Measuring Income Inequality in Pakistan: Constructing Distributional National Accounts

Ali Ahmed*

Department of Economics
University of California, Berkeley

Undergraduate Honors Thesis
Advised by Gabriel Zucman

Abstract

We combine household surveys and recently released tax data to reconstruct the pre-tax income distribution of Pakistan from 2012 to 2015. According to our estimates, the top 1% income share is 30.2%, the top 0.1% share is 13.4% and the top 0.01% share is 5.1%. These top income shares are significantly higher than those in neighboring India. Instead, Pakistan's top income distribution is similar to that of the Middle East and Brazil. The middle 40% share is 30.0% and the bottom 50% share is at a mere 11.6%. We hope that this research informs the debate on Pakistani income inequality and provides a starting point for further analysis towards full distributional national accounts.

*Email address: ali.ahmed@berkeley.edu. I would like to thank Khaleel, Kulsum, and Alya Ahmed for their support throughout this process. I would also like to acknowledge helpful advice from Gabriel Zucman and Emmanuel Saez. All computer codes related to this report, including final combined microdata are available here.

1 Introduction

In 1965, the chief economist of Pakistan and later policy director of the World Bank, Mehboob Al-Haq, famously alleged that 22 families dominated the economic and financial life of Pakistan. Although unsubstantiated, he claimed that they controlled two-thirds of industrial assets, 80% of banking and 79% of insurance assets in the industrial domain. The speech became a rallying cry against an economic system that seemed to be concentrating income and wealth in the rapidly emerging economy.[1]

Yet since his declaration, there has been a notable absence of research examining income distribution within Pakistan. Outside of aggregate statistics using limited survey data, little has been done to construct such an account. Without knowing such a distribution, policy-makers and academics lack the ability to understand the distributive impact of economic growth or redistributive policies.

This is not uncommon for developing countries. Data quality is poor, and even when survey data is available, it provides vast under-estimates of top incomes. Pakistan in particular presents a unique challenge due to the considerable size of its informal economy which is excluded from basic economic statistics, including aggregate GDP. Estimates show that this unrecorded activity ranges from 20% - 30% of the total economy (Arby, Malik and Hanif, 2010).

Luckily, inequality research has since advanced to include methodology that corrects for poor survey data quality by using tax data. Since Kuznets' (1953) original attempt, a number of academics have used tax data to construct series of top income shares (Alvaredo et al., 2011-2016). In addition, following seminal papers by Piketty, Saez and Zucman (2018), and Piketty, Yang and Zucman (2018), extensive progress has been

[1] See for instance https://web.archive.org/web/20130722001847/http://www.mhhdc.org/html/system˙blame.htm.

made to develop a more robust methodology for the combination of tax data and survey data to construct distributional national accounts.

In 2013, in response to public pressure for tax transparency, Pakistan first began to release tax directories. Each directory contains more than 20,000 pages in PDF format, detailing the amount of taxes paid by individuals, corporations and associations of persons for the fiscal year. We are able to modify existing PDF-to-text parsers to efficiently convert, scrape and organize this data into a usable format, unlocking a new resource for the research community to expand on.

This paper attempts to leverage this data to provide a more precise understanding of Pakistani income inequality. Our primary objective is to combine tax data with existing individual survey data to create a snapshot of the pre-tax income distribution of Pakistan. Our secondary objective is to expand beyond this singular snapshot, and provide estimates of the dynamics of this distribution throughout the 21st century. We hope that this provides a starting point for the researchers at WID.world and across the economics community to build better estimates of Pakistani income inequality, eventually leading to a full distributional national account.

To achieve our primary objective, we organize and clean individual survey microdata from the Household Integrated Economic Survey (HIES) module of the Pakistan Social and Living Standards Measurement Survey (PSLM). We the combine the appropriate tax data with the survey data at the tax exemption level. This exercise has its limitations. We observe a small but significant sample of taxpaying individuals from the population of eligible taxpayers (only 10.3% in fiscal year 2015). In addition, we must estimate the taxable income of individuals from an aggregate amount of taxes paid. Unlike in developed countries, rampant corruption in tax assessment is to be suspected. However, we are still able to construct lower-bound estimates for taxable income, given the relative simplicity of the Pakistani tax code. The survey data is comparatively better quality as it allows for interviewers to assess income from both informal and formal sources. The survey data also includes a thorough capital income assessment if the income-to-consumption ratio of a household is below a certain ratio.

In brief: we provide new estimates on income inequality. To our knowledge, this is the first time an income distribution for Pakistan has been generated using tax, survey and national accounts since Haq (1964). This provides a much more precise illustration of income inequality than aggregate statistics. Our findings are summarized in table 1a, with a more in-depth discussion on results in section 4. According to our estimates for 2015, the top 1% income share was at 30.2%, on par with the Middle East and Brazil. The top 0.1% share was at 13.4% and top 0.01% share at 5.1%. The middle 40% share is at 30.0%, comparable to that of India which stands at 29.2%. The bottom 50% share is at 11.6%, more than South Africa and the Middle East, slightly less than India and China, and on par with the United States.

Table 1A - Distribution of National Income in Pakistan using Combined Microdata, 2015

	Labor Force	Average Annual Income (PKR)	Average Annual Income (USD)	Income Share
Full Population	65,422,753	279,601	$1,985	100%
Bottom 50%	32,711,376	64,927	$460	11.6%
Middle 40%	26,169,101	209,513	$1,487	30.0%
Top 10%	6,542,275	1,633,230	$11,595	58.4%
Top 1%	654,227	8,437,088	$59,903	30.2%
Top 0.1%	65,422	37,474,289	$266,067	13.4%
Top 0.01%	6,542	142,210,514	$1,009,694	5.1%

We stress there are strong limitations to our data sources and that this analysis would be strengthened substantially with more transparent income and wealth statistics. It is known that the government retains data such as reported taxable income of taxpayers, and wealth statements for each taxpayer that is not included in the tax directory.

Releasing those statistics for research, along with taxable income levels would provide academics with much more accurate data. That being said, we consistently use lower-bound estimates for all assumptions to address these gaps.

This paper is a starting point for researchers to expand on. We borrow heavily from the methodology used to understand Indian income inequality (Chancel and Piketty, 2017), and hope that this analysis can be expanded on given our new tax data source. We have taken great care in constructing readily available and readable computer codes[2] for others to contribute to and run alternative scenarios on given alternate assumptions.

The paper is organized as follows. Section 2 examines the literature regarding Pakistani income inequality. Section 3 presents our data sources and methodology. Section 4 presents our results and compares them with other studies that perform similar analysis. We also present limitations. Section 5 concludes.

2 Literature Review: Pakistani Inequality

As previously mentioned, academic research surrounding income inequality in Pakistan has been limited due to data quality issues, not dissimilar to those faced by other developing countries. In recent years however, new research constructing income distributions has emerged for countries like Chile (Atria, Flores, Sanhueza and Mayer, 2018), China (Piketty, Yang and Zucman, 2017) and neighboring India (Chancel and Piketty, 2017) using novel methodology combining tax and survey data. This research is inspired by those efforts.

[2] See https://github.com/ali-wetrill/DistributionalNationalAccounts-Pakistan for an interactive walkthrough of this research. The datafiles folder contains 5 jupyter notebooks (with .ipynb extensions) that contain both code and text that run through the methodology described in this paper. The taxpayerData and hiesSurvey folders contain the tax and survey data for others to use. miscData contains other data sources that are used in the notebooks. The finalData folder contains cleaned and organized microdata from the SurveyCalculations.ipynb file and TaxCalculations.ipynb. The interior finalCombinedMicrodata folder contains the combined data from the CombiningData.ipynb file. A more detailed account on how to understand the data files is available in the README.md file displayed when opening the link.

The focus of most research on Pakistan has revolved around industrialization and economic transition. As this distributional national accounts (DINA) research is very recent, research regarding Pakistani income inequality has generally relied on survey data and aggregate statistics, with several notable exceptions.

Some of the earliest research measuring income inequality came from Khadija Haq who obtained access to tax data and constructed income distribution (Haq, 1964). The study examines the change in income distribution from the country's independence in 1947 to 1958, using calculations for taxable income in lieu of more detailed data. We draw on this in our study, creating potential taxable incomes in a similar manner. The top 1% of the distribution captured roughly 16% of income, while the bottom 50% capture roughly 20% during this time period. A table of these early results available in appendix 1A. Income shares across the time period for the top 1% and bottom 50% are available in appendix 1B and 1C for reference. During this time, the income shares calculated for West Pakistan (what is now Pakistan) barely changed. This reflects the policies of General Ayub Khan during the country's purported 'Golden Period' of economic development. At the time Pakistan was hailed as a model for development, with countries such as South Korea seeking to emulate Pakistan's growth model.

In later years, several studies create summary statistics drawing from HIES grouped survey data. Bergan (1967) generates a Gini coefficient of 0.381 during the same year as the end of Haq's study. This number is replicated by Jeetun (1978), who goes further by calculating Kuznets Total Disparities Measure and other summary statistics. The limitations of this early HIES grouped data are apparent. General lack of granularity and accounting for informal income are made clear in Kemal (1981). Further efforts to calculate aggregate inequality statistics were performed on later HIES grouped survey data by de Kruijk and Leewan (1985), and de Kruijk (1986), but results were still based on household income rather than individual income.

Although using underlying household units was seen as a better measure of income inequality at the time, given the prevalence of extended family households and culture of income sharing, Ahmad (2000) shows that this biases inequality measurements. The study uses HIES microdata to show that high-income households predictably have more individuals than low-income households. His results also serve as inspiration for our research, as our HIES survey microdata is also provided in household and individual units.

We see several other studies in recent years that use novel methods to better account for aggregation bias in grouped data and household microdata that lacks individual units including Ercelawan (1988), Kemal (1994), Jafari and Khattak (1995) and Choudhry (1995).

With the exception of Haq (1964), inequality research has relied solely on HIES data, with most on HIES grouped data. In addition, most research has been confined to aggregate statistics rather than distributional analysis. Studies using HIES grouped data in particular are sensitive to bias from the intervals chosen by the makers of the survey (Siever, 1979), and most other studies use the household as a frame of reference, which does not provide an accurate accounting of inequality. Even with HIES microdata, the shortcomings of household survey data in observing the evolution of inequality are well known; because of underreporting and under-sampling, even with the requisite weighting surveys fail to properly capture inequality dynamics at the top of the distribution (Atkinson and Piketty, 2007, 2010).

Our contribution to this literature is twofold. Firstly, as mentioned in our introduction, we combine household surveys with newly released tax data in order to correct top incomes in the survey-based distribution, painting a more accurate picture of inequality than previous studies. Secondly, we unlock this tax data for further research. To my knowledge, this is the first study to use tax and survey data to construct an income distribution since Haq (1964).

3 Data Sources and Methodology

3.1 Data Sources

Here we present the data[3] used to produce the top income distribution for 2012 - 2015 (the years covered by the tax data). We also present the other data sources used to estimate income growth rates in our attempt to expand our tax data to cover 2004 - 2011. We use survey data as well, which accounts for the rest of the distribution and covers the entire time period. Finally, we present a plethora of miscellaneous data sources that provide comparative context to our results.

3.1.1 Federal Board of Revenue Tax Directory

In 2013, the Pakistan Federal Board of Revenue (FBR) began to release Tax Directories. Each directory reports a name, amount of tax paid, and national tax number (NTN) or computerized national identity card (CNIC) number[4] serving as a unique identifier for each fiscal year. The directory covers those who filed returns digitally. Manually filed returns were entered into the directory by the FBR[5]. The directory includes voluntary payments, which are classified by the FBR differently from withholding tax and collection on demand.[6] The directory holds such voluntary payment

[3] All data sources can be found under the datafiles folder.

[4] In fiscal year 2013 the NTN number was included. From fiscal years 2013 - 2016, this was replaced with an NTN / CNIC column, corresponding to the integration of the NTN system with the CNIC system in an effort to stop tax evasion. The CNIC system is more comprehensive as it covers all individuals with a national identity card.

[5] A large proportion of taxpayers in Pakistan are non-filers, who pay withholding tax. Across fiscal years 2013 - 2017, this represents 60-70% of direct taxes paid. We note that this data is not included in the tax directory, only information on filers. Undoubtedly the government has information on the income of these individuals, who are likely to be salaried earners. Our analysis would benefit greatly, and top income shares would be revised downward, if this data were made available.

[6] More information on these definitions can be found in the FBR Yearbook. Voluntary payments are classified as a tax payment according to the appropriate tax rates. Collection on demand (CoD) occurs when the individual hasn't paid, and the tax collector arrives at their household. CoD generally includes a fine, but is subject to corruption. The fact that the directory is made up of only voluntary payments from

information about corporations, associations of persons (AOPs)[7], and individuals. It should be noted that the FBR states that a considerable number of manually filed returns could not be entered due to missing identifiers.[8] As a result, we may underestimate the number of filers and derived statistics on taxable income and revenue.

We will now examine our specific sample of tax data. We define *TD Tax Filers* as the individual tax filers listed in the tax directory. This includes all individuals who filed taxes digitally, and significant portion of those who filed manually. It is important to note that our definition is less than or equal to the total number of tax filers because (as stated above) not all manually filed returns were able to be entered into the directory. We are unable to find the number of unrecorded manual filers. We define *TD Taxpayers* as TD Tax Filers who paid more than 0 in taxes. We can examine how large these numbers are when comparing them to the labor force and the number of individuals who are above the tax exemption limit. We also reach the conclusion that most of this population is paying personal income tax, and in the cases where the tax paid is capital gains or property tax using personal income tax is an appropriate lower-bound estimate of the individual's income (see section 3.1.6 for this discussion). We can see the raw numbers of tax filers and taxpayers in the tax directory for all fiscal years below, in table 2.

filers (mostly digital) is positive for our analysis. Although it is a smaller sample without CoD, this classification suggests our tax sample is not as biased by corruption. Even so, we emphasize that every number we use is a lower-bound estimate.

[7] This is categorization can be thought of as a joint-filing for which we do not know how many individuals file together. The FBR has a definition for this here.

[8] The exact phrasing stated in the directory is "Manually filed Returns have been entered into the system. There may be some discrepancies at the level of filing or data entry. Moreover, a considerable number of manually filed Returns could not be entered into the system despite best efforts due to missing identifiers on the Returns."

Table 2 - Number of TD Tax Filers and TD Taxpayers

Fiscal Year	Labor Force	Eligible Taxpayer Population	Number of TD Tax Filers	Number of TD Taxpayers
2013	60,357,706	6,029,151	728,346	422,744
2014	62,025,233	3,788,656	780,462	494,375
2015	64,070,380	5,097,884	1,000,718	692,360
2016	66,149,979	7,800,316	1,135,764	803,412
2017	67,602,890	8,788,375	1,680,404	1,087,486

Let's look at these numbers as a percentage of the labor force. If the tax directory is judged to be representative of the entire population of individual filers and payers, the number of taxpayers and tax filers in Pakistan has increased dramatically over the past few years as shown in table 3. Only 1.2% of the labor force filed returns in fiscal year 2013. By fiscal year 2017, that number more than doubled to 2.5%. However, only about two-thirds of filers actually pay any taxes, resulting in taxpayers being a mere 1.6% of the labor force.

Table 3 - FBR Tax Return Summary Statistics

Fiscal Year	Labor Force	TD Tax Filers (% of Labor Force)	TD Taxpayers (% of Labor Force)	TD Taxpayers (% of TD Tax Filers)
2013	60,357,706	1.21%	0.7%	58.0%
2014	62,025,233	1.26%	0.8%	63.3%
2015	64,070,380	1.56%	1.08%	69.2%
2016	66,149,979	1.72%	1.21%	70.7%
2017	67,602,890	2.49%	1.61%	64.7%

The rate of increase in the taxpaying population over such a short period of time is impressive. The absolute proportion of taxpayers is similar to levels observed in France and the United States in the mid 1910s, but much lower than the proportion in the interwar period (about 10-15%) and in the decades following World War II, where the proportion soared to more than 50% in these two countries (Piketty, 2001; Piketty and Saez, 2003). The current figure is similar to India in the mid-1990s (the country's proportion of individual taxpayers is nearer 7% today), and slightly below China today (Chancel and Piketty, 2017; Piketty, Yang and Zucman, 2017; CIA World Factbook).

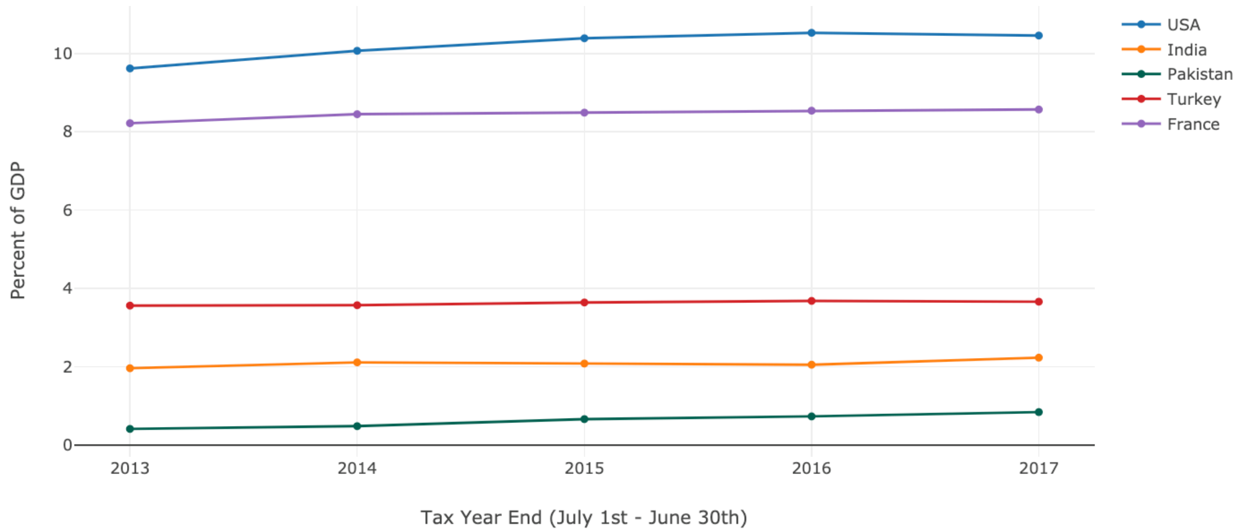
However, our sample is still significant. Pakistani tax code has a tax exemption limit for incomes below 350,000 PKR in fiscal year 2013, and 400,000 PKR from fiscal year 2014 - 2017. This means that the full population of taxpayers should be roughly the top 10% of income earners, according to the HIES survey data (the exact percentiles are listed in appendix 3). If we calculate the number of taxpayers covered by the tax directory over the proportion of the labor force that represents the full population of *eligible* taxpayers (judged by HIES survey data), our sample represents as much as 12.4% of the full population of these top incomes. This calculation is replicated for all fiscal years in table 4. It is possible that these individuals are also skewed towards the lower end of the top income distribution, as earlier editions of the FBR Yearbook (a summary of tax collection created by the FBR) states most income taxes paid are from salaried individuals.

Table 4 - FBR Tax Directory Sample in Context

Fiscal Year	Eligible Taxpayer Population	TD Taxpayer Sample (% of Eligible Taxpayer Population)
2013	6,029,151	7.01%
2014	3,788,656	13.05%
2015	5,097,884	13.58%
2016	7,800,316	10.3%
2017	8,788,375	12.37%

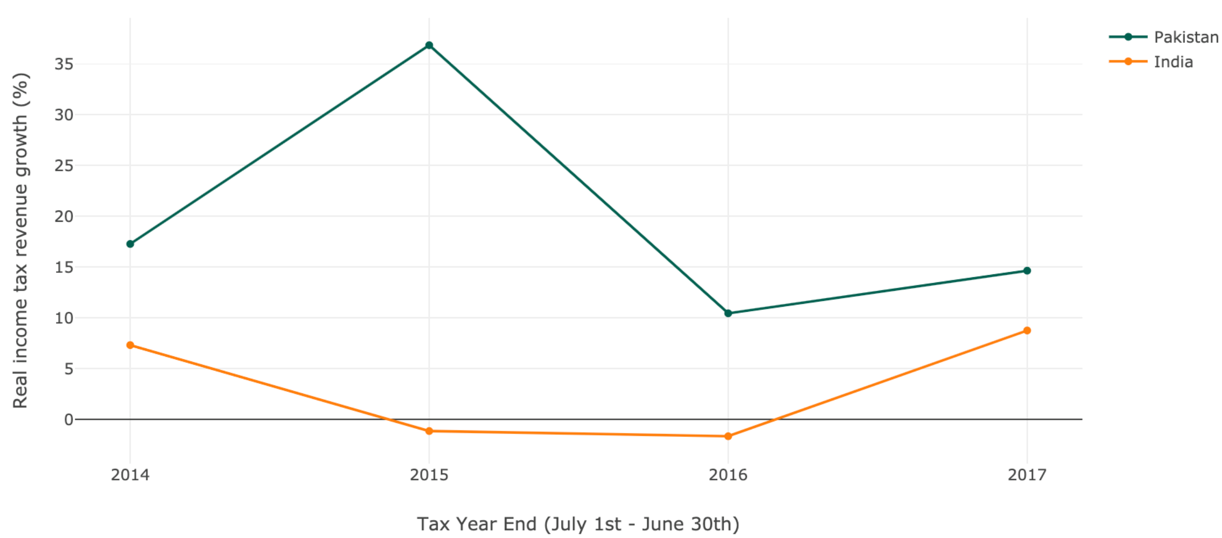
We can compare personal income tax revenue as a proportion of GDP across countries to put this data in perspective in figure 1. We see that countries like Turkey and India with a similar personal income tax environment having more personal income tax revenue as a percent of GDP than Pakistan (this is assuming that the entire sample is paying under personal income tax, meaning Pakistan's number is an underestimate).

Figure 1 - Personal Income Tax Revenue (% of GDP)



Countries such as France and the United States, which have nearly their entire population contributing personal tax, form an upper bound. Pakistan's income tax revenue as a proportion of GDP is half of India's, where Chancel and Piketty (2017) were able to perform a distributional analysis of inequality. Regardless, the rate of increase of this statistic, and the taxpaying population in general compared to countries like India, tell us that in subsequent years this tax directory data source will become more valuable and more accurate.

Figure 2 - Personal Income Tax Revenue Growth, India and Pakistan



Many of our summary statistics are from the FBR Yearbook that is published alongside the tax directory. It includes summary statistics and provides more context to the tax directory. We include these yearbooks in our computer codes[9], along with the original tax pdfs and the text versions.

3.1.2 PSLM HIES Survey Data

The Household Integrated Economic Survey (HIES) is a module of the Pakistan Social and Living Standards Measurement Survey (PSLM) conducted by the Pakistan Bureau of Statistics (PBS). Since 2004, the HIES has been part of the PSLM with microdata publicly released on the PBS website[10]. We use this data to estimate bottom incomes. The survey covers years 2004 - 2016. It is important to note that the HIES survey has been conducted yearly since 1963. We do not have privileged access to the microdata before 2004 but further research and collaboration with the PBS could greatly enhance our sense of distributional national accounts, allowing us to generate a long-run analysis of inequality.

The number of households covered by the survey varies dramatically from 2004 - 2016. It ranges from 14,707 households with 21,170 income earning individuals in 2004 to 78,505 households with 122,664 income earning individuals in 2014. More information can be found in Appendix 2. The universe of the survey consists of all provinces in Pakistan, excluding northern military restricted areas. The sampling frame[11] specifies large-sized cities selected for urban households as well as rural areas. The survey is completed by an interviewer, and conducted in person (it is not a mail-in sample).

---

[9] The copies of relevant FBR yearbooks are located under the supplementTaxInformation folder here.

[10] The original data from the entire PSLM survey can be found at http://www.pbs.gov.pk/content/microdata. Our cleaned and organized HIES data can be found under finalData while a copy of all of the relevant microdata under the HIES module can be found under datafiles/hiesSurvey.

[11] More information on the sampling frame can be found in the introduction PDFs released with each survey. Many are kept under *data_files/hiesSurvey* within our files. The selection of primary and secondary sampling units changes across years, especially with regard to cities selected to represent urban areas.

The detail of the HIES survey is extraordinarily granular. It provides information on employment income, non-employment income, and various categories of other income and other annual earnings for which occupation and industry categories are not specified. We assume this is put in place to capture income from the informal economy, and indeed over 22% of individuals in 2015 reported income in these categories of the survey. The survey data also includes financial assets owned by the household, and other forms of household capital income. Outside of income, it also provides an extensive accounting of household profile and consumption data (ranging from daily used items, clothing, food, education and healthcare).

Notably, although the survey is weighted, we see dramatically different growth rates in income across surveys (see section 3.2.3 for a more detailed discussion and table 5 for the corresponding rates), revealing that each survey is a self-contained sample representative of that year. This limits our findings as we cannot compare growth rates across years.

3.1.3 World Inequality Database, National Accounts Data

We use GDP and National Income data from the World Inequality Database (WID.world). The GDP data is based on UN-SNA main tables for our sample until 2014. Later years are computed based on GDP growth rates from the IMF World Economic Outlook. WID.world derives National Income from IMF Balance of Payments Statistics, the World Bank, and the IMF Economic Outlook. Again, it is important to note that none of these statistics include the informal economy. This leads to an overestimate for any number calculated as a percentage of GDP or national income.

In any given year, there is a gap between national income reported from WID.world and the weighted income reported in PSLM HIES survey data. Even with requisite weighting, the survey data covers less than 30% of national income in fiscal year 2015.

One of the reasons most commonly cited to explain this gap is under-reporting and under-sampling of top incomes in survey data (Banerjee and Piketty, 2005). A detailed analysis of this can be seen in the case of India, where top incomes account for a large percentage of this under-reporting (Chancel and Piketty, 2017). As we aim to provide better estimates for top incomes in a combined set of microdata, we naturally correct the survey data to better fit national accounts data. We can see this in figure 8. We discuss these results in section 4.5.

A proportion of the remaining gap can be plausibly explained by missing top incomes in our tax data. However, the goal of this research is not focused on solving this gap and as has been suggested, we cannot assume that top incomes account for most of the difference (Lakner and Milanovic, 2015). National accounts statistics include imputed rents, and contain inconsistencies related to the informal economy (Arby, 2008). This means our sample will never quite cover 100% of national income derived from national accounts data.

3.1.4 World Bank Population and Labor Force Data

We define the population of working individuals as the labor force of Pakistan. We take this data from the World Bank, which defines the labor force as people ages 15 and older who are currently employed, and people who are unemployed but seeking work. They derive this using data from the International Labor Organization, ILOSTAT database and World Bank population estimates. We use this information when analyzing the taxpaying population. We also collect World Bank population estimates for comparative summary statistics (including percentage of the population below 30).

3.1.5 UNU-WIDER Government Revenue Dataset, OECD Dataset

In analyzing the tax data, we perform a number of comparisons with other countries. We take data on personal income tax revenue as a percentage of GDP for Turkey, the United States and France from the OECD website. We take direct taxes as a percentage of GDP from the UNU-WIDER government revenue dataset. Each dataset has included their number for country GDP during the year they calculate the percentage. When comparing percentages, we re-calculate gross amount collected and divide by the same GDP metric.

3.1.6 Ernst and Young Worldwide Tax and Immigration Guide

From the amount of taxes paid for each individual, we back-calculate the amount of taxable income. For this exercise, we make use of the EY worldwide tax and immigration guide[12] to find exemption limits and tax rates.

The FBR categorizes income as earned from primary occupation, capital gains, property, and miscellaneous activities. There is some variability in the tax schedule across the 2013 - 2017 fiscal years which we account for, but most taxes remain the same. Income from capital gains is taxed at 0% if the asset is held for more than two years. If not, the tax varies from 8 - 12.5%. Property taxes are similar as after two years of holding the asset the tax rate levied is 0%. The main tax category, primary occupation, is taxed under two different schedules - salaried and non-salaried income. If more than 50% of your income comes from salary, you are subject to the salaried tax schedule. The guide also provides information on what is considered deductible, which can be used to calculate actual income from taxable income. Since we cannot make any assumptions about this, we use the lower bound and assume that an individual's taxable income is the same as their pre-tax income.

[12] Copies of the guides can be found under datafiles/taxpayerData/supplementalTaxInformation.

In our calculations, we assume all of the income earned by individuals falls under primary occupation. We calculate salaried and unsalaried taxable income estimates. It is important to note that the property and capital gains made by individuals that are not taxed (due to under-reporting or tax-dodging) would not appear in the directory, and therefore we would not capture this gain in our taxable income estimate. In addition, if we assume the 8 - 12.5% rate is the only tax paid by an individual, then on average our calculations overestimate an individual's taxable income, thus leading to higher national income shares for top incomes than if we use the main salaried and unsalaried tax schedules. In this way and following this tax guide, we can reasonably assume that using our taxable income estimates are underestimates of top pre-tax incomes.

3.1.7 Miscellaneous

There are several data sources that we use to supplement the information from above. For instance, UNU-WIDER and OECD datasets only have direct income tax revenue information on India, but not personal income tax revenue. We find that number from the Indian Ministry of Finance. We use brochures from the FBR to verify the EY tax and immigration guide tax exemption limits and fill in the gap for fiscal year 2012 (as no guide was published). There are a few other comparative statistics we use to put our data in context, including China's personal income tax revenue as a percent of GDP. These are all included in the same section as other data sources.[13]

3.2 Methodology

We draw on a wealth of knowledge from researchers at WID.world. Following a seminal paper on United States Distributional National Accounts by Pikketty, Saez and Zucman (2018), researchers at WID.world have been developing methodology to combine tax, survey and national accounts data. This has been conveniently compiled into

[13] These miscellaneous datasets are included in datafiles/miscData.

Distributional National Accounts (DINA) guidelines that present a workaround for countries with limited data (Alvaredo et al., 2016). We also draw heavily on methods used in estimating income distribution in India, as the country is comparable to Pakistan in terms of data quality issues (Chancel and Piketty, 2017).

3.2.1 Loading, Cleaning and Organizing PLSM HIES Survey Data

We extract the relevant income information from our survey data. The HIES survey data covers years 2004 - 2016. We first extract individual employment information, and sum the various categories of income to create total individual income. Again, this includes income from primary and secondary occupations, pensions, rent, remittance and other. We then load the weights assigned to primary sampling units. We merge this tables with individual employment information to create household and individual income tables, although we only use individual income tables for our analysis. We also merge individual profile data which includes relation to head of household, age, sex, and birth year for summary statistics. Finally, we combine the different sources of reported income for the individual to create a balance sheet (which includes total gross income).

Many of the years include a pre-computed balance sheet which includes a summation of individual income and expenditure levels, along with an income to expenditure ratio. If that balance sheet is available, we make use of it. If it is not available, the above paragraph details how we create that balance sheet.

If the ratio of income to expenditure for a household is below 0.85, the survey does a thorough capital accounting of household assets. A new balance sheet for the household is made which considers income from financial assets, agricultural assets etc. We do not make use of this data because it is only calculated on the household level. This potentially skews our results to show more inequality, as this income supplements individual income within households considered below the (roughly) top 10% (or whatever junction percentile applies to that year). Further analysis would take this income and distribute among different members of the household.

Finally, we perform corrections to the data. Keeping in line with years that have individual balance sheets available, we remove all incomes where the individual was 'not seeking employment' and reported 0 PKR in income. For the year 2012, a few rows do not have any weights associated with them. We remove these rows. Finally, even for years with balance sheets, some data is incorrectly entered. We have to manually calculate the income of the individual as we did with years where no balance sheet is available.

This gives us a pre-tax income distribution which becomes more inaccurate for higher percentiles. We will now use the tax directory data source to bring us closer to true distribution by correcting top pre-tax incomes.

3.2.2 Deriving Top Incomes from the Tax Directory

The tax data presents a unique technical challenge as it is released in PDF format. Each year contains nearly 20,000 pages of tax returns, of which more than 90% are individual filings. We first convert the tax directories into text format. We cannot use standard PDF tools like Tabula or PDF2Text libraries[14] which preserve the ordering of the entries as they are too slow. We instead download a script from pdfminer.six[15], which is derived from the pdfminer library but is more efficient and compatible with python 3. We then run this script.[16] Once we have performed the conversion, we then create

[14] These are standard libraries used by computer scientists to convert PDF to machine readable text documents. By our last estimate, it took around 6 hours to convert a single year's tax directory. We considered the strategy of using Optical Character Recognition (OCR), but instead found a modified existing package that we were able to further modify for batch processing. As a result, we were able to reduced our PDF to text processing time to under 4 minutes per tax directory.
[15] One can download the library here https://github.com/pdfminer/pdfminer.six
[16] In datafiles/TaxCalculations.ipynb we have the code to run the conversion. For other's convenience, we have created a folder under datafiles/taxpayerData/txtCreateFolder, where one can put any PDFs. One can then run the code in the notebook, and text files will be created in the folder from the PDFs.

custom parsers to skim through each text file and extract the information on taxes paid, and NTN / CNIC unique identifiers.

This population of eligible taxpayers, defined by the tax exemption limit, lies at the top of the income distribution of the survey data (the tax data represents the top 11.8% of individuals in 2015, and climbs as high as the top 6.11% across our sample. See appendix 3 for the corresponding junction percentiles). We are only given a small sample of these eligible taxpayers, as only 10.3% of this population contributes to the tax base in fiscal year 2015 (see discussion of this in section 3.1.1). In addition, we are only given the raw amount of taxes paid by those individuals, without categorization for which type of tax the individual paid.

What we would like the tax data to show is total fiscal income which is defined as the total personal income that should be reported by individuals if all eligible taxpayers reported revenues to the FBR. In Pakistan, as in the case of India, we do not observe this because of the limited tax base (Chancel and Piketty, 2017). As a result, unlike Piketty, Saez and Zucman (2018) we cannot get a precise estimate of pre-tax income. We instead resort to constructing possible taxable incomes as a lower-bound estimate. This follows a well-documented pattern of using taxable incomes to estimate top income shares, used originally by Kuznets (1941, 1953), Atkinson (2005), and Piketty (2001, 2003). We face limitations due to corruption in tax assessment, but again we stress that these are lower-bound estimates.

We calculate these taxable incomes assuming all workers are either salaried or non-salaried, representing the two main tax schedules released by the FBR. A more through discussion of how we derive taxable income can be found in section 3.1.6, where we explain the semantics of the Pakistani tax code. By assuming the entire population is either salaried or non-salaried, we represent two lower-bound assumptions of taxable incomes. We transform each tax paid into these two possible taxable incomes which we take as a lower-bound estimate for pre-tax income.

3.2.3 Combination of Tax and Survey Data

Our original objective is to construct a profile of total personal income that matches national income data. With survey data, we are missing reliable estimates for top incomes. Several strategies can be used to correct for top incomes in survey data, including the modification of weights or multiplication of top income levels (Burkhauser et al., 2016). We prefer to assume a similar methodology as that detailed in the India paper (Chancel and Piketty, 2017) and in the DINA guidelines (Alvaredo et al., 2016) by assuming that surveys are reliable to a certain percentile, and tax data is reliable above that percentile.

We first need to calculate the percentile where we will combine the tax data with the survey data. We choose the income exemption level for each tax year as those individuals in the tax directory must have an income above that level. We compute the percentiles of the full survey data and find the percentile at which income is above this exemption level. This forms our junction percentiles that are shown in appendix 3.

We now seek to calculate top income growth rates. Our secondary goal was to establish a longer run picture of income distribution over time. We only have taxes from fiscal years 2012 - 2017, but survey data from 2004 - 2016. In the absence of tax data, we try to estimate various growth rates that could serve as suitable candidates for extrapolating our tax data. Evidence from Chancel and Piketty (2017) suggests that using the growth rates of the top incomes in the survey (defined by the incomes above the junction percentile) is best given no other data to draw from. However, after examining these growth rates, the average growth rates of the survey data, and national income growth rates (as well as their effect on our results), we see that all options are dubious at best.

Table 5 - Income Growth Rates (Survey and National Income)

Year	Taxpayer Income Growth Rate (Salaried)	Taxpayer Income Growth Rate (Non-Salaried)	Average Income Growth Rate (Survey)	Average Income Growth Rate (National Income)
2005	-1.55%	-1.55%	9.86%	7.16%
2006	2.83%	2.83%	-7.44%	5.05%
2007	1.39%	50.32%	39.64%	2.93%
2008	69.58%	55.19%	29.57%	2.23%
2010	-37.4%	-53.86%	15.4%	5.15%
2011	50.15%	50.15%	17.57%	3.31%
2012	58.7%	58.7%	36.17%	3.91%
2013	-10.59%	-10.59%	15.49%	4.34%
2014	53.07%	53.07%	23.46%	4.64%
2015	-17.46%	-17.46%	-7.86%	4.95%

These rather erratic growth rates tell us that the survey is a self-contained sample that does a better job at describing incomes within each year rather than across years. It limits our ability to analyze growth rates and changes across years where no tax data is available to ground our estimates. Our best option is to use the average survey growth rate, to keep the distribution in-line with average movements in income distribution in the survey. We stress that this is flawed, meaning only our results from 2012 onward are robust. Years before 2012 should be viewed as an imperfect estimate.

We also intended to use these growth rates to align tax data with survey data, as the fiscal year definition is from July 1st - June 30th. We cannot due to lack of reliable growth rates. Instead we test alternative scenarios by constructing series using both the fiscal year ahead of the survey year and before. For example, for survey year 2014, we test income shares by combining the data with taxes from fiscal year 2014 and fiscal

year 2015. Note for survey year 2012 we just use the tax data for fiscal year 2012. We see this does not significantly affect income shares for years 2012 - 2016, where tax data is available.
We again stress that more alternative scenarios and better proxies for growth rates of top incomes are needed for analysis across years 2004 - 2011. When we combine our survey and tax data at the junction percentile, any income that has shifted below the tax exemption threshold due to our selected income growth rate is adjusted to be at the threshold. We do this because we know the individual paid taxes, which means their income must lie above that threshold at a minimum. We find that this skews top 1%, top 0.1% and top 0.01% income shares towards lower numbers (and that this movement is significant from 2004 - 2011).

We then combine the data at the junction percentile to create our final sample of combined microdata[17] for visualization. To do so, we remove incomes above the junction percentile in the survey data, append the derived incomes from our tax data to our survey and evenly distribute the weights.

The weights now represent how we think of our sample of taxpayer data. Evenly distributing the weights assumes that our sample is representative of the population of eligible taxpayers. We believe that the actual tax data trends towards lower incomes among the eligible taxpayer population. If this is to be believed, our weight distribution is an underestimate. Further analysis should explore alternate scenarios with different distributions of the weights. This would generate a more robust range for us to examine.

With this combined microdata, we are able to produce more accurate visualizations of income distribution across the sample.

[17] This combined microdata is available in datafiles/finalData/finalCombinedMicrodata.

4 Results

4.1 Pre-Tax Income Distribution

Our results begin to paint a clearer picture of distributional national accounts. We observe pre-tax income shares of various segments of the population in table 1a. A significant share of national income goes towards the top 10% and top 1% of the population. We take notice of the extreme disparity in average incomes, characteristic of developing economies with open access to international markets.

Table 1A - Distribution of Pre-Tax National Income in Pakistan using Combined Microdata, 2015

	Labor Force	Average Annual Income (PKR)	Average Annual Income (USD)	Income Share
Full Population	65,422,753	279,601	$1,985	100%
Bottom 50%	32,711,376	64,927	$460	12%
Middle 40%	26,169,101	131,418	$933	18%
Top 10%	6,542,275	1,633,230	$11,595	57%
Top 1%	654,227	8,437,088	$59,903	30%
Top 0.1%	65,422	37,474,289	$266,067	13%
Top 0.01%	6,542	142,210,514	$1,009,694	5%

Our combined tax and survey microdata provide significant re-adjustments to distributions in the original PSLM HIES survey data, as shown in figures 3a, 3b and 3c. We observe that in the original sample, the top 10% of income earners accounted for nearly 35% of national income. In our corrected data, that number jumps to 58.4%. We see similar jumps when comparing across top 1% income shares, and reductions when comparing across middle 40% shares. These figures reflect the significance of tax data in changing the distribution reported in survey data.

Figure 3A - Top 10%, Bottom 90% Pre-Tax Income Distribution, 2015

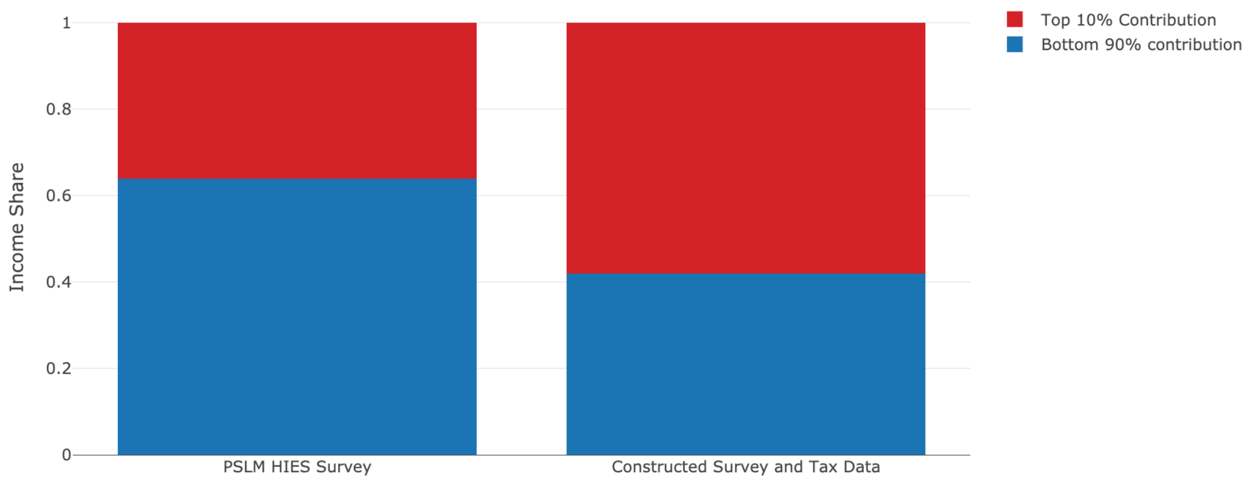


Figure 3B - Top 1%, Bottom 99% Pre-Tax Income Distribution, 2015

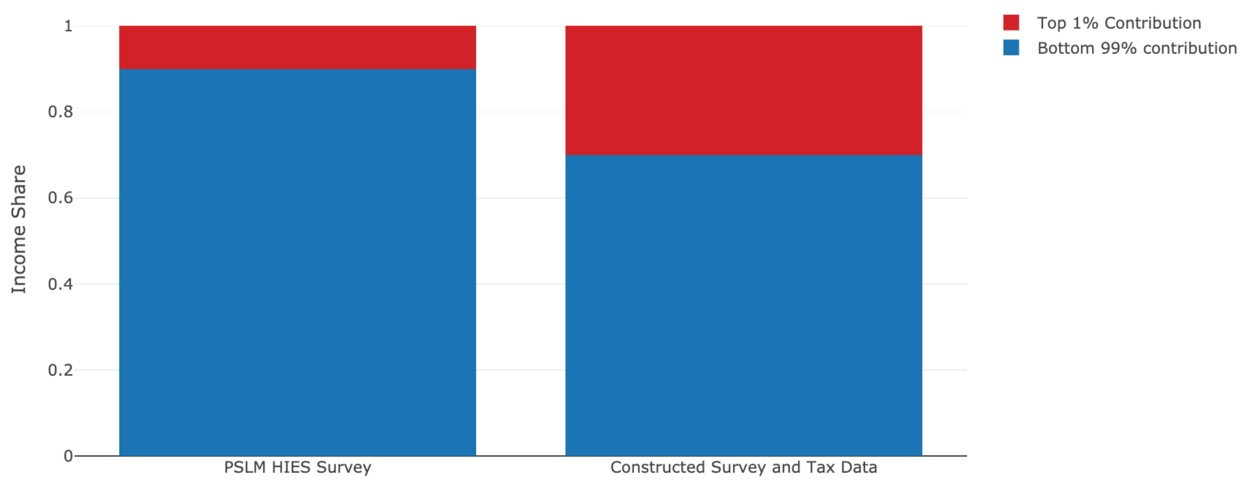
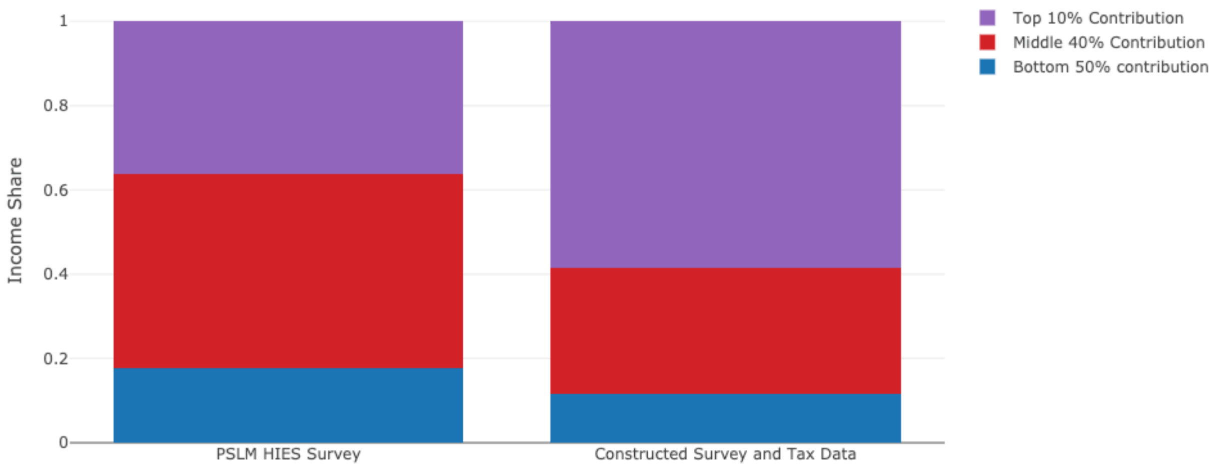


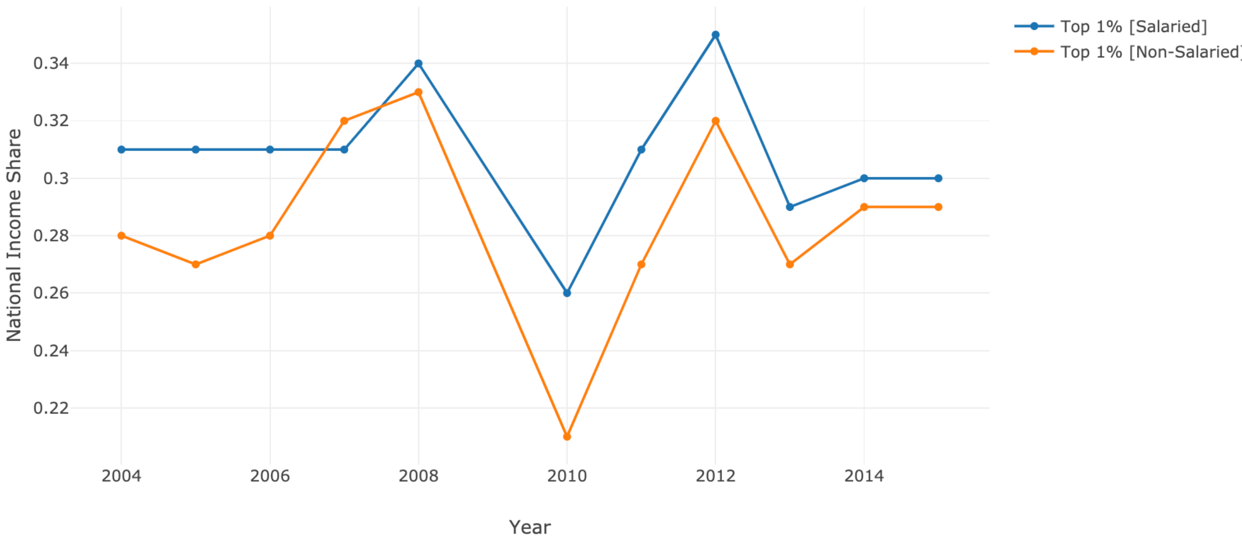
Figure 3C - Top 10%, Middle 40%, Bottom 50%, Pre-Tax Income Distribution, 2015



4.2 Top Income Shares

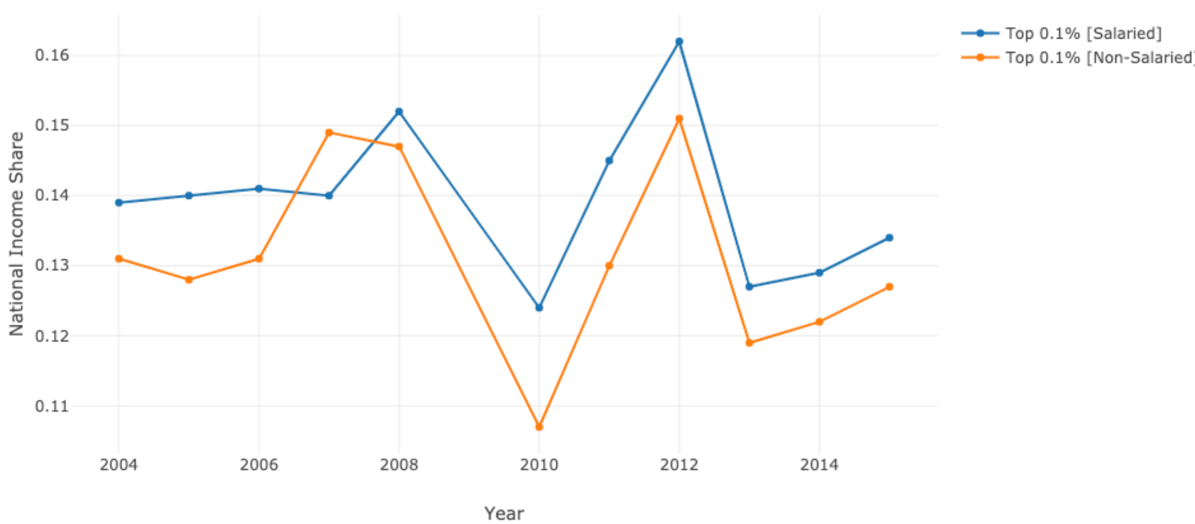
In examining the set of time series in figures 4a, 4b, 4c and 4d, it is again important to note that only years 2012 - 2015 are robust. The previous years are based off of 2012 tax data with estimated growth rates derived from survey data. We see a generally high level of income captured by the top 1%, 0.1% and 0.01% of the population. We do not have data for 2009, but we see some sort of decline during the Great Recession. Again, the magnitude should not be considered until after 2011. Note in the following figures we show pre-tax national income shares based on taxable income calculations for salaried and non-salaried earners. The true lower-bound lies somewhere between the two estimates.

Figure 4 - Top 1% Pre-Tax National Income Shares, Pakistan



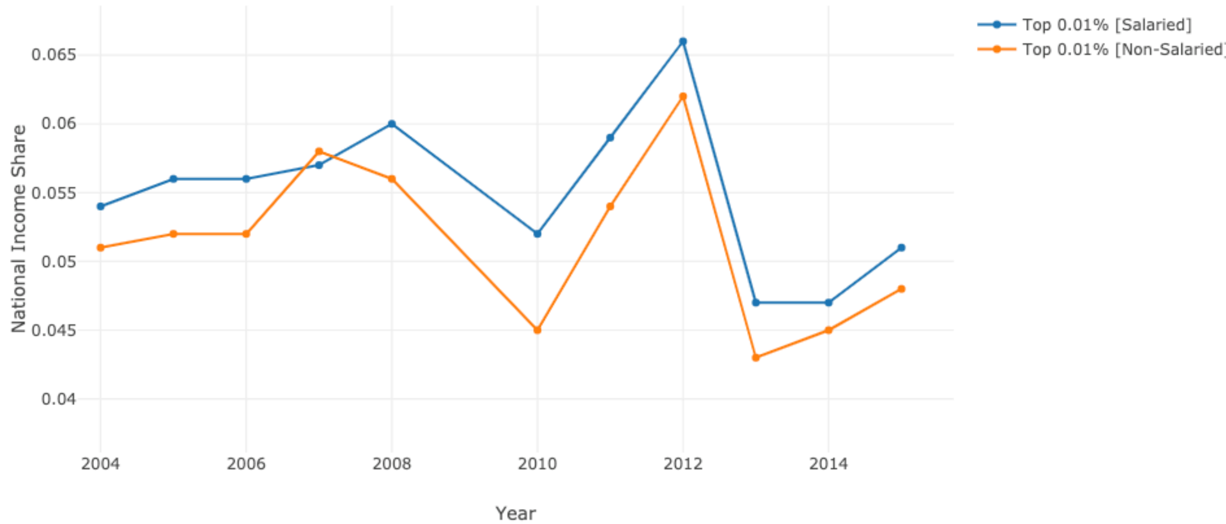
We see the share of income captured by the top 1% of the population reaching 30 - 35% of national income in recent years. If we assume most of the top 1% are salaried earners, this number is nearer the higher end. We put this number in perspective in section 4.4.

Figure 4B - Top 0.1% Pre-Tax National Income Shares, Pakistan



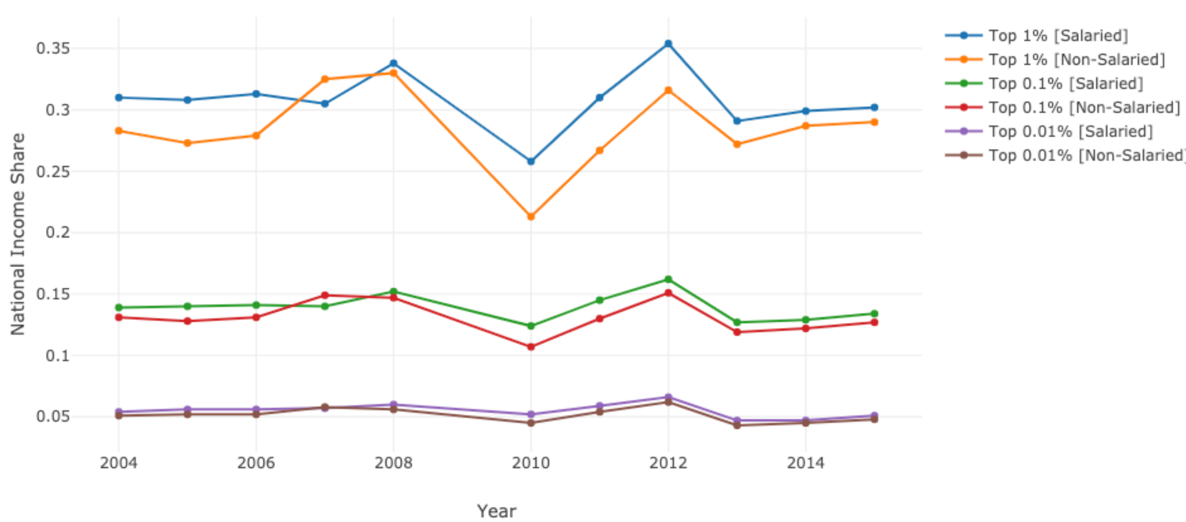
A surprisingly large amount of that income is captured by the top 0.1% of the population. Anywhere from 11 - 16% of national income goes to this population alone.

Figure 4C - Top 0.01% Pre-Tax National Income Shares, Pakistan



Finally, the top 0.01% of the population presents the least volatile estimates. This portion of the population presently captures anywhere from 4 - 6% of national income.

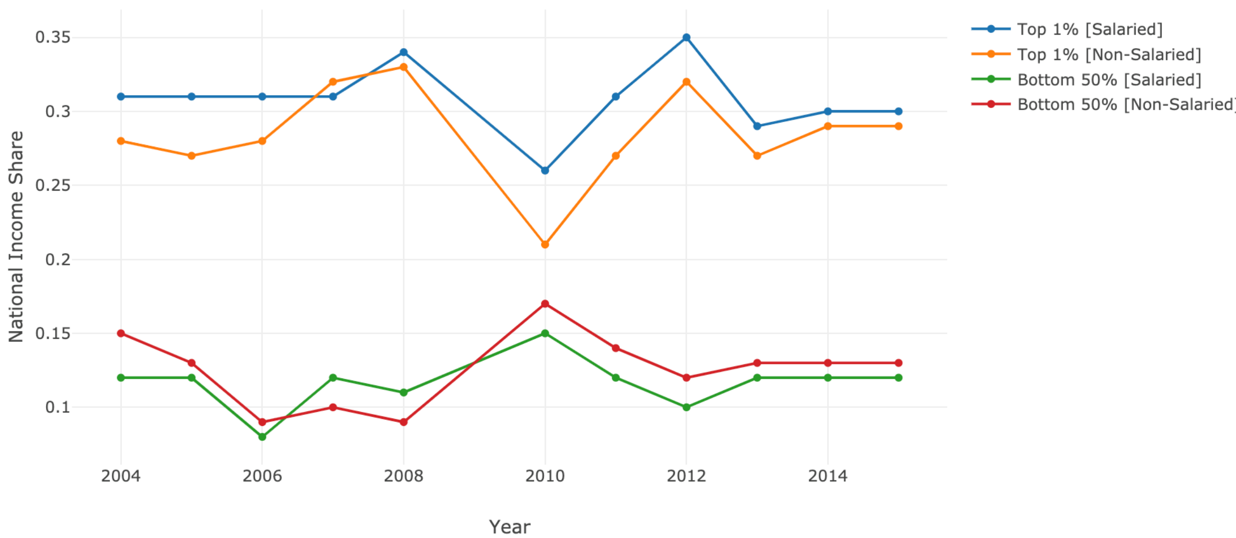
Figure 4D - Top 1%, Top 0.1%, Top 0.01% Pre-Tax National Income Shares, Pakistan



4.3 Bottom 50% and Middle 40% Income Shares

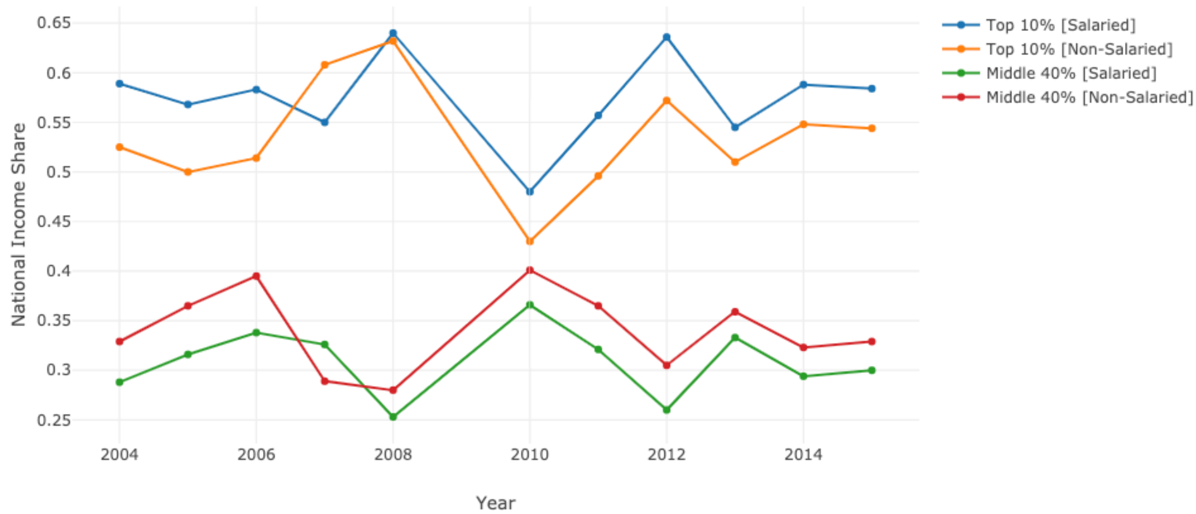
The income share held by the bottom 50% of the population is shockingly low. It ranges from 10 - 13% of national income. We can see the difference when plotting against the top 1% of the population in figure 5, which (as seen above) accounts for 30 - 35% of national income.

Figure 5 - Top 1% and Bottom 50% Pre-Tax National Income Shares, Pakistan



The middle 40% of the distribution holds nearly twice as much of the national income share as the bottom 50%, with around 25 - 35% of national income. The top 10% holds a much an overwhelming 51 - 64% of national income. We can see the gap in these two shares in figure 6. The middle 40% income share is similar to that of other developing countries. We also put these numbers in perspective in the following section.

Figure 6 - Top 10% and Middle 40% Pre-Tax National Income Shares, Pakistan



4.4 Comparing Pakistan with Similar Countries

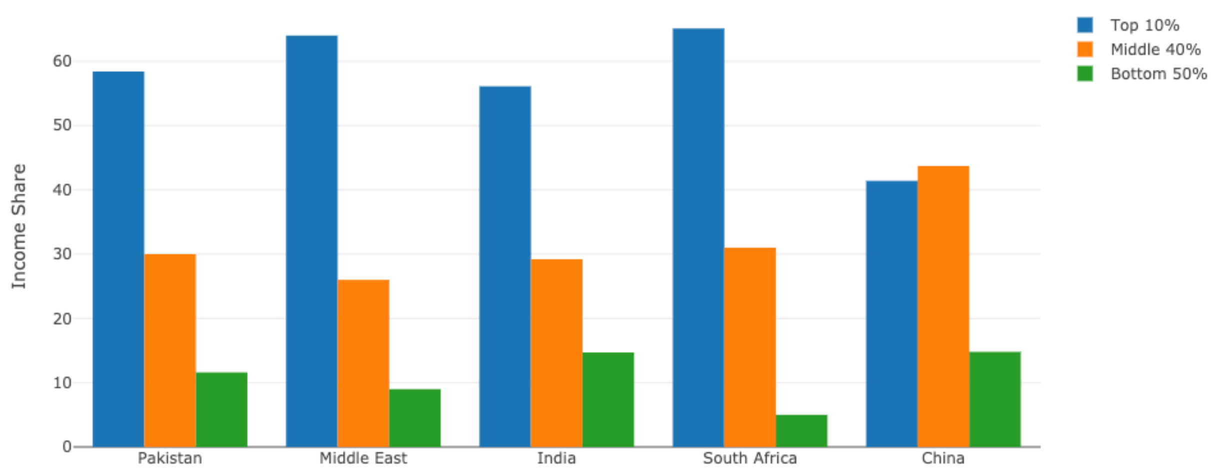
To put this data in context, we have provided similar statistics in table 6 that have been compiled by researchers at WID.world. We compare our findings, data sources and methodology to recent papers constructing distributional national accounts. This includes research by Chancel and Piketty (2017), Alvaredo, Assouad and Piketty (2017), Piketty, Yang and Zucman (2018), Piketty, Saez and Zucman (2018), and Assouad, Chancel and Morgan (2018). This research covers India, the Middle East, China, the United States, and South Africa and Brazil respectively. For Pakistani income inequality in table 6, we average the non-salaried and salaried income share projections.

Table 6 - Comparative Inequality, Pre-Tax National Income Shares 2015

	Pakistan	India	Middle East	South Africa	Brazil	Turkey	China	USA
Bottom 50%	11.6	14.7	9.0	5.0	13.9	14.9	14.8	12.5
Middle 40%	30.0	29.2	26.0	31.0	30.6	32.5	43.7	40.5
Top 10%	58.4	56.1	64.0	65.1	55.6	52.6	41.4	47
Top 1%	30.2	21.3	30.5	19.2	28.3	21.5	13.9	20.2
Top 0.1%	13.4	8.2	None	None	None	None	6.3	9.3
Top 0.01%	5.1	3.4	None	None	None	None	3.2	4.4

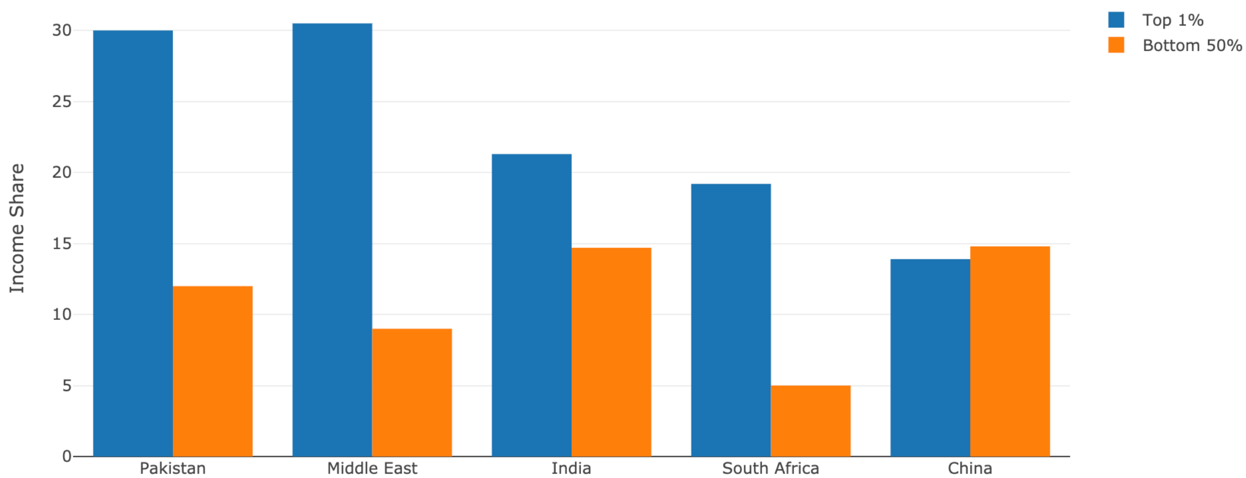
As compared to neighboring India, our results show similar levels of income share at the top 10% and middle 40% level. However, at the very top of the income distribution, nearer the top 1% and top 0.1% of the population, significantly more income is held by those portions of the population than in other countries. This reflects a more unequal society in Pakistan.

Figure 7A - Comparative Inequality: Top 10%, Middle 40%, Bottom 50% Pre-Tax Income Shares, 2015



When we compare differences in income shares across select countries, as in figure 7a, Pakistan seems to have a similar share of national income going to the middle 40%. The difference between top 10% and bottom 50% national income share is not as large as the Middle East or South Africa, but still considerable compared to India.

Figure 7B - Comparative Inequality: Top 1% and Bottom 50% Pre-Tax Income Shares, 2015

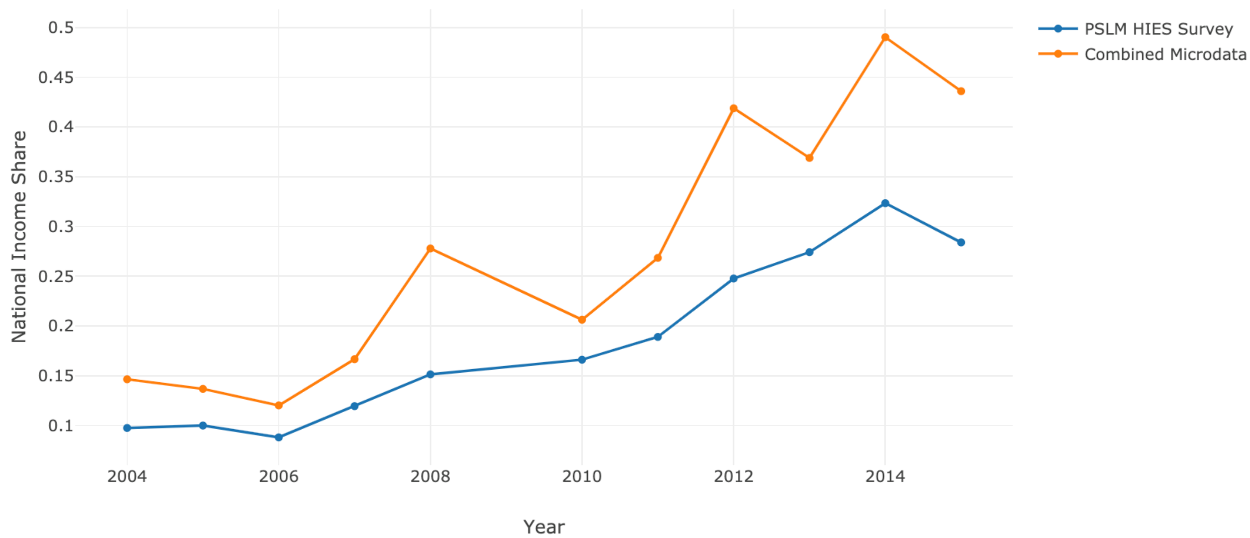


Within the top 1% of the distribution however, we can see that Pakistan is comparable to the Middle East. The bottom 50% has a similar proportion of income as compared to other countries.

4.5 Share of Income Gap Explained by Top Incomes

As discussed regarding national accounts data (section 3.1.3), there is a large gap between total survey incomes and national income reflected in figure 8. We compare our combined microdata with tax data corrections to the survey microdata. Our corrections account for an additional 11 - 17% of national income between 2012 - 2016. This sheds light on the contribution of top income shares to national income that is unaccounted for in the survey data. Although we can only be sure of results from 2012 - 2015, if top incomes grew by the same amount as bottom survey incomes, this share rises from just 4% in 2004.

Figure 8 - Share of National Income for Survey and Combined Microdata



4.6 Limitations

As mentioned in earlier sections, there are several measurement issues. We will recount a few key limitations here. All measurements from 2004 - 2011 suffer from the problem of estimating a growth rate for top incomes. We use the average survey growth rate in lieu of better data. Therefore, those numbers are not reliable, but should be treated as a ballpark estimate for income concentration. The distribution calculated from 2012 onwards is considered much more accurate as both tax and survey data exists.

Even then, most of the limitations lie around the tax data. For uncertain calculations, we consistently choose underestimates. Not all the manually filed tax returns have been included, leading to an underestimate of top incomes. In deriving taxable income from taxes paid, we assume the sample consists of either solely salaried or solely non-salaried workers, when in reality a mix would lie in between. If we assume taxes are paid from other sources of income (from the categorization explained in section 3.1.6, this could include capital gains or property), then their true taxable income is higher than our current estimates. We do not make any assumptions about deductions, meaning we assume taxable income is equal to personal income. This is certainly an underestimate.

Regarding the survey data, there is a large share of income that is unaccounted for when comparing to national income. Our combined microdata removes some of the gap, but the erratic growth rates across years suggests larger issues with the HIES survey data. Therefore, we cannot accurately estimate income growth rates across our sample. We assume the HIES survey data is correct for a given year, but not across years. Whether the survey data accurately captures income from the informal economy is questionable, although we believe the necessary steps have been taken to account for it in survey design.

Alternative weighting scenarios have not been considered when combining the tax and survey data. Our current equal weighting assumes that the sample of taxpayers in the tax directory is representative of all eligible taxpayers. Further research should explore these different scenarios and examine their effects on the full income distribution.

There are also alternative data sources that could improve or corroborate our estimation of top incomes that is left for future research. Land records are fully digitized across Pakistan's largest provinces. A cursory reading of summary statistics surrounding this data shows that land is highly concentrated among Pakistani elites. Exploring this concentration and estimating income earned from rents would provide a valuable addition to this study. We encourage research to explore that dataset and add to our findings.

5 Conclusion

We construct an account of Pakistani income inequality using household surveys and newly released tax data to produce better estimates of the full distribution of pre-tax income between 2012 - 2015. We find a level of income inequality similar to that of the Middle East and slightly more than neighboring India. Our combined dataset provides a significant correction to household survey data that is usually used in research regarding Pakistani income inequality. In addition, our corrections explain a non-trivial amount of the discrepancy between combined household survey income and national income derived from national accounts data. We find our results are robust from 2012 - 2015 given we use lower-bound assumptions in all of our calculations. We also attempt to estimate income shares from 2004 - 2011 using survey growth rates on 2012 tax data, and although we can provide a general estimate, we cannot provide robust findings for income distribution during this time period.

According to our estimates for 2015, the top 1% income share was at 30.2%, on par with the Middle East and Brazil. The top 0.1% share was at 13.4% and top 0.01% share at 5.1%. The middle 40% share is at 30.0%, a statistic comparable to similar developing countries. The bottom 50% share is at 11.6%, more than South Africa and the Middle East, less than India and China, and on par with the United States. We stress the need for more accurate data on incomes from the tax and statistics authorities of Pakistan to produce better results and a time series of income distribution. We cannot construct a complete system of distributional national accounts without the cooperation and expertise of the official tax and statistical agencies of Pakistan. With this data, we can construct a more accurate and long-term series of accounts that can in turn lead to a better understanding of the impact of redistributive policies and broader economic growth in Pakistan. We can also better understand the role informal economy by closing the gap between estimates of income derived from micro and macro data.

Our findings provide a snapshot rather than a longer-term account. Two key contributions of this paper are the creation of the new tax dataset, which has been converted and parsed from a PDF directory into usable microdata, and our combined final microdata. We hope that this research and new data (1) informs the debate on Pakistani income inequality and (2) provides a starting point for further research.

References

Ahmad, M., & Shirazi, N. S. (2000). Estimation of Distribution of Income in Pakistan, Using Micro Data [with Comments]. *The Pakistan Development Review*, 807-824

Alvaredo, F., Assouad, L., & Piketty, T. (2017). DP12405 Measuring inequality in the Middle East 1990-2016: The World's Most Unequal Region?

Alvaredo, F., Atkinson, A. B., Chancel, L., Piketty, T., Saez, E., & Zucman, G. (2016). Distributional National Accounts Guidelines: Conceps and Methods used on WID.world. WID.world Working Paper 2016/1

Arby, Muhammad Farooq. (2008). Some issues in the national income accounts of Pakistan (rebasing, quarterly and provincial accounts and growth accounting).

Arby, Muhammad Farooq, Muhammad Jahanzeb Malik, and Muhammad Nadim Hanif. "The size of informal economy in Pakistan." (2010).

Assouad, L., Chancel, L., & Morgan, M. (2018). Extreme Inequality: Evidence from Brazil, India, the Middle East, and South Africa. In *AEA Papers and Proceedings* (Vol. 108, pp. 119-23).

Atkinson, A. B., & Piketty, T. (Eds.). (2007). Top incomes over the twentieth century: a contrast between continental european and english-speaking countries. OUP Oxford.

Atkinson, A. B., & Piketty, T. (Eds.). (2010). Top incomes: A global perspective. Oxford University Press.

Atria, J., Flores, I., Mayer, R., & Sanhueza, C. (2018). Top Incomes in Chile: A Historical Perspective of Income Inequality (1964-2015). *URL: http://ignacioflores. com/pdf/top-incomes-chile.pdf.*

Banerjee, A., & Piketty, T. (2005). Top Indian Incomes, 1922–2000. The World Bank Economic Review, 19(1), 1-20.

Bergan, Asbjorn (1967) Personal Income Distribution and Personal Savings Pakistan, 1963- 64. *The Pakistan Development Review* 7:2, 160-212.

Burkhauser, R. V., Hérault, N., Jenkins, S. P., & Wilkins, R. (2016). What has been happening to UK income inequality since the mid-1990s? Answers from reconciled and combined household survey and tax return data (No. w21991). National Bureau of Economic Research.

Chancel, L., & Piketty, T. (2017). Indian income inequality, 1922-2014: From British Raj to Billionaire Raj?

Choudhary, M. K. (1995) National and Provincial Size Distribution of Income in Pakistan. *Journal of Economics* 2, Sep.

de Kruijk, Hans, (1986) Inequality in the Four Provinces of Pakistan. *The Pakistan Development Review* 25:4, 685-706.

de Kruijk, Hans, and M. Van Leeuwen (1985) Changes in Poverty and Income Inequality in Pakistan During 1970s. *The Pakistan Development Review* 24:3-4, 407-422.

Ercelawn, A. (1988) Income Inequality in Pakistan During the 70s: Issues in Estimation. Applied Economic Research Centre, Karachi. (Discussion Paper No.92)

Haq, Khadija (1964) A Measurement of Inequality in Urban Personal Income Distribution in Pakistan. *The Pakistan Development Review* 4:4.

Jafari, S. Y., and Azizullah Khattak (1995) Income Inequality and Poverty in Pakistan. *Pakistan Economic and Social Review* 33:1&2.

Jeetun, A. (1978). *Trends in inequality of income distribution in Pakistan*. Applied Economics Research Centre, University of Karachi.

Kemal, A. R. (1981) Income Distribution in Pakistan: A Review. Pakistan Institute of Development Economics, Islamabad. (Research Reports Series No. 123)

Kemal, A. R. (1994) Structural Adjustment, Employment, Income Distribution and Poverty. *The Pakistan Development Review* 33:4, 901-911.

Kuznets, S., & Jenks, E. (1953). *Shares of upper income groups in income and savings*. National Bureau of Economic Research, Inc.

Milanovic, B., & Lakner, C. (2015). Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession. World Bank Economic Review.

Piketty, T. (2001). Les hauts revenus en France au XXème siècle. Grasset.

Piketty, T., & Saez, E. (2003). Income inequality in the United States, 1913–1998. The Quarterly journal of economics, 118(1), 1-41.

Piketty, T., E. Saez, & G. Zucman (2018). "Distributional National Accounts: Methods and Estimates for the United States", Quarterly Journal of Economics, 133(2): 553-609.

Piketty, T. Yang, L., Zucman, G. (2017). Capital Accumulation, Private Property and Rising Inequality in China, 1978-2015. NBER working paper 23368, June 2017.

Seiver, D. A. (1979) A Note on the Measurement of Income Inequality with Interval Data. *Review of Income and Wealth* 25:2, June 1979.

Table 1A
Distribution of Pre-Tax National Income in Pakistan using Combined Microdata, 2015

	Labor Force	Average Annual Income (PKR)	Average Annual Income (USD)	Income Share
Full Population	65,422,753	279,601	$1,985	100%
Bottom 50%	32,711,376	64,927	$460	11.6%
Middle 40%	26,169,101	209,513	$1,487	30.0%
Top 10%	6,542,275	1,633,230	$11,595	58.4%
Top 1%	654,227	8,437,088	$59,903	30.2%
Top 0.1%	65,422	37,474,289	$266,067	13.4%
Top 0.01%	6,542	142,210,514	$1,009,694	5.1%

Table 1B
Distribution of Pre-Tax National Income in Pakistan using HIES Survey Data, 2015

	Labor Force	Average Annual Income (PKR)	Average Annual Income (USD)	Income Share
Full Population	65,422,753	182,103	$1,292	100%
Bottom 50%	32,711,376	64,930	$461	17.8%
Middle 40%	26,169,101	209,508	$1,487	46.0%
Top 10%	6,542,275	658,201	$4,673	36.2%
Top 1%	654,227	1,761,974	$12,510	9.7%
Top 0.1%	65,422	4,054,818	$28,789	2.3%
Top 0.01%	6,542	8,128,723	$57,713	0.5%

Table 2 - Number of TD Tax Filers and TD Taxpayers

Fiscal Year	Labor Force	Eligible Taxpayer Population	Number of TD Tax Filers	Number of TD Taxpayers
2013	60,357,706	6,029,151	728,346	422,744
2014	62,025,233	3,788,656	780,462	494,375
2015	64,070,380	5,097,884	1,000,718	692,360
2016	66,149,979	7,800,316	1,135,764	803,412
2017	67,602,890	8,788,375	1,680,404	1,087,486

Table 3 - FBR Tax Return Summary Statistics

Fiscal Year	Labor Force	TD Tax Filers (% of Labor Force)	TD Taxpayers (% of Labor Force)	TD Taxpayers (% of TD Tax Filers)
2013	60,357,706	1.21%	0.7%	58.0%
2014	62,025,233	1.26%	0.8%	63.3%
2015	64,070,380	1.56%	1.08%	69.2%
2016	66,149,979	1.72%	1.21%	70.7%
2017	67,602,890	2.49%	1.61%	64.7%

Table 4 - FBR Tax Directory Sample in Context

Fiscal Year	Eligible Taxpayer Population	TD Taxpayer Sample (% of Eligible Taxpayer Population)
2013	6,029,151	7.01%
2014	3,788,656	13.05%
2015	5,097,884	13.58%
2016	7,800,316	10.3%
2017	8,788,375	12.37%

Table 5
Income Growth Rates (Survey and National Income)

Year	Taxpayer Income Growth Rate (Salaried)	Taxpayer Income Growth Rate (Non-Salaried)	Average Income Growth Rate (Survey)	Average Income Growth Rate (National Income)
2005	-1.55%	-1.55%	9.86%	7.16%
2006	2.83%	2.83%	-7.44%	5.05%
2007	1.39%	50.32%	39.64%	2.93%
2008	69.58%	55.19%	29.57%	2.23%
2010	-37.4%	-53.86%	15.4%	5.15%
2011	50.15%	50.15%	17.57%	3.31%
2012	58.7%	58.7%	36.17%	3.91%
2013	-10.59%	-10.59%	15.49%	4.34%
2014	53.07%	53.07%	23.46%	4.64%
2015	-17.46%	-17.46%	-7.86%	4.95%

Table 6
Comparative Inequality, Pre-Tax National Income Shares 2015

	Pakistan	India	Middle East	South Africa	Brazil	Turkey	China	USA
Bottom 50%	11.6	14.7	9.0	5.0	13.9	14.9	14.8	12.5
Middle 40%	30.0	29.2	26.0	31.0	30.6	32.5	43.7	40.5
Top 10%	58.4	56.1	64.0	65.1	55.6	52.6	41.4	47
Top 1%	30.2	21.3	30.5	19.2	28.3	21.5	13.9	20.2
Top 0.1%	13.4	8.2	None	None	None	None	6.3	9.3
Top 0.01%	5.1	3.4	None	None	None	None	3.2	4.4

Figure 1
Personal Income Tax Revenue (% of GDP)

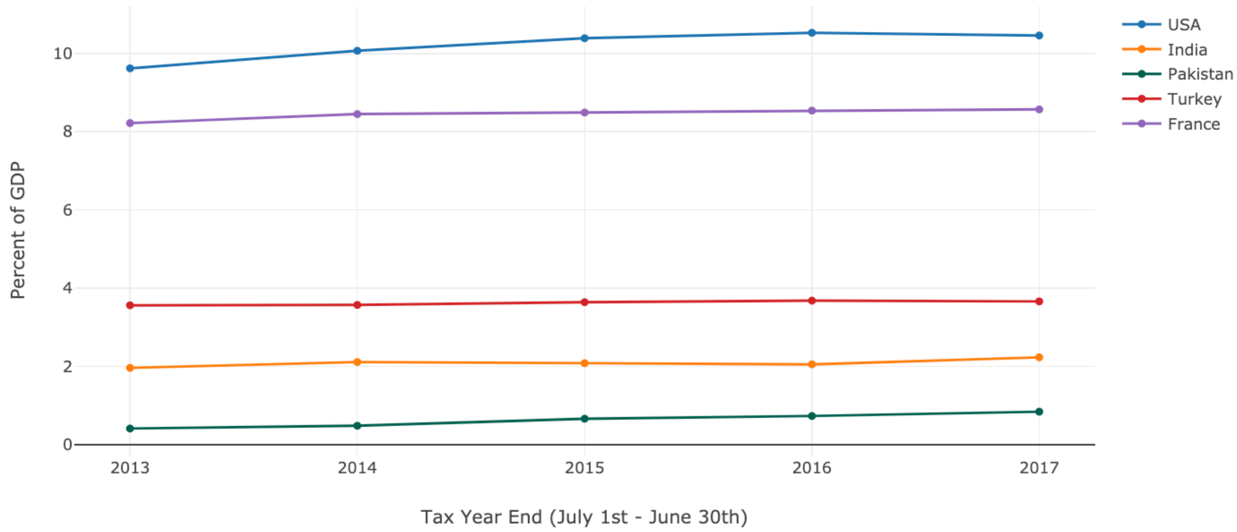


Figure 2
Personal Income Tax Revenue Growth, India and Pakistan (% of GDP)

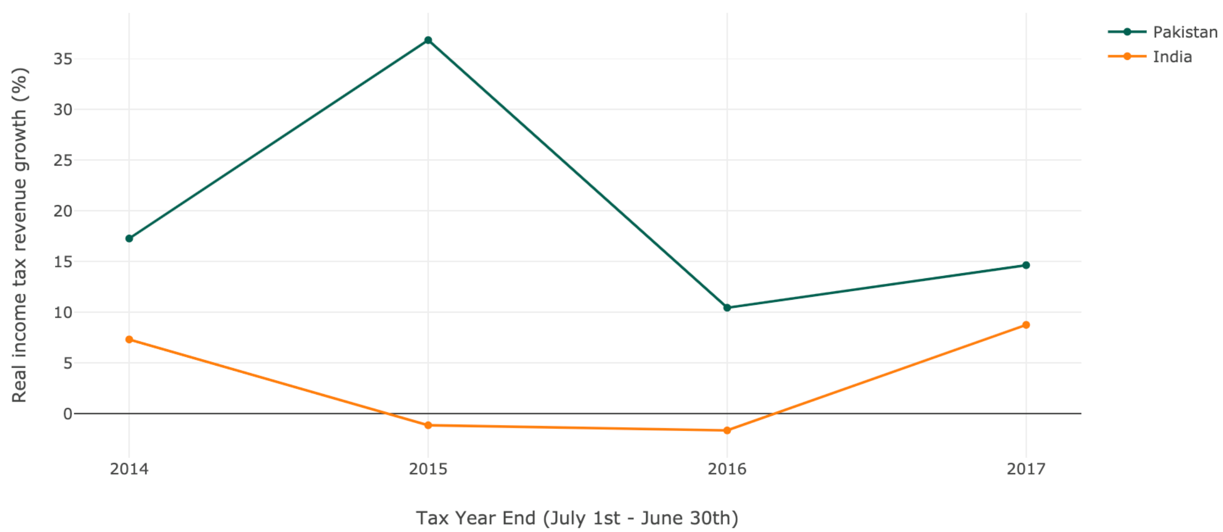


Figure 3A
Top 10%, Bottom 90%, Pre-Tax Income Distribution, 2015

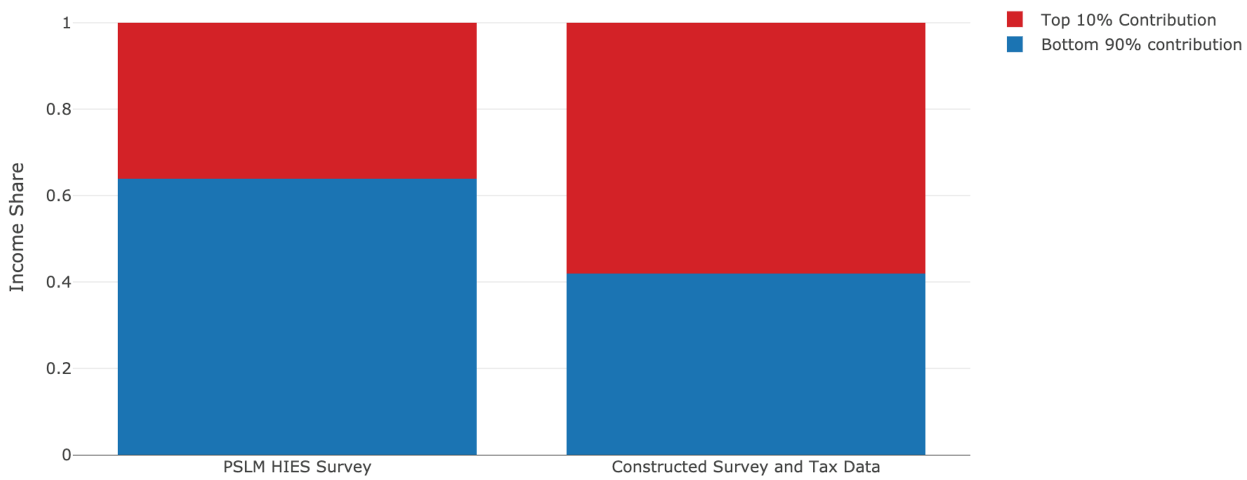


Figure 3B
Top 1%, Bottom 99%, Pre-Tax Income Distribution, 2015

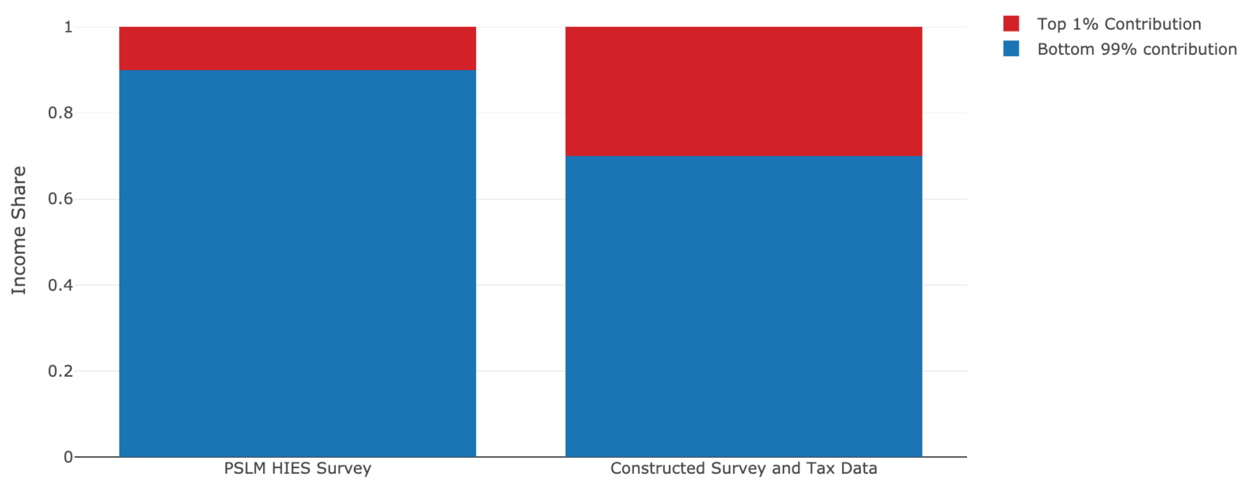


Figure 3C
Top 10%, Middle 40%, Bottom 50%, Pre-Tax Income Distribution, 2015

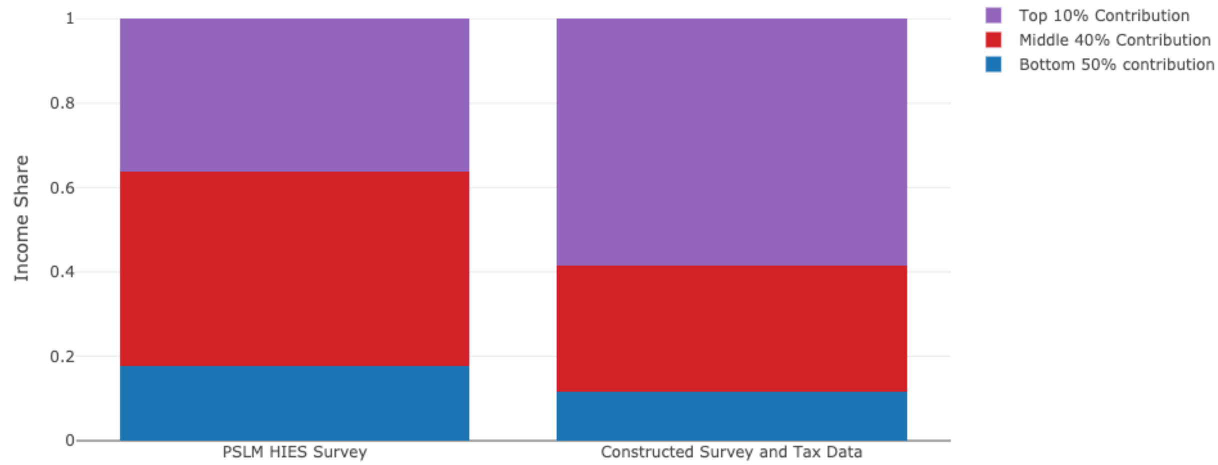


Figure 4A
Top 1% Pre-Tax National Income Shares, Pakistan

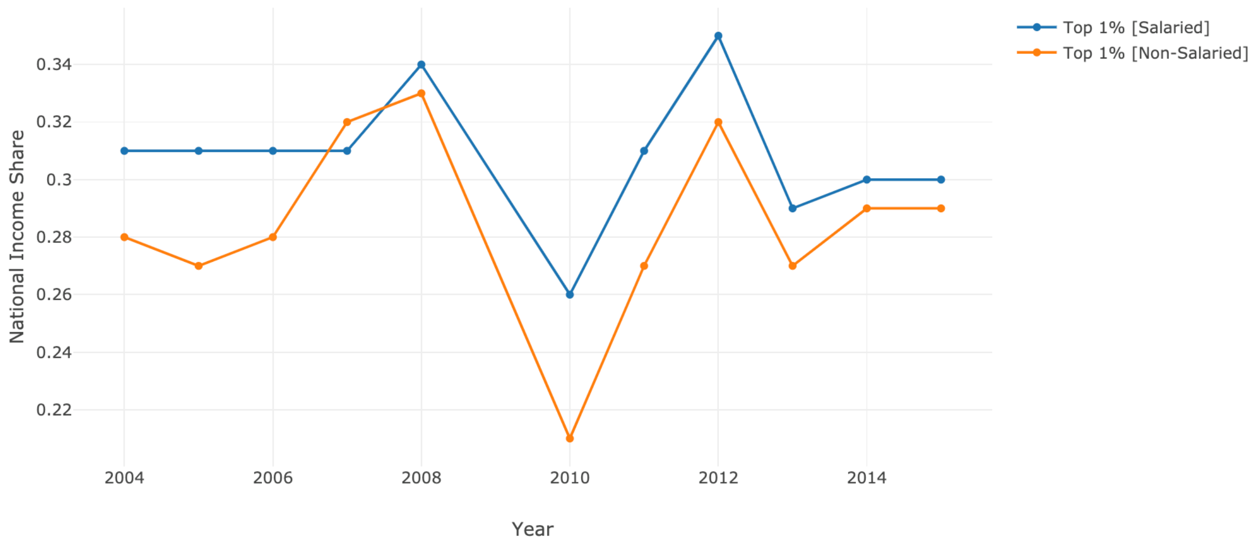


Figure 4B
Top 0.1% Pre-Tax National Income Shares, Pakistan

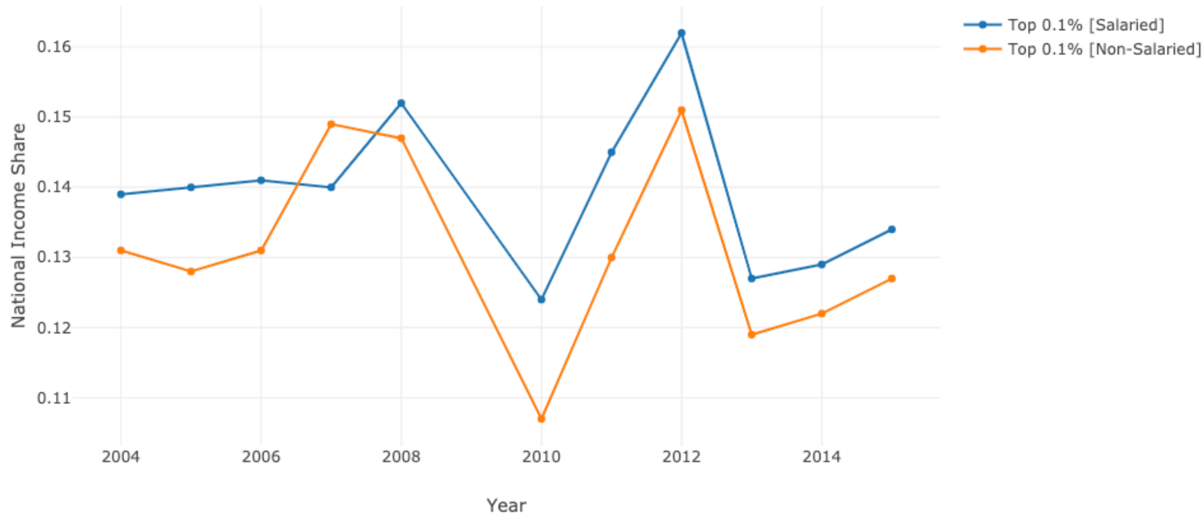


Figure 4C
Top 0.01% Pre-Tax National Income Shares, Pakistan

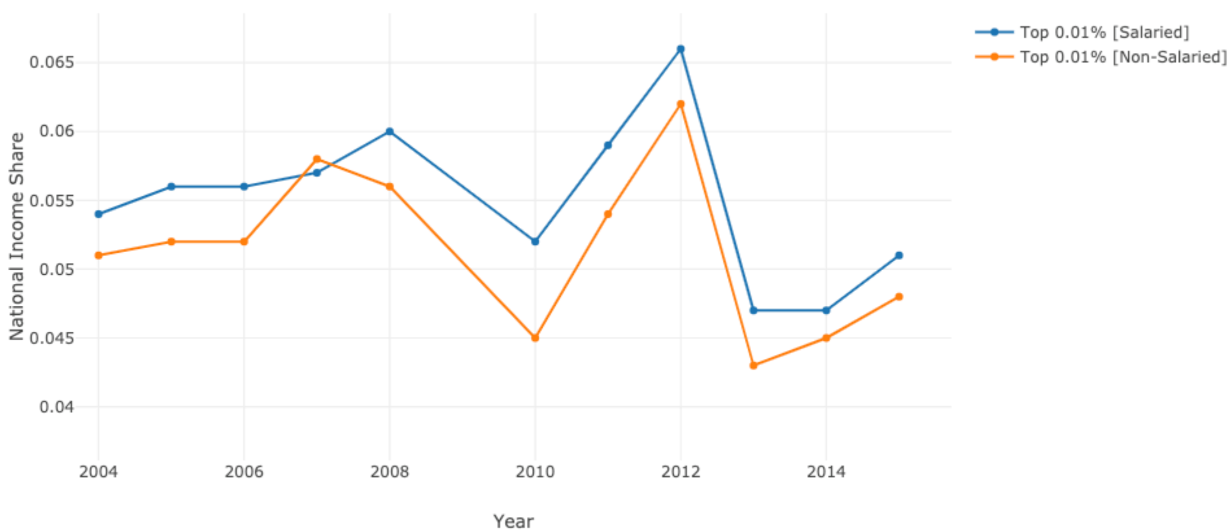


Figure 4D
Top 1%, Top 0.1%, Top 0.01% Pre-Tax National Income Shares, Pakistan

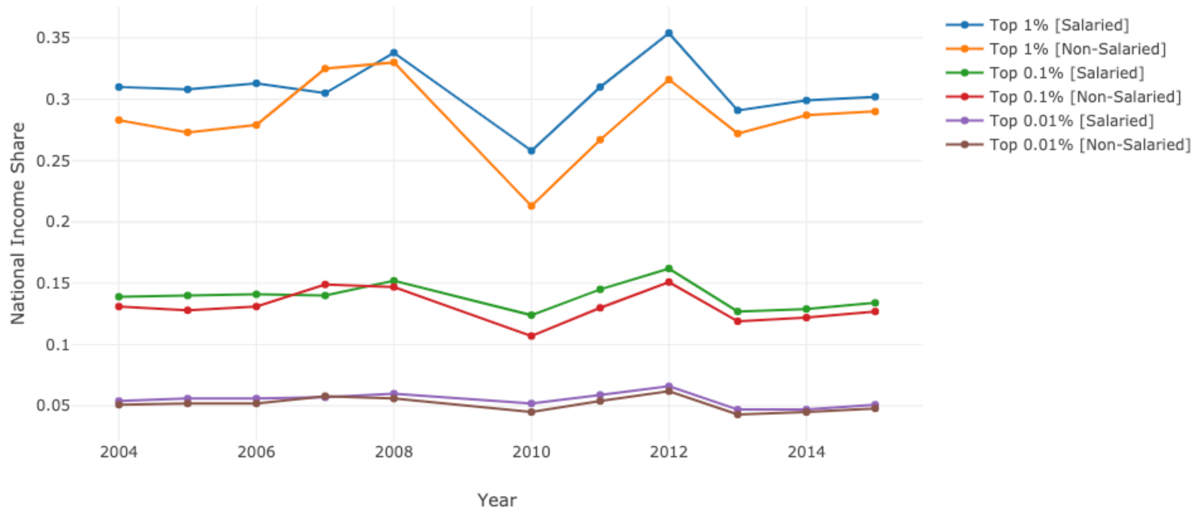


Figure 5
Top 1% and Bottom 50% Pre-Tax National Income Shares, Pakistan

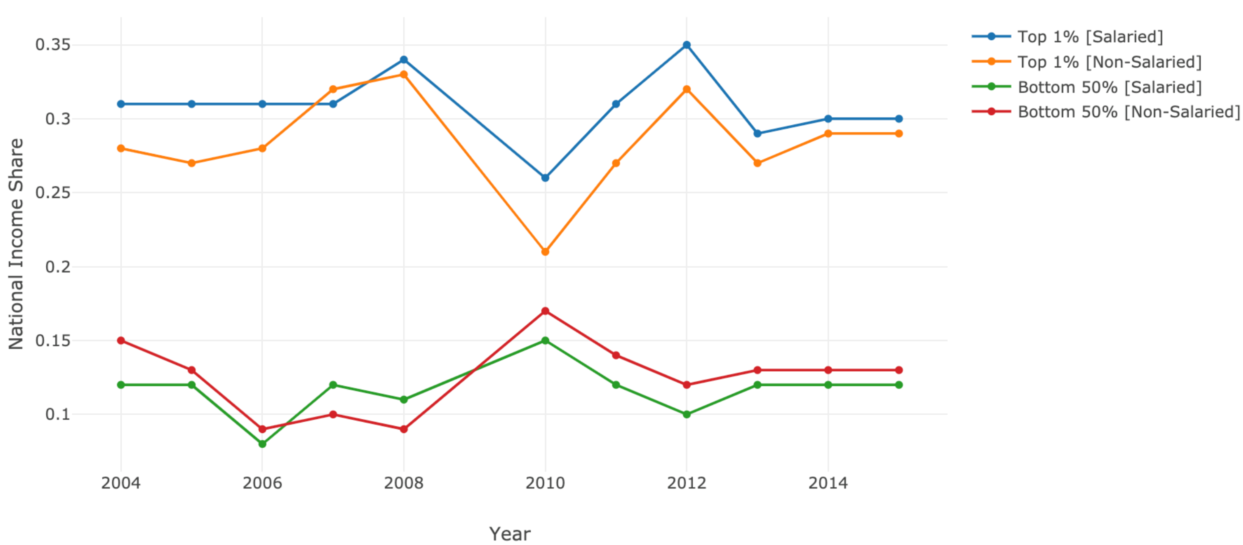


Figure 6
Top 10% and Middle 40% Pre-Tax National Income Shares, Pakistan

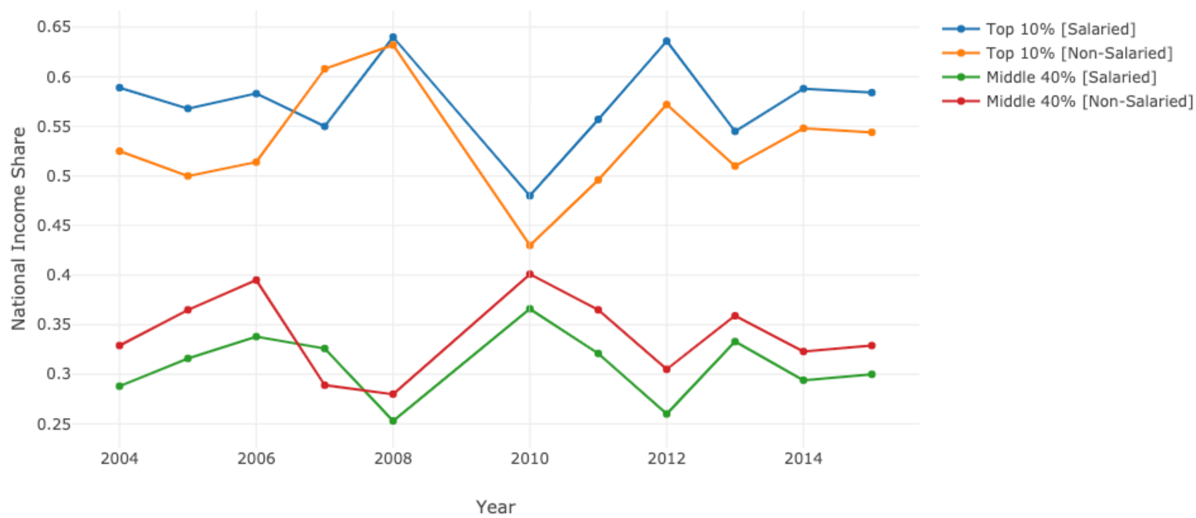


Figure 7A
Comparative Inequality: Top 10%, Middle 40%, Bottom 50% Pre-Tax Income Shares, 2015

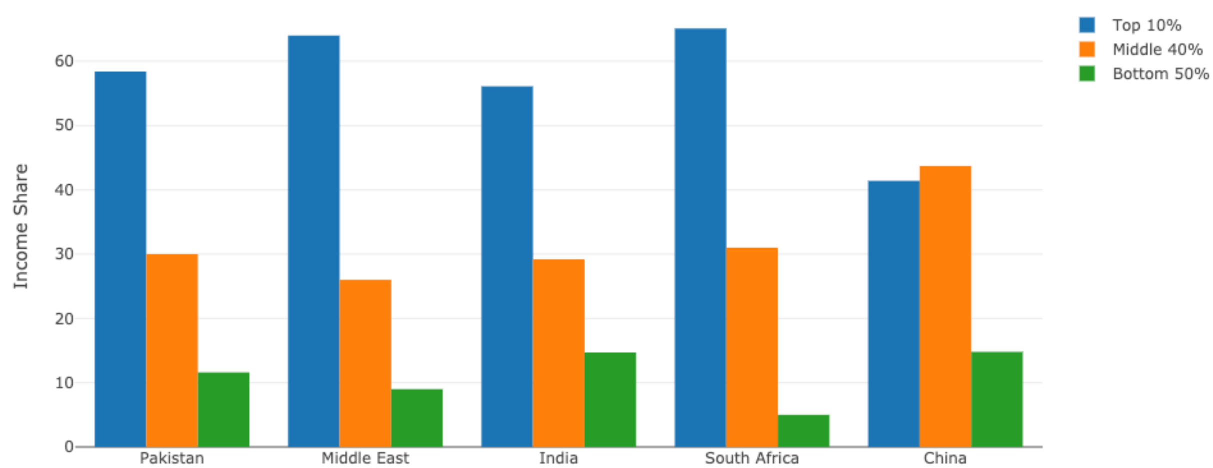


Figure 7B
Comparative Inequality: Top 1% and Bottom 50% Pre-Tax Income Shares, 2015

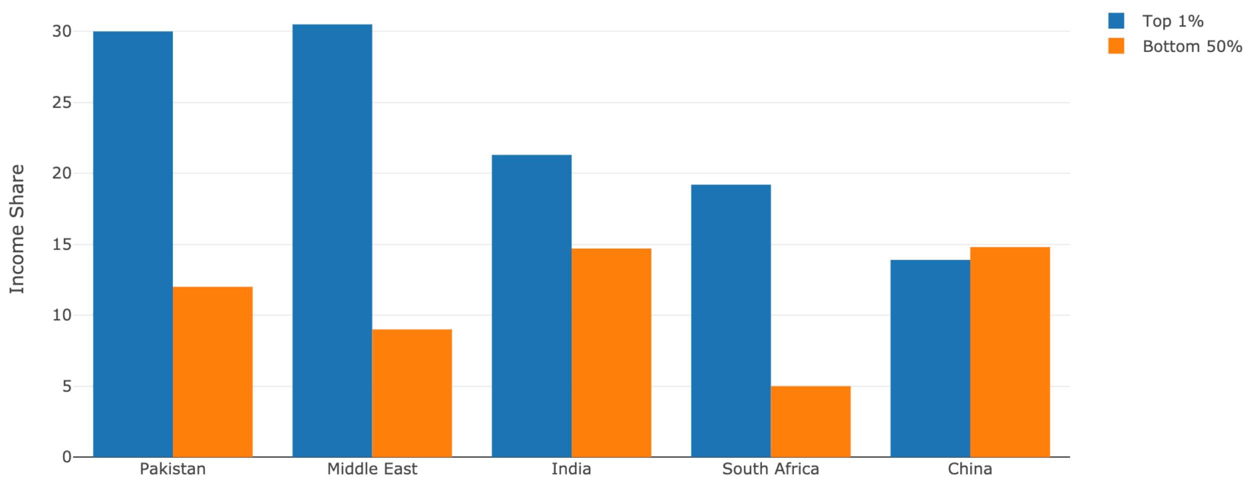
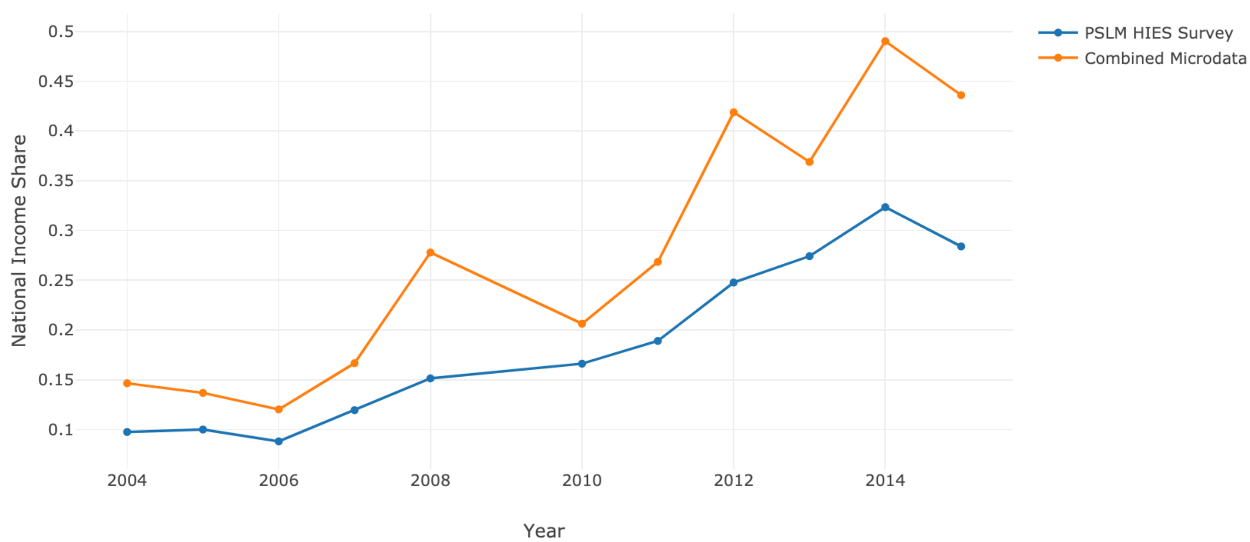


Figure 8
Share of National Income for Survey and Combined Microdata



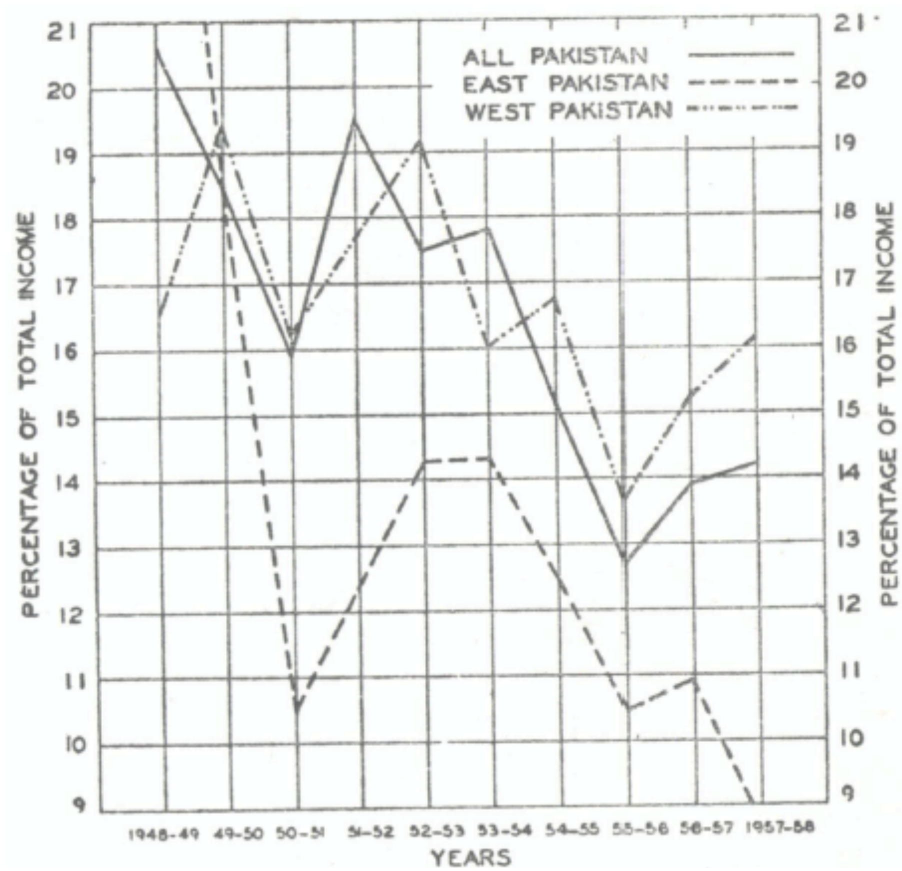
Appendix 1A
Patterns of Pre-Tax Income Distribution in East and West Pakistan, Haq (1964)

Year	Region	Top 1%	Top 10%	Top 20%	Bottom 50%
1950-51	East Pakistan	10.5	43.1	58.0	19.2
	West Pakistan	16.2	44.3	58.4	18.9
1957-58	East Pakistan	9.6	34.5	47.4	24.5
	West Pakistan	16.1	41.8	55.4	21.8

Note that West Pakistan is now present-day Pakistan. East Pakistan is now present-day Bangladesh.

Source: Haq, K. (1964). A measurement of inequality in urban personal income distribution in Pakistan. *The Pakistan Development Review*, 4(4), 623-664.

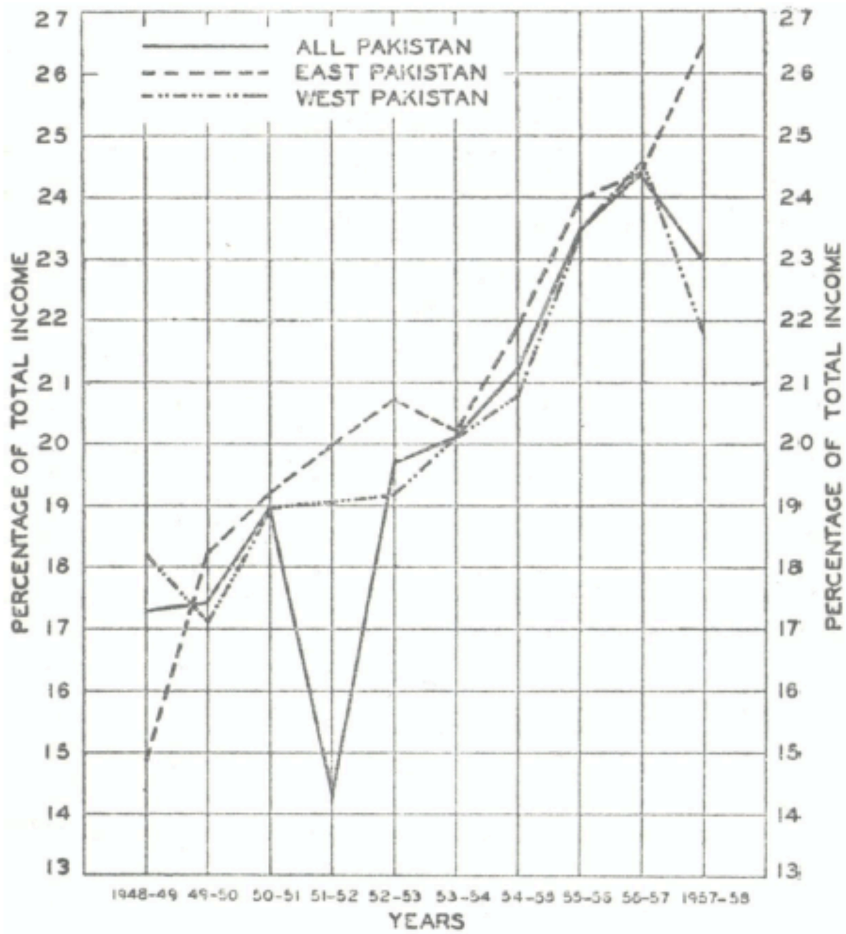
Appendix 1B
Top 1% Pre-Tax Income Share, 1948 - 1958



Note that West Pakistan is now present-day Pakistan. East Pakistan is now present-day Bangladesh.

Source: Haq, K. (1964). A measurement of inequality in urban personal income distribution in Pakistan. *The Pakistan Development Review*, 4(4), 623-664.

Appendix 1C
Bottom 50% Pre-Tax Income Share, 1948 - 1958



Note that West Pakistan is now present-day Pakistan. East Pakistan is now present-day Bangladesh.

Source: Haq, K. (1964). A measurement of inequality in urban personal income distribution in Pakistan. The Pakistan Development Review, 4(4), 623-664.

Appendix 2
PSLM HIES Household Survey Sampling Size

Year	Number of Households	Number of Individuals	Number of Households	Ratio
2004	14,707	21,170	14,707	1.44
2005	15,450	25,303	15,450	1.64
2006	70,156	108,427	70,156	1.55
2007	15,511	24,900	15,511	1.61
2008	71,491	108,698	71,491	1.52
2010	16,341	25,202	16,341	1.54
2011	15,807	25,362	15,807	1.60
2012	75,333	117,639	75,333	1.56
2013	17,891	29,321	17,891	1.64
2014	78,505	122,664	78,505	1.56
2015	24,238	38,552	24,238	1.59

Appendix 3
Junction Percentiles

	Year	Percentile	Tax Exemption Limit (PKR)
0	2004	80.3	80,000
1	2005	85.74	100,000
2	2006	85.47	100,000
3	2007	89.65, 78.28*	150,000, 100,000*
4	2008	83.61, 68.43*	150,000, 100,000*
5	2010	93.81	300,000
6	2011	90.25	300,000
7	2012	90.01	350,000
8	2013	93.89	400,000
9	2014	92.04	400,000
10	2015	88.21	400,000