of the observable biases. This I do by re-weighting the CPHS dataset. Then, using the adjusted weights, I can directly estimate the distribution of total incomes (labour + non-labour).

The basic idea is the following. Let $X = \{x^1, x^2, \ldots, x^M\}$ be a set of control (auxiliary) variables with $\overline{X} = \{\overline{x}^1, \overline{x}^2 \ldots, \overline{x}^M\}$ the set of their a-priori known means, p_j the prior (original) design weight of CPHS households and q_j the calibrated (adjusted) weight I want to generate, with $j = 1, 2, \ldots, N$ being the observations in the CPHS sample. Without loss of generality, let p_j and q_j denote *normalized* weights such that $\sum_{j=1}^{N} p_j = \sum_{j=1}^{N} q_j = 1$. Finally, let $D(\boldsymbol{p}, \boldsymbol{q})$ be some distance function between the two vectors of weights, satisfying most importantly: (i) $D(\boldsymbol{p}, \boldsymbol{q})$ is strictly convex (ii) $D(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ (iii) $D(\boldsymbol{p}, \boldsymbol{p}) = 0$. Imposing a few more restrictions on $D(\boldsymbol{p}, \boldsymbol{q})$, Deville and Särndal (1992) showed that the re-weighting problem at hand

[6] Since incomes seasonally fluctuate over a financial year, controlling for the month of survey ensures that the labour incomes that I match households on correspond to roughly the same time frame.

[7] With both CEM and MDM, a small fraction ($\sim$ 5%) of households do not get matched. In the case of CEM, I impute non-labour incomes from the nearest cell and in the case of MDM, I resort to CEM to match the un-matched households.

can be re-stated as the following constrained optimization problem:

$$\begin{cases} \min\limits_{\{\boldsymbol{q}\}} \;\; D(\boldsymbol{p}, \boldsymbol{q}) & \textit{subject to} \\ \\ \;\; q_j \geq 0 \qquad\qquad\quad \forall\, j & \text{[Non-negative weights]} \\ \sum_{j=1}^{N} q_j x_j^m = \overline{x}^m \quad \forall\, m & \text{[Control-variable means]} \\ \sum_{j=1}^{N} q_j = 1 & \text{[Normalized population]} \end{cases} \tag{15}$$

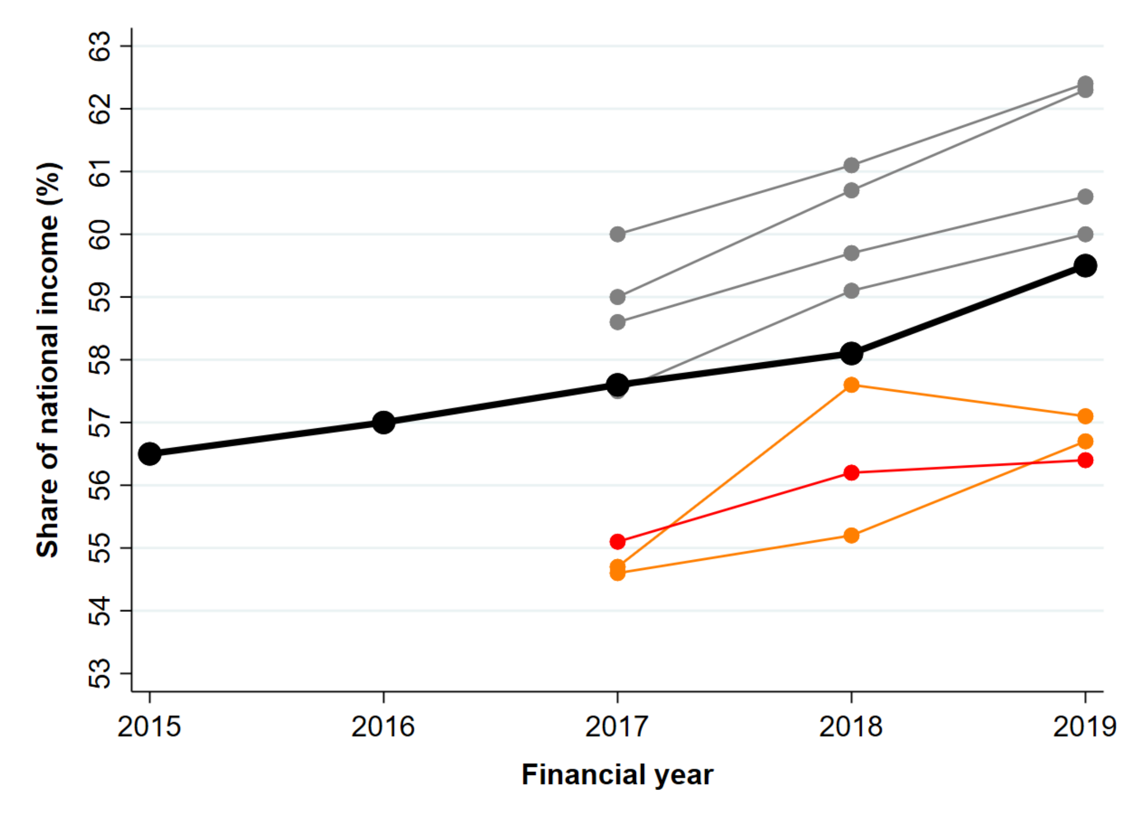
With respect to choice of control variable means, I use the same set of co-variates as those listed above for the multi-variate distance matching. Much like statistical matching, this approach too relies on a selection on observables type of assumption but applied to a non-response model for CPHS observations. This means that the control variables in X must not only be correlated with incomes but also with the sampling bias in CPHS. A key variable likely to be correlated with both incomes and the sampling bias - *distance of residence from main street* - is not included since it is not observed in either survey. Moreover, the adjusted weights fix the marginal distributions of each of the individual control variables $(x_1, \ldots, x_{10})$ but not their joint distribution. On both counts, re-weighting is likely to only deliver an approximate (partial) correction to the bias in CPHS. Nonetheless, the method provides a useful alternative benchmark. Moreover, in addition to including *mean* labour income, I also include the *deciles* of labour incomes in order to correct not just the mean but also provide for a correction across the entire distribution of labour incomes (not just at the mean).

5.3 Robustness Results

As noted above, my benchmark series relies on converting consumption reported in PLFS to NSS-comparable consumption and then applying a scaling-ratio to move from consumption to incomes. To see how reliable this approach is, I make two sets of comparisons. First, I compare my benchmark estimates with those from matching and re-weighting based strategies which do not rely on consumption-based scaling. Second, I also present some results regarding the robustness of my benchmark strategy to alternate scaling factors given that the two rounds of IHDS data allow for construction of a few different ratios (see section 3.2). Lastly, given that both CPHS and PLFS are simultaneously required to implement the alternate estimation strategies, I am restricted to focus on the three years 2017-2019 for which both survey sources are available.

Figure 9 presents estimates of top 10% shares from 8 different approaches. The estimates from the two matching methods (CEM and MDM) are presented in orange while the estimates from the re-weighting procedure are in red. Not only are the estimates from the two alternate strategies very similar to each other but they are also quite close to my benchmark estimates (in black), albeit only by a small difference. This is rather re-assuring and suggests that while not perfect, the approach of scaling consumption to derive an income distribution might be a reasonable approximation in the absence of the other data. Next, the various estimates in gray correspond to my benchmark strategy but using alternate scaling ratios. Two things to note. Firstly, the trend looks similar across the various estimates. Second, the benchmark is the most conservative in the sense that it results in relatively lower shares than the other alternate scaling ratios (among the various gray lines). Overall, these comparisons suggest that while there is ofcourse some degree of uncertainty, the survey-based measurement of incomes is largely robust to alternate approaches.

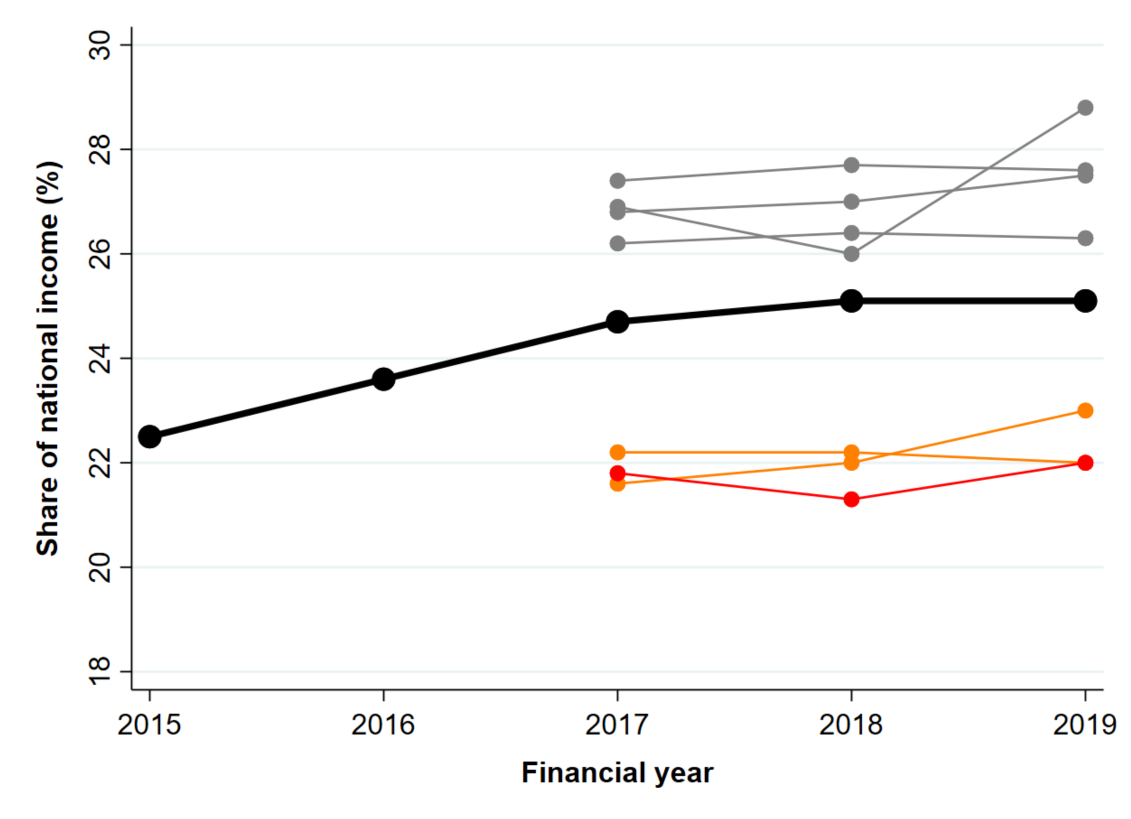
Figure 9: **Robustness: Top 10% shares from 8 alternate strategies**



Note: *Author's estimates using 8 alternate estimation strategies. The black line corresponds to the benchmark estimates, the orange lines are estimates from two different matching methods (CEM and MDM) and the red-line are estimates from re-weighting CPHS. The gray lines correspond to the benchmark strategy but using alternate consumption-to-income scaling ratios.*

Figure 10 presents a similar robustness exercise but for the top 1% shares using estimates from 8 different approaches. The overall picture is quite similar to that for the top 10% shares. The matching and re-weighting estimates (orange and red respectively) deliver slightly lower top 1% shares in the range of 22%-23% in 2019-20 compared to 25% as per the benchmark. On the other hand, alternate income-to-consumption scaling ratios (gray lines) deliver even higher top 1% shares. Therefore, I consider these estimates as reflecting a relatively moderate scenario. Nonetheless, given that all these scenarios rely on the relatively tax tabulations, a certain degree of uncertainty inherently remains till better data for tracking top incomes becomes available.

Figure 10: **Robustness: Top 1% shares from 8 alternate strategies**



Note: *Author's estimates using 8 alternate estimation strategies. The black line corresponds to the benchmark estimates, the orange lines are estimates from two different matching methods (CEM and MDM) and the red-line are estimates from re-weighting CPHS. The gray lines correspond to the benchmark strategy but using alternate consumption-to-income scaling ratios.*

Figures 17 and 18 present similar robustness results for the bottom 50% and middle 40%. The results are qualitatively the same.

6 Discussion

Tracking the dynamics of income inequality in India in recent years has proved challenging owing to suppression of various data sources by the Government of India as well as issues relating to coverage and quality with those that are available. In this paper, I combine available tax tabulations, new survey sources and national accounts data to extend the income inequality series in Chancel and Piketty (2019) to the most recent years (2015-2019) when data challenges are most severe. I then use these estimates to study the distributional dynamics of growth between 2000 and 2020. I find that between 2000-01 and 2019-20, income share of the top 10% grew by 50% from 40% to nearly 60%. Even more so, the income shares of the top 1% grew by two-thirds from 15.1% to 25.1%. As a consequence, the top 10% captured as much as 67.8% of the aggregate national income growth (real-terms) between 2000 and 2020 with the top 1% alone capturing nearly as much the entire bottom 90% did. This extreme levels of concentration of income and growth would put India as one of the most unequal countries in the world (Assouad et al., 2018). These results also suggest that economic policies in recent years has done little to reverse the highly unequal distribution of incomes in India.

Nonetheless, owing to the various measurement challenges, the estimates presented here are far from perfect at this stage. This is primarily because availability of *both* surveys and tax data has been patchy in recent years severely hindering the ability to study incomes. The absence of tax tabulations beyond 2017-18 and the reliance on projections add an added degree of uncertainity. Therefore my current results are best seen as a first-step towards a more clearer understanding of income dynamics in the recent years. In future updates I intend to incorporate additional data sources which are likely to improve the quality of measurement. To begin with, a new survey source has become available. The People Research for India's Consumer Economy (PRICE) has conducted a national household survey that collected detailed information on both capital and labour incomes. This would not only allow to improve the survey-based measurement of incomes, in particular capital incomes which are currently imputed from CPHS data, but also to extend the current series beyond 2019. Moreover, as recent work by Singh (2022) suggests, the wealthy in India are able to hide a non-trivial share of their taxable incomes implying that inequality estimates based on income tax tabulations are likely to understate true inequality. To account for these issues as well as the absence of tax data post-2017, I intend to make use of alternate source of wealth dynamics. For instance, data from the Forbes rich list suggests that the net worth of the 100 richest Indians grew by 356% in real terms between 2014 and 2022. Additionally, the National Housing Bank has been publishing city-wise real estate prices which could be used to shed light on a part of capital incomes. It would appear that incorporating these various new sources is likely to result in even higher inequality estimates than currently estimated.

References

Abraham, V. and Sasikumar, S. K. (2017). "Declining wage share in India's organized manufacturing sector: Trends, patterns and determinants". ILO Asia-Pacific Working Paper Series.

Anand, I. and Kumar, R. (2022). The sky and the stratosphere: Concentrated wealth in india during the last (lost) decade. SSRN.

Anand, I. and Thampi, A. (2021). "The Crisis of Extreme Inequality in India". *Indian Journal of Labour Economics*, 64:663–683.

Assouad, L., Chancel, L., and Morgan, M. (2018). Extreme inequality: Evidence from brazil, india, the middle east, and south africa. *AEA Papers and Proceedings*, 108:119–23.

Atkinson, A. B. (2007). "Measuring Top Incomes: Methodological Issues". In: Atkinson, A. B. and Piketty. T. (eds.), "Top Incomes over the Twentieth Century: A Contrast Between European and English-Speaking Countries", Oxford: OUP.

Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature*, 49(1):3–71.

Banaji, J. (2022). Indian big business. *Phenomenal World*, December 20.

Banerjee, A. and Piketty, T. (2005). Top indian incomes, 1922-2000. *World Bank Economic Review*, 19(1):1–19.

Bardhan, P. (2022). The 'new' india. *New Left Review*, 136, July/August.

Bharti, N. (2018). Wealth inequality in india 1961-2012. WID.World Working Paper N 2018/14.

Blanchet, T., Flores, I., and Morgan, M. (2022a). "The Weight of the Rich: Improving surveys using tax data". *The Journal of Economic Inequality*, 20(1):119–150.

Blanchet, T., Fournier, J., and Piketty, T. (2022b). "Generalized Pareto Curves: Theory and Applications". *Review of Income and Wealth*, 68(1):263–288.

Bourguignon, F. (2018). Simple adjustments of observed distributions for missing income and missing people. *The Journal of Economic Inequality*, 16(2):171–188.

Chancel, L. and Piketty, T. (2019). "Indian Income Inequality, 1922-2014: From British Raj to Billionaire Raj". *Review of Income and Wealth*, 65:S33–S62.

Chodorow-Reich, G., Gopinath, G., Mishra, P., and Narayanan, A. (2019). Cash and the Economy: Evidence from India's Demonetization*. *The Quarterly Journal of Economics*, 135(1):57–103.

Damodaran, H. (2020). From 'entrepreneurial' to 'conglomerate' capitalism. *Seminar*, 734, October.

Das, A. and Usami, Y. (2017). Wage rates in rural india, 1998–99 to 2016–17. *Review of Agrarian Studies*, 7(2).

Deaton, A. and Dreze, J. (2002). Poverty and inequality in india: A re-examination. *Economic and Political Weekly*, 37(36):3729–3748.

Deaton, A. and Kozel, V. (2005). "Data and Dogma: The Great Indian Poverty Debate". *The World Bank Research Observer*, 20(1):177–199.

Deville and Särndal, C. (1992). "Calibration Estimators in Survey Sampling". *Journal of the American Statistical Association*, 87(418):376–382.

Drèze, J. (2023). Distress is there to see. *Indian Express*, 23 May.

Drèze, J. P. and Somanchi, A. (2021). "Not Having the Have Nots ". *Economic Times*, 21 June.

Garbinti, B., Goupille-Lebret, J., and Piketty, T. (2018). Income inequality in france, 1900–2014: Evidence from distributional national accounts (dina). *Journal of Public Economics*, 162:63–77. In Honor of Sir Tony Atkinson (1944-2017).

Ghatak, M. and Mukherjee, U. (2019). The mirage of modinomics. *The India Forum*, February 22.

Ghatak, M., Raghavan, R., and Xu, L. (2022). "Trends in Economic Inequality in India". *India Forum*.

Gupta, A., Malani, A., and Woda, B. (2022). "Inequality in India Declined During Covid". *India Forum*.

Jain, T., Mukhopadhyay, A., Prakash, N., and Rakesh, R. (2022). Science education and labor market outcomes in a high stem economy. *Economic Inquiry*, 60(2).

Jayadev, A. and Narayan, A. (2018). The evolution of india's industrial labour share and its correlates. CSE Working Paper 2018-4, Azim Premji University.

Kapoor, A. and Duggal, J. (2022). "The State of Inequality in India Report". Institute for Competitiveness (Commissioned by Economic Advisory Council to Prime Minister of India).

Khera, R. and Yadav, M. (2020). “What pay ratios in NIFTY50 companies tell us about income inequality in India”. *Ideas for India*.

Korinek, A., Mistiaen, J., and Ravallion, M. (2006). Survey nonresponse and the distribution of income. *The Journal of Economic Inequality*, 4(1):33–55.

Kuznets, S. (1953). Shares of upper income groups in income and savings. National Bureau of Economic Research, Cambridge MA.

Morris, S. and Kumari, T. (2019). Overestimation in the growth rates of national income in recent years? – an analyses based on extending gdp04-05 through other indicators of output. IIMA Working Paper No. 2019-01-01, Indian Institute of Management, Ahmedabad.

Nagaraj, R. (2020a). Understanding india’s economic slowdown. *The India Forum*, January 20.

Nagaraj, R. (2020b). Understanding india’s economic slowdown: Need for concerted action. *The India Forum*, January 20.

Pai, S. and Vats, A. (2023). “Indus Valley Report 2023”. Blume Ventures.

Pareto, V. (1896). Cours d’économie politique.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press.

Piketty, T., Saez, E., and Zucman, G. (2017). Distributional National Accounts: Methods and Estimates for the United States*. *The Quarterly Journal of Economics*, 133(2):553–609.

Press Information Bureau (2020). Household consumer expenditure survey. Released by Ministry of Statistics Programme Implementation, Government of India. Dated: 15 November 2020.

Sangani, P. (2021). Tech firms may hike pay by 120% to hire, retain niche talent. *Economic Times*, December 30.

Singh, R. (2022). “Do the Wealthy Underreport their income? Analysing Relationship between Wealth and Reported Income in India”. CDE-DSE Working Paper No. 331.

Somanchi, A. (2021). “Missing the Poor, Big Time: A Critical Assessment of the Consumer Pyramids Household Survey”. SocArxiv Working Paper, August 11.

Subramanian, A. (2019a). India’s gdp mis-estimation: Likelihood, magnitudes, mechanisms, and implications. CID Faculty Working Paper No. 354, Harvard Kennedy School.

Subramanian, S. (2019b). "Letting the Data Speak: Consumption Spending, Rural Distress, Urban Slow-Down, and Overall Stagnation". *The Hindu Centre for Politics and Public Policy*, 11 December.

Subramanian, S. (2019c). "What is happening to Rural Welfare, Poverty, and Inequality in India?". *The India Forum*, 29 November.

Subramanian, S., Anand, A., and V., D. (2020). New welfarism of india's right. *Indian Express*, December 22.

Verghese, A. (2021). Behind the gates. *The Hindu*, December 24.

World Inequality Lab (2021). "Distributional National Accounts Guidelines: Methods and Concepts Used by the World Inequality Database".

Appendix

A.1. Inadequacy of data on new e-filing portal

As discussed in the main text, the Government of India abruptly stopped releasing tax tabulations after FY 2017-18. Instead, it began putting out some numbers on an online dashboard on its new e-filing portal.[8] This includes figures for tax filers by a few income brackets. In principle, this data could be used for our purposes. However, there are numerous concerns regarding the consistency of the data as well as its comparability with the tax tabulations used for the earlier period. I highlight the main concerns here.

First, the estimates for total filers on the e-filing portal do not match the totals in the ITRS tabulations for the years for which data from both sources are available (2012-2017). Second, it appears that this difference may at least partly be accounted by the fact that the e-filing portal only reflects data to online tax-filers. Third, the data on the e-filing portal is aggregated by filing data but it is unclear if the data on the e-filing portal is segregated by the true assessment years of the form. Fourth, the e-filing portal is supposed to be 'real-time' tracking of tax filings with monthly updates. But strangely, the data on the portal refreshes in April every year (corresponding to the start of a new financial year) even though tax filings deadlines have often been as late as July. Fifth, compared to the ITRS which provided break-up of tax-filers for 25 fairly granular brackets, the e-filing portal provides it only for 6 very aggregated brackets. Sixth, while the ITRS tabulations provided information on both total filers and total incomes assessed in each of the brackets, the e-filing portal does not provide information on total incomes assessed. This makes extracting an income distribution from the tables even harder.

For these reasons, I deem the data from the new e-filing portal inadequate for my purposes at this stage. Instead, I prefer to project forward data from the tax tabulations for the years 2018 and 2019.

[8] Available here - `https://www.incometax.gov.in/iec/foportal/statistics-data`.

A.2. Projection of tax data for 2018 and 2019

To project tax data beyond 2017, I proceed in the following manner. Letting $b \in \{1, 2, \ldots, 25\}$ denote the 25 income brackets for which the tax data is reported and $R_{b,t}$ the total returns in the bracket b in year t, the mean annual growth rate of total returns for each income bracket is given by:

$$\overline{g}_b = \frac{1}{6} \sum_{t=2012}^{2017} \left(\frac{R_{b,t} - R_{b,t-1}}{R_{b,t-1}} \right) \quad ; \ \forall \, b \tag{16}$$

Note, the $R_{b,t}$ used for estimation here are the total returns in each bracket after accounting for non-filers. Hence, the estimated growth rates $\overline{g}_b$ implicitly account for the growth in both filers and non-filers. The total returns in each bracket for the years 2018-19 and 2019-20 are then imputed in a two-step procedure as follows:

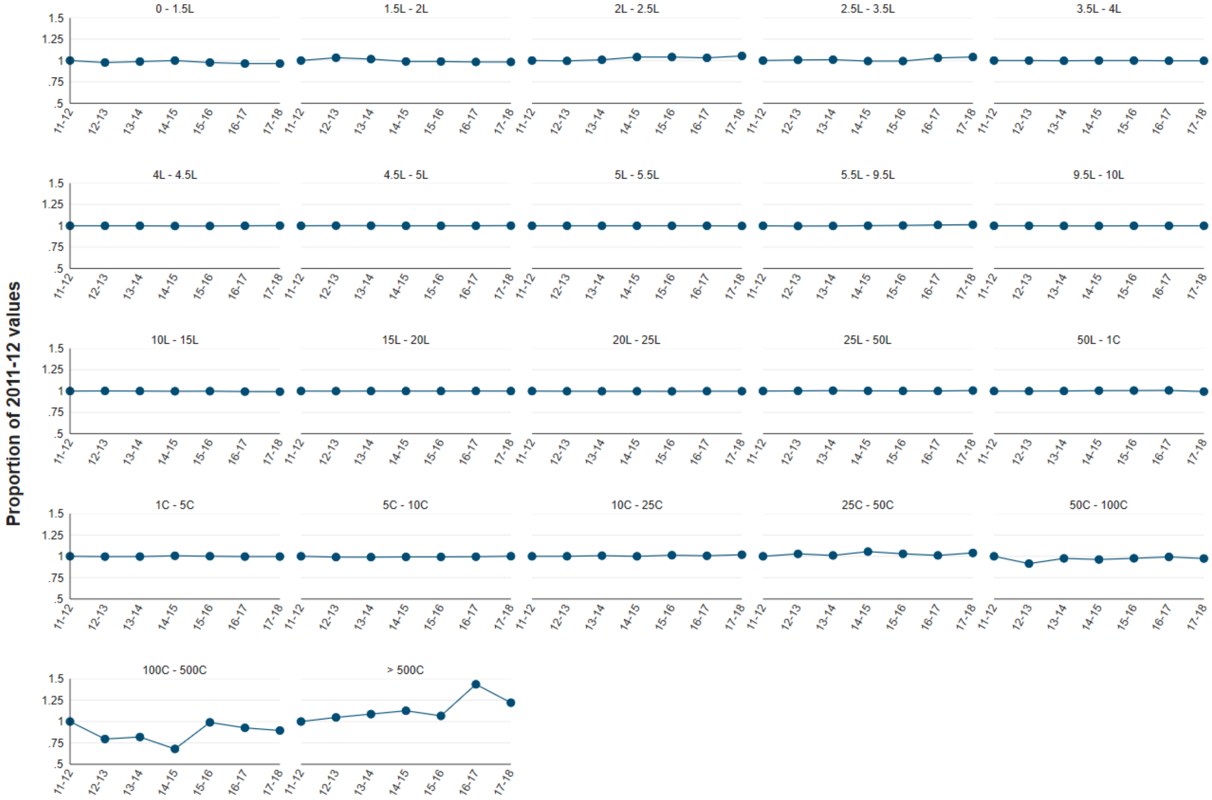
$$\begin{aligned} \widehat{R}_{b,2018} &= (1 + \overline{g}_b) \, R_{b,2017} \quad ; \ \forall \, b \\ \widehat{R}_{b,2019} &= (1 + \overline{g}_b) \, \widehat{R}_{b,2018} \quad ; \ \forall \, b \end{aligned} \tag{17}$$

As shown in Figure 11, average incomes reported in each tax bracket remained essentially constant over the period 2011-12 to 2017-18. I simply take a mean over the period to impute average incomes ($\widehat{\overline{y}}_b$) for each bracket for 2018 and 2019. With estimates for $\widehat{R}_{b,t}$ and $\widehat{\overline{y}}_b$ in place, total incomes in each bracket are simply given by:

$$\widehat{Y}_{b,t} = \widehat{R}_{b,t} \times \widehat{\overline{y}}_b \ ; \forall \, b \text{ and } \forall \, t \in \{2018, 2019\} \tag{18}$$

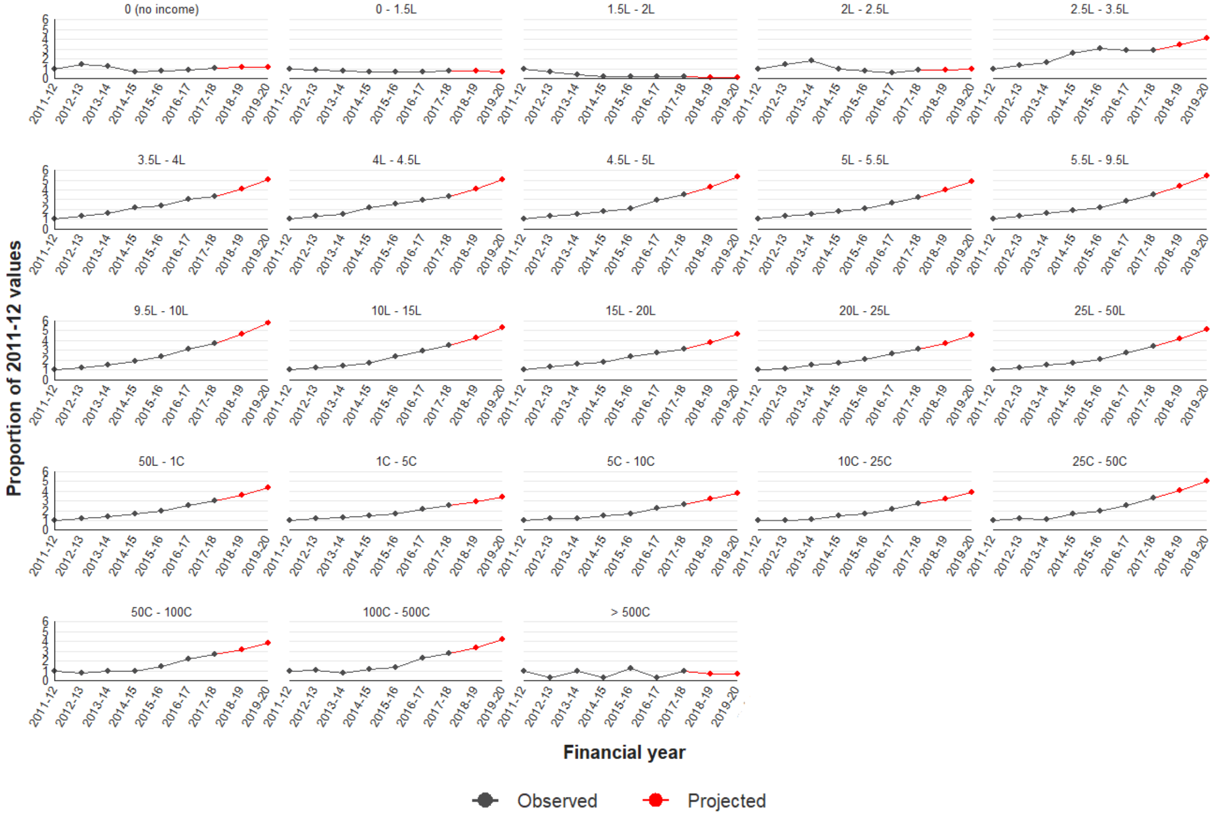
These projections essentially rely on the assumption that the trends in growth rates of total returns in each bracket during the 2011-2017 period allows us to identify the growth rates in 2018 and 2019 as well. The assumption may hold reasonably if I believe the nature of economic growth did not abruptly change in 2018-19 and not otherwise. Nonetheless, as shown in Figure 12, our projections imply a very smooth growth rate of tax-filers in each bracket, in line with what simple polynomial projections from pre-trends would suggest.

Figure 11: **Constant average incomes in each tax bracket**



Note: *Author estimates based on ITRS data. The figure shows the average incomes reported in each tax bracket between 2011 and 2017, with values for 2011 normalized to 1. We see that average incomes in each bracket were essentially constant through this period.*

Figure 12: **Projections of tax data for 2018 and 2019**



Note: *Author estimates based on ITRS data. The figure presents the growth of total returns filed in each tax brackets. The values in 2011-12 are normalized to 1. The blue dots till 2017 are what are actually observed in the data and the red dots are my projections for the years 2018 and 2019 based on the observed growth rates between 2011-2017.*

A.3. Longer horizon trends, 1980-81 to 2019-20

Figures 13 to 16 present income inequality over an extended time frame going back to 1980. These help contextualize the results for the most recent years to the estimates in the longer-run. In doing so, we see a steeper decline in the bottom 50% and middle 40% shares (Figure 13). At the same time, top 10% have doubled (Figure 14) and top 1% have grown 5 times (Figure 15). Moreover, looking at the growth rates of cumulative incomes confirms that majority of this growth has been captured by the very top. Figure 16 suggests that incomes of the very top percentile grew to the tune of 2500%-3000% compared to less than 250% for majority of the population.

Figure 13: **Bottom 50% and Middle 40% shares, 1980-2020**

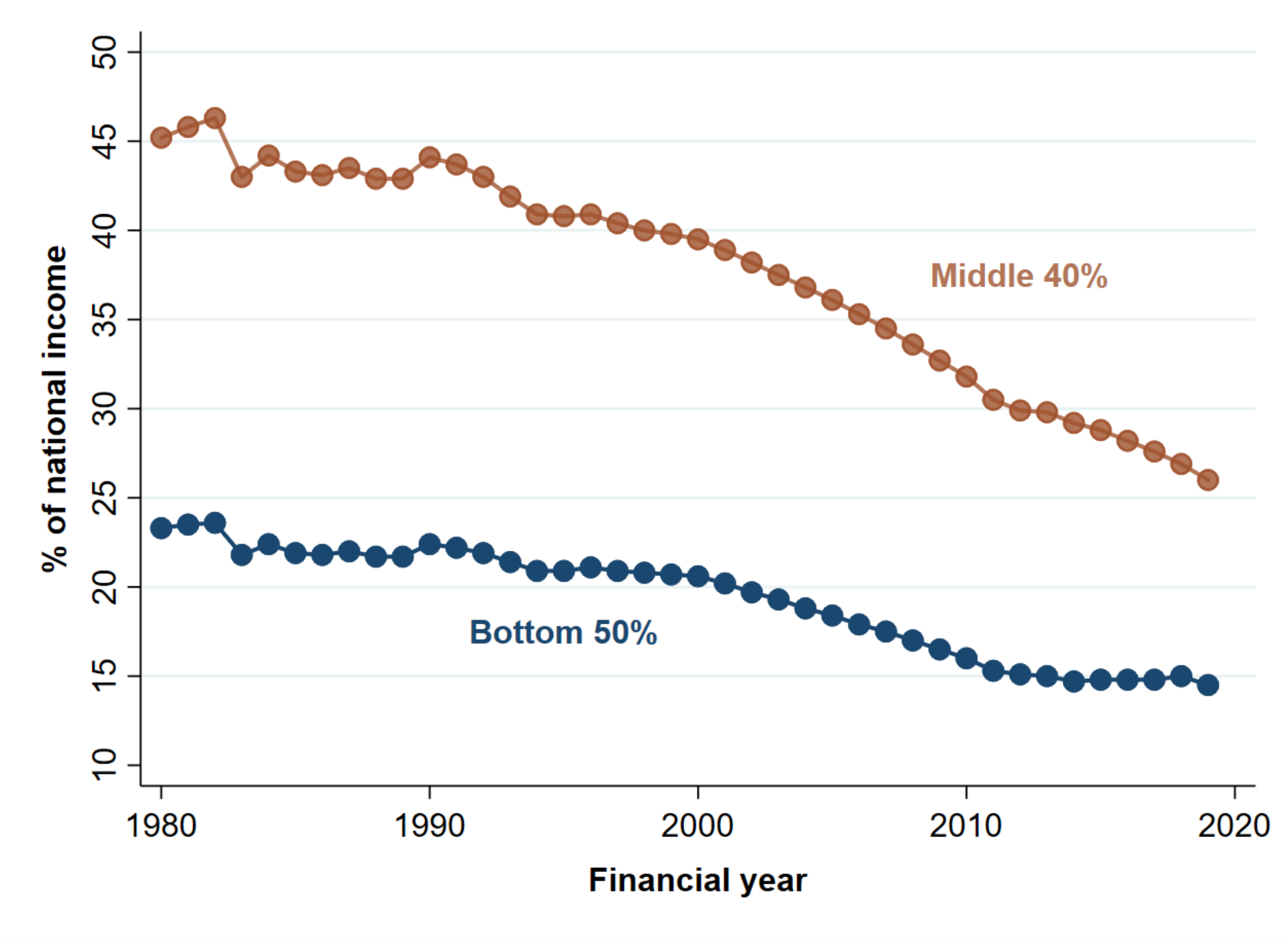


Figure 14: **Top 10% shares, 1980-2020**

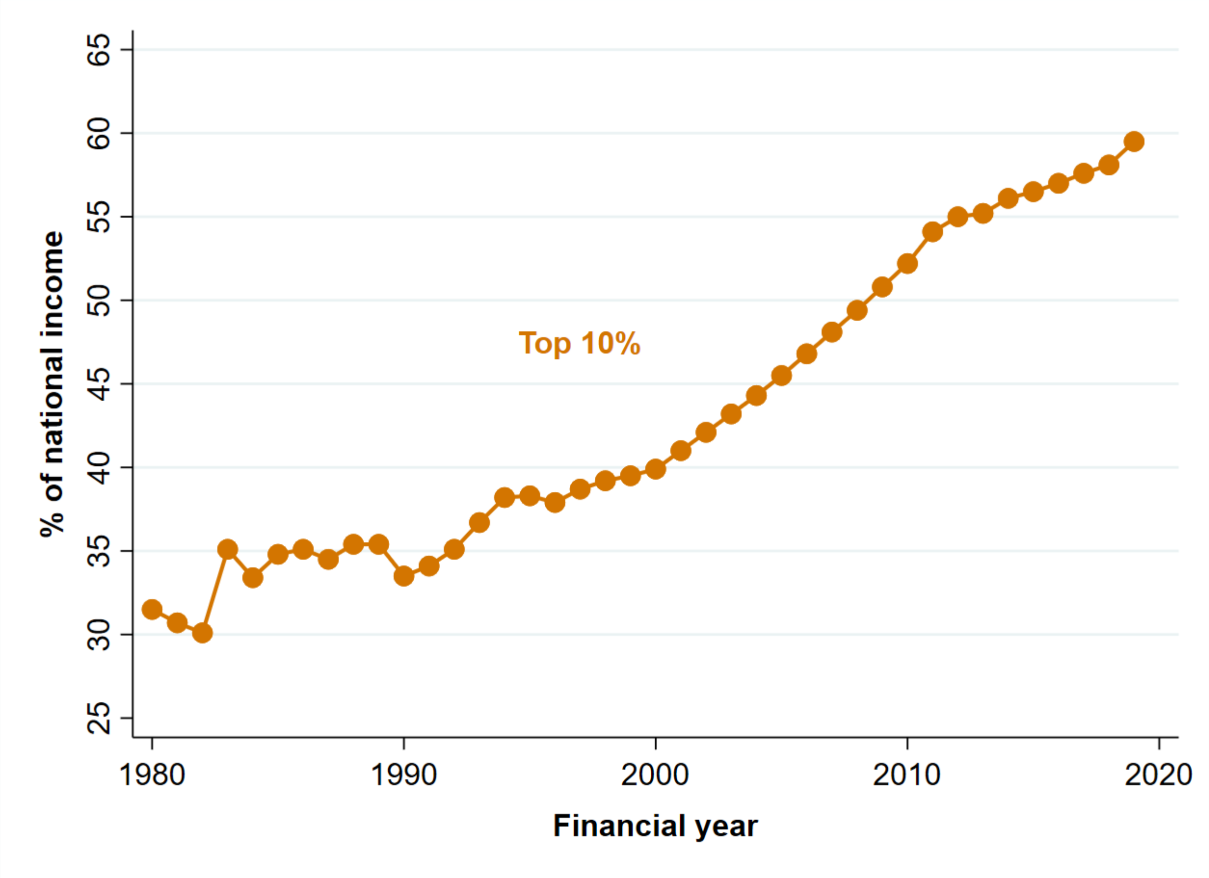


Figure 15: **Top 1% shares, 1980-2020**

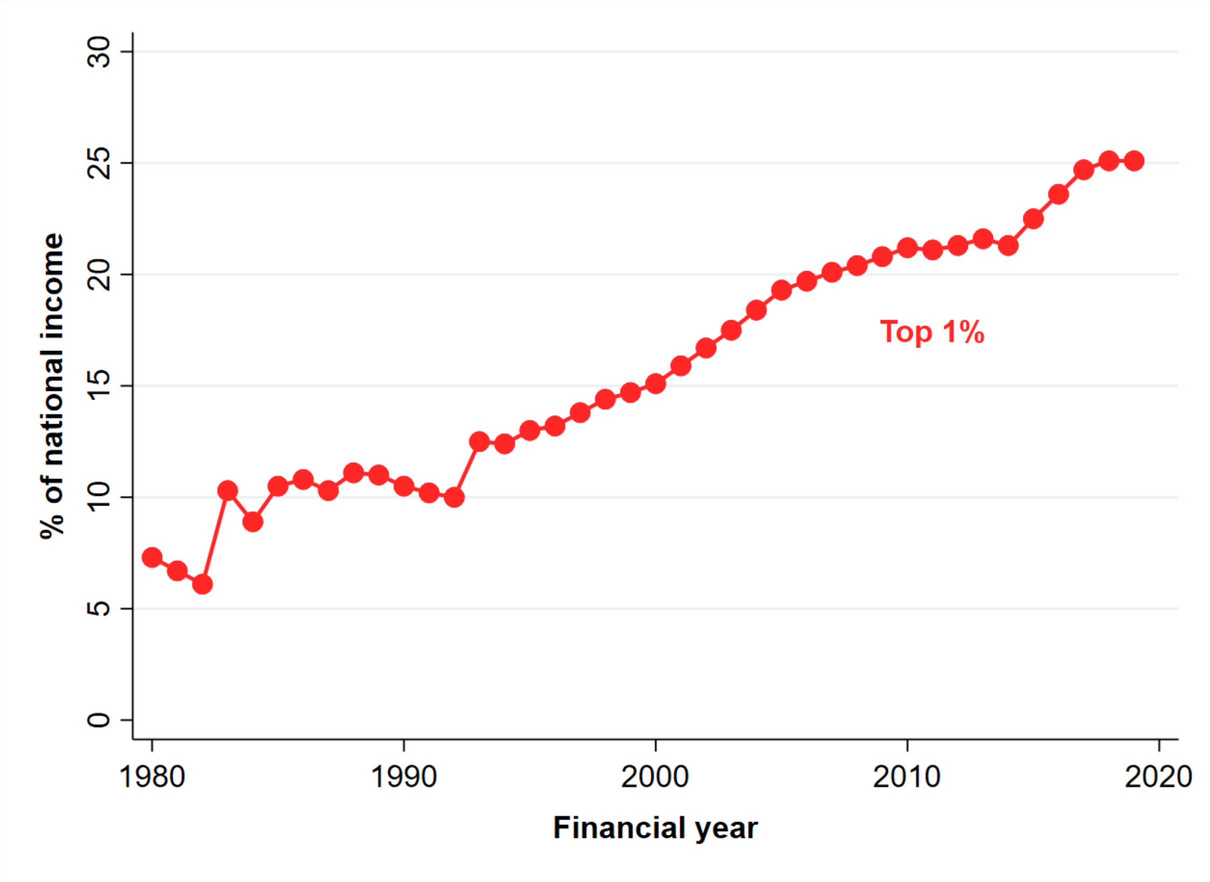
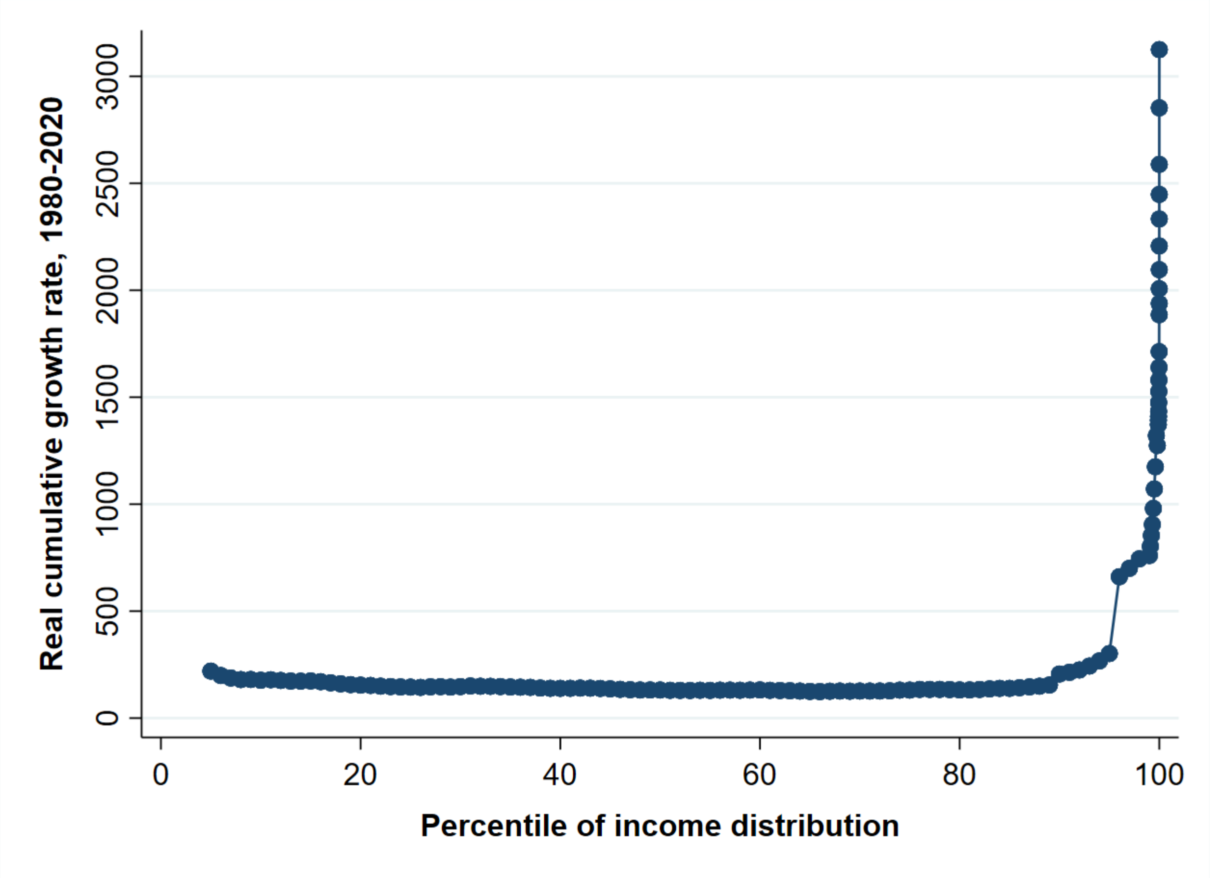


Figure 16: **Real growth rates of income, 1980-2020**



A.4. Further robustness checks

In the text, as robustness checks, I present top 10% and top 1% shares from 8 different estimation strategies (Figures 9 and 10). Here I present similar results for the bottom 50% and middle 40%.

Figure 17 presents the results for the bottom 50%. As we can see, the benchmark estimates (black) for the bottom 50% shares are higher than all other estimation strategies in recent years including matching (orange), re-weighting (red), and alternate income-to-consumption scaling ratios (gray).

Figure 17: **Bottom 50% shares from 8 alternate strategies**

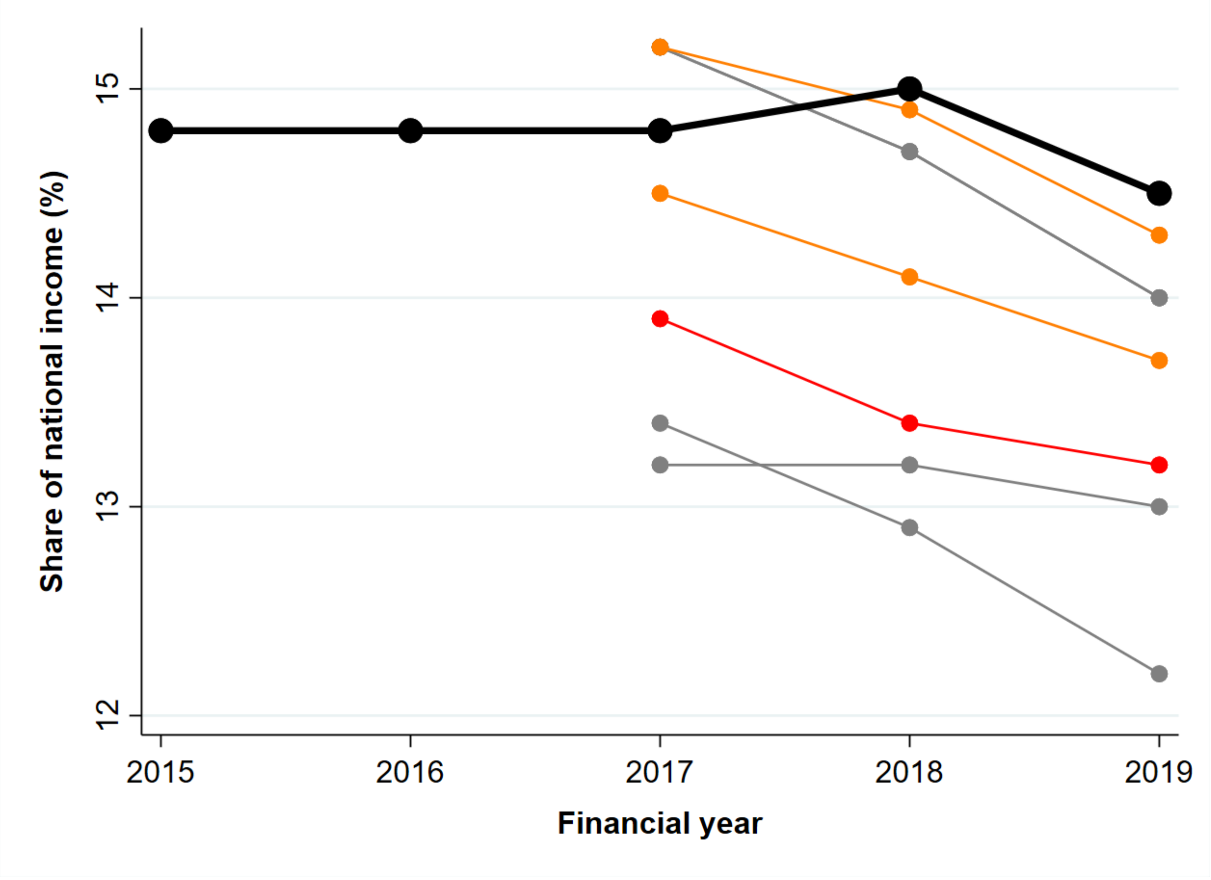
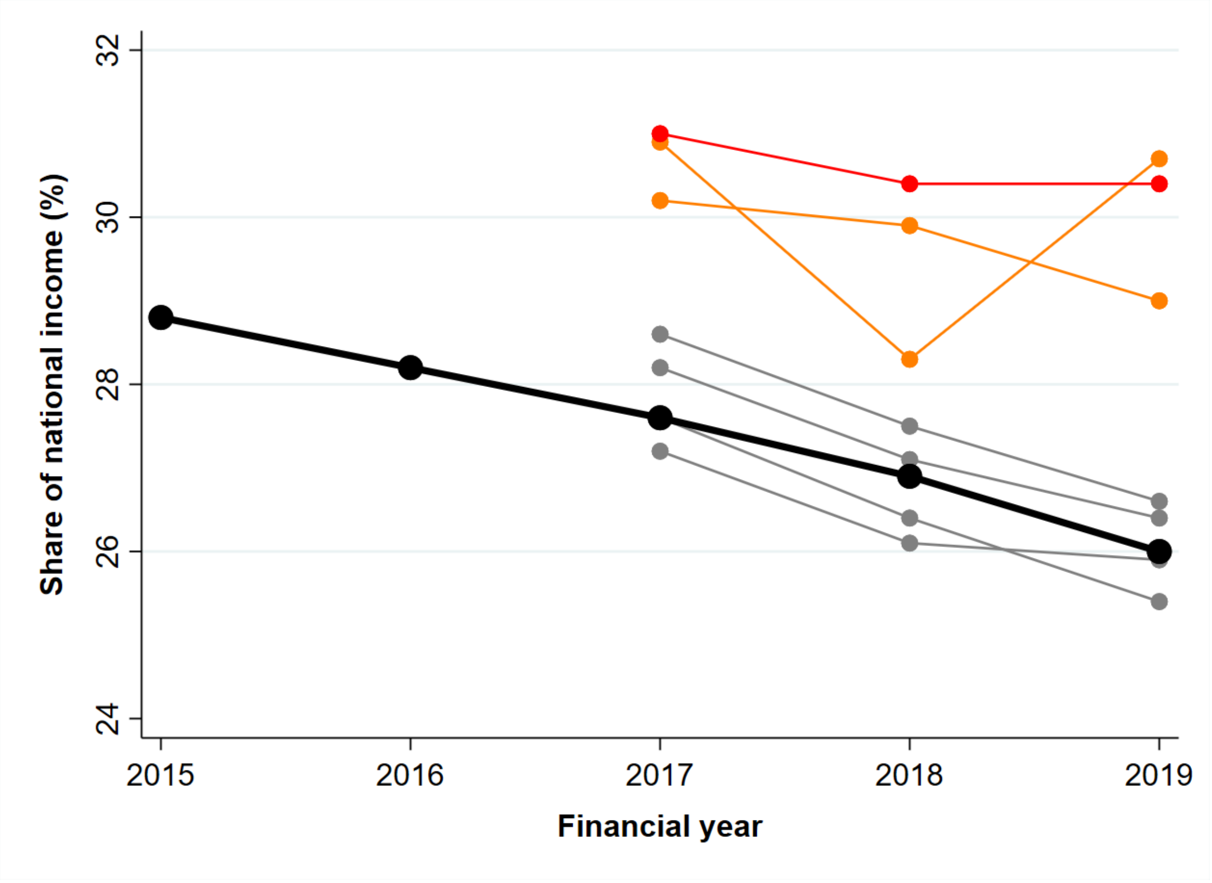


Figure 18 presents the results for the middle 40%. Here we see that the matching and reweighting estimates deliver slightly higher middle 40% shares compared to the benchmark but the trends of a decline are observed all across the board. Interestingly, the various different consumption-to-income scaling ratios deliver quite similar results. Once again, these can be read as a relative middle ground

Figure 18: **Middle 40% shares from 8 alternate strategies**


Share of national income (%)
Financial year