Open Data and Open Science : Insights from the World Inequality Lab

Thomas Piketty[a], Anmol Somanchi[b]

This paper examines the significance of open data and open science through the experiences of the World Inequality Lab in developing, maintaining, and disseminating the World Inequality Database. We emphasize the central role of reliable public statistics in facilitating rigorous analysis of income and wealth distributions globally. Despite some improvements in recent years with regards to availability of inequality statistics, significant disparities across countries remain, as demonstrated by the Inequality Transparency Index. We argue that a robust commitment by governments to systematically collect and openly disseminate detailed fiscal and socio-economic data is essential not just to tackle inequality but for evidence-based policy debates more broadly. Ultimately, access to reliable socio-economic data is not just a technical requirement but a democratic imperative that empowers societies to better understand—and in turn effectively tackle—the pressing challenges of our times.

Données ouvertes et science ouverte : Perspectives du Laboratoire mondial sur les inégalités

Cet article examine l'importance des données ouvertes et de la science ouverte à travers l'expérience du World Inequality Lab dans le développement, la maintenance et la diffusion de la base de données mondiale sur les inégalités. Nous soulignons le rôle central de statistiques publiques fiables pour faciliter une analyse rigoureuse de la répartition des revenus et des richesses à l'échelle mondiale. Malgré quelques améliorations ces dernières années en matière de disponibilité des statistiques sur les inégalités, d'importantes disparités persistent entre les pays, comme le montre l'Indice de transparence des inégalités. Nous soutenons qu'un engagement ferme des gouvernements à collecter systématiquement et à diffuser ouvertement des données budgétaires et socio-économiques détaillées est essentiel, non seulement pour lutter contre les inégalités, mais aussi pour des débats politiques plus larges et fondés sur des données probantes. En fin de compte, l'accès à des données socio-économiques fiables n'est pas seulement une exigence technique, mais un impératif démocratique qui permet aux sociétés de mieux comprendre – et donc de relever efficacement – les défis urgents de notre époque.

Keywords: open data, inequality, public statistics, fiscal data, national income accounts

[a] Paris School of Economics and World Inequality Lab. *Correspondance :* Paris School of Economics, 48 Bd Jourdan, 75014 Paris, France.*E-mail :* thomas.piketty@psemail.eu

[b] Paris School of Economics and World Inequality Lab. *Correspondance :* Paris School of Economics, 48 Bd Jourdan, 75014 Paris, France.*E-mail :* anmol.somanchi@psemail.eu

Mots clés : données ouvertes, inégalités, statistiques publiques, données fiscales, comptes nationaux

Classification *JEL* : C81, C82, E01.

1. INTRODUCTION

We would like to begin by congratulating the *Revue économique* on its 75th anniversary. We wholeheartedly welcome the decision of the journal to move to an open-access model and are glad to contribute to this Special Issue on "disciplinary openness and open science" celebrating this transition. This brief contribution aims to reflect on some aspects of the debate surrounding open data and open science, based on the work at the World Inequality Lab (WIL) relating to the creation, management, and dissemination of the World Inequality Database (WID).

The last two decades brought with them a renewed interest in distributional concerns in the economics discipline. Starting with long-run studies on France (Piketty [2001]) and the United States (Piketty and Saez [2003]) at the turn of the century, a rich and vibrant literature investigating historical inequality trends flourished. Within a few years, a motley group of researchers were able to put together a collective volume that used hitherto unutilized tax data to study dynamics of top incomes over the twentieth century in the American and Anglo-Saxon world (Atkinson and Piketty [2007]). This collective project was subsequently expanded to include a broader canvas of countries, both in Europe but also India, China and a few other Asian economies (Atkinson and Piketty [2010]). It was at this stage that the need was felt for collating all the data series in one place and making them openly and freely accessible to researchers, citizens, and policymakers.

With contributions from multiple researchers spread across the globe, the World Top Incomes Database (WTID) was created and launched in January 2011. Over the next decade and a half, the scope and ambition expanded to include not only income, but also wealth, and more recently gender gaps and carbon emissions. The focus also shifted from top incomes and wealth only to studying the *full distributions*, bottom to top. As a growing number of middle- and lower-income economies were added to the database, it also became clear that tax data, which tends to cover a small fraction of the population in these countries, would not suffice. It had to be creatively combined with household surveys and national accounts. To this effect, the Distributional National Accounts (DINA) guidelines were drafted (Alvaredo, Atkinson, *et al.* [2016]) and a novel method for efficiently combining tax and survey data was developed (Blanchet, Flores, and Morgan [2022]). Today, thanks to the spirited contributions from over 200 "WID Fellows"

Acknowledgements : We are grateful to Ignacio Flores and David Margolis for helpful discussions. Anmol Somanchi acknowledges financial support from the French National Agency for Research (ANR-17-EUR-0001).

from all over the world, the WID includes full distributions (127 generalized percentiles) of income and wealth for over 100 countries estimated based on the consistent, comparable, and harmonized DINA framework. Very recently, the database was also updated to include a decomposition of the income series into pre-tax and post-tax (Fisher-Post and Gethin [2023]), allowing for a richer understanding of not only inequality dynamics but also tax-and-transfer systems. Last but not least, in addition to all distributional statistics at the country-level, the WID also includes key macroeconomic aggregates (national accounts, exchange rates, etc.), a quasi-complete decomposition of national income into its various subcomponents, and population aggregates decomposed by gender and age. These developments are intended to bring the WID one step closer to being a one-stop shop for all growth and inequality data.

It would only be fair to say that open data and open access have been core to the philosophy driving the work at the WIL right from its inception. This includes not just public release of data series, along with their corresponding working papers and replication code, but also the development and dissemination of tools and statistical packages to facilitate wider use of cutting-edge empirical methods. A good example is the “Generalized Pareto Interpolation” method, developed by researchers at the WIL (Blanchet, Fournier, and Piketty [2022]), to extract a smooth distribution from tabulated data. Besides the ‘`gpinter`’ package for *R*, the WIL website also hosts a simple and intuitive online tool for those without any familiarity with statistical software. At the same time, for those who are familiar with *Stata* and *R*, *all variables* in the *entire database* available in the WID can be downloaded (bulk or otherwise) using the respective ‘`wid`’ packages. These initiatives have not only encouraged greater research transparency but also enabled and fostered vibrant academic and public debates on inequality issues. Of course, there is still work to be done to further enhance transparency, improve the quality of metadata, and make the statistics more accessible to those less used to working with ratios and percentages.

With this brief background of the WIL and its work in place, we can now move on to the focus of the rest of our article – governments as providers of reliable public statistics. Most of empirical work at the WIL relies on crucial statistical information that is collected and published by public statistical agencies. This is not only the case with tax data but also household surveys, national accounts, and demographic statistics. It goes without saying that much gratitude is owed to the local statistical agencies that are often operating under various constraints, including limited finances and personnel. The WIL remains very open to collaborating with them to improve the distributional aspect of statistics, particularly national accounts. At the same time, in many countries, either these crucial public statistics are not systematically collected at all, or governments are hesitant to release them to the public. Even when such data is available, its quality and richness varies significantly across countries. In the next section, we provide a global snapshot of the state of open data with regards to inequality statistics.

2. INEQUALITY TRANSPARENCY INDEX

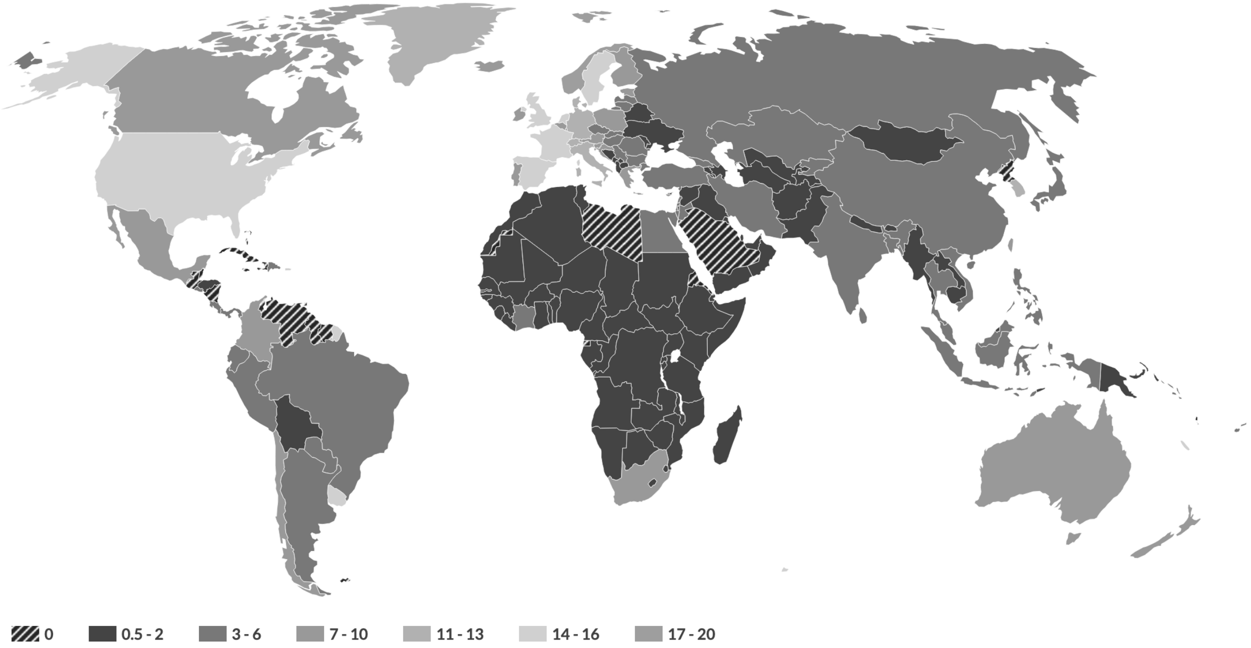
Public access to quality data on income and wealth distributions is a prerequisite for evidence-based policy debates. In this current digital age, access to this basic information should be considered a public good. In this spirit, on the occasion of the 2019 Human Development Report, the WIL partnered with the United Nations Development Programme (UNDP) to develop and release the Inequality Transparency Index. The index, ranging from 0-20, scores each country's income and wealth data on four key parameters: frequency, access, availability of microdata, and data quality. We refer the reader to Burq and Chancel [2020] for technical details about the index.

The latest index scores for all countries from 2023 are presented in Figure 1, with dark gray reflecting low transparency and light gray high transparency. Overall, the predominance of darker gray shades suggests that when it comes to inequality statistics, we remain very much in the 'dark ages'. Or to put it differently, at a time when a growing number of multinational corporations like Google, Facebook, Visa, Mastercard have access to various intimate details of our lives, citizens in most countries hardly have access to even the most basic statistics concerning the distribution of income and wealth (Alvaredo, Chancel, *et al.* [2019]). What's worse, there are 18 countries with scores of exactly 0, marked with black-white stripes in Figure 1. These include Cuba, Libya, North Korea, Saudi Arabia or Venezuela, where rather unfortunately we have absolutely no reliable data on inequality. For the majority of the African continent, even where scores are above 0, they are almost always below 2. In these countries, researchers and citizens alike must make do with whatever little statistical crumbs that local governments make available.

On the more encouraging side of things, the clear star performer in terms of inequality transparency is Norway with a score of 17.5. Indeed, the quantum, quality, and richness of the public statistics coming out of Norway is commendable, at least in comparison to other countries. Following close behind Norway are Spain, United States, United Kingdom, Uruguay, France, Sweden and the Netherlands with scores ranging from 14.5 to 16. These are all countries where high quality data is available but some scope for improvement remains.

Most of the remaining countries, mainly in Asia and Latin America, register scores in the range of 3 to 6. These are countries like India, where some form of income and wealth data is available, but the coverage and quality remain low. For instance, given the low incomes for majority of the population and the relatively high tax exemption thresholds, India's tax data only covers about 10% of adults in recent years. In addition, there is no reliable household survey capturing *all sources* of incomes. Consequently, estimating inequality levels and trends becomes a far more complicated exercise than it should be and forces us to apply a range of corrections to the raw data. In these countries, there is massive scope for improving the quality of inequality statistics. Often, these improvements are possible solely using the data that public agencies are already collecting. For example, almost all tax departments carefully record information on the various sources of incomes of tax units. And yet, sometimes, the way the tabulated

Figure 1 – *Inequality Transparency Index, 2023*



data is presented does not allow for decomposing gross incomes into its various components. This severely restricts the richness of the analysis and prevents careful diagnosis of factors driving the observed trends.

Overall, as things stand today, much remains to be done to enhance transparency of public statistics. This is not only necessary for carefully documenting social and economic facts but in turn also for democratic debates on the legitimacy of economic and social arrangements. As we argue in the next section, it is high time that governments start fulfilling their responsibility as providers of reliable public statistics.

3. GOVERNMENTS AS PROVIDERS OF PUBLIC STATISTICS

Building on the pioneering works of Simon Kuznets and Anthony Atkinson, one of the key innovations of the work at the WIL is the extensive use of tax tabulations, and more recently tax micro-files, to shed sharper light on the dynamics of top incomes. Prior to this, for a variety of reasons (including non-availability of tax data), household surveys served as the main source for studying income (and wealth) inequality around the world. As it turns out, surveys are not the most reliable source for studying inequality due to several well-documented reasons including differential non-response, misreporting, and most importantly, non-coverage of the ultra-rich in usual samples (Bourguignon [2018]; Korinek, Mistiaen, and Ravallion [2006]). In contrast, for many Western economies, tax data provides a much more comprehensive picture of income and wealth for nearly the entire population, including the ultra-rich. Not surprisingly, inequality measures derived from tax data often point to higher levels of concentration than those derived from survey data, particularly at the very top of the distribution—see Burkhauser, Feng, *et al.* [2012] for the United States, Burkhauser, Hérault, *et al.* [2017] for the United Kingdom, Bartels and Metzing [2018] for Germany, and Basu [2025] for India.

Given the scale and coordination necessary, public authorities alone can be expected to generate reliable fiscal data, while maintaining the relevant data protection and privacy standards. In many countries, national statistical agencies are doing a commendable job in producing such data, including anonymized micro tax files, available to researchers and citizens alike. At the same time, in many countries, this data remains entirely elusive. This is unfortunate given that such data is highly policy-relevant, not just with reference to income inequality, but more generally for the study of growth dynamics, social mobility, and the performance of tax-and-transfers system, to name but a few.

The same applies to wealth statistics. In a handful of countries where some form of wealth or estate tax is in force, we have some data to study the evolution of wealth distributions. In most other countries, we only have some form of household wealth surveys, which, just like income surveys, are known to severely miss the right tail of the distribution (Vermuelen [2016]). As a result, citizens and researchers have been forced to rely on wealth rankings published by luxury magazines to shed light on the very top of the distribution. This is a shame given that such rankings involve little transparency about the concepts and methods they use and rely in practice on limited and unsystematic information sources. Unfortunately, this situation is likely to persist as long as public authorities do not fulfill their role as providers of reliable, transparent and systematic statistical information.

We take this opportunity to make a spirited call for wider and more democratic access to fiscal data around the world. At the very least, governments around the world should publish data akin to Tables 1 and 2. This applies not just to income tax tabulations but also, crucially, to wealth statistics. Even in the absence of a comprehensive and well-administered wealth tax, it is common practice for tax administrations in most countries to collect a large quantity of information about wealth. This includes information on real estate assets through the registration of real-estate transactions. This also includes information on company ownership and financial portfolios that should be automatically transmitted from banks to tax administrations to properly control and audit the taxation of financial income flows. The problem in many countries is that this information is typically not used in a systematic manner. We recommend that governments release, along with income tax tabulations, a detailed set of wealth tabulations indicating the numbers of asset holders and amounts of their assets for many wealth brackets and asset classes (real estate, equity, bonds, etc.). This should also include top wealth groups, including multi-millionaires, billionaires and multi-billionaires, so that citizens and experts can have an informed discussion on the basis of well-defined public statistics (rather than by using rich lists published by magazines on the basis of unclear concepts and methods). At this stage, the quality of public information on top wealth groups is highly insufficient, including in countries like Norway and other Nordic countries, where data quality is severely limited by the intensive use of tax havens and offshore wealth by the wealthiest individuals, as highlighted by Alstadsæter, Johannesen, and Zucman [2019].

4. A SPOTLIGHT ON INDIA

Given its geographical size and population, now highest in the world, the distribution of economic growth in India has significant consequences on global inequality dynamics which in turn are crucial to our understanding of global economic arrangements. This makes the careful measurement of income and wealth inequality in India an important exercise.

Given the advances in large-scale data collection, computing power, and more recently artificial intelligence, it would be safe to say we are living in the age of "big data". On the other hand, if one were to go by the recent Indian experience, it would seem that this may instead be described as the age of "big opacity". Over the last decade and a half, a whole range of key public statistics have either become unavailable or their quality has become suspect. This applies to national accounts data, GDP deflators, household surveys, tax tabulations, and more.

Among other things, this regression of public statistics raises numerous complications for measurement of inequality, as amply highlighted in Bharti *et al.* [2024].

Table 1 – *Income and Tax Composition by Net Income Bracket*

Net Income Bracket ($)	**Persons**	**Total Income**	**Labor Income**	**Total Capital Incomes**			**Total Income Taxes**			**Total Wealth Taxes**	
				Housing asset income	Equity and net interest income	Pension and life insurance income	Personal income tax	Corporate income tax	Capital gains tax	Wealth and property tax	Inheritance and estate tax
0–10k	…	…	…	…	…	…	…	…	…	…	…
10k–20k	…	…	…	…	…	…	…	…	…	…	…
20k–30k	…	…	…	…	…	…	…	…	…	…	…
30k–40k	…	…	…	…	…	…	…	…	…	…	…
40k–50k	…	…	…	…	…	…	…	…	…	…	…
50k–70k	…	…	…	…	…	…	…	…	…	…	…
70k–100k	…	…	…	…	…	…	…	…	…	…	…
100k–150k	…	…	…	…	…	…	…	…	…	…	…
150k–200k	…	…	…	…	…	…	…	…	…	…	…
200k–400k	…	…	…	…	…	…	…	…	…	…	…
400k–600k	…	…	…	…	…	…	…	…	…	…	…
600k–800k	…	…	…	…	…	…	…	…	…	…	…
800k–1m	…	…	…	…	…	…	…	…	…	…	…
1m–10m	…	…	…	…	…	…	…	…	…	…	…
10m–100m	…	…	…	…	…	…	…	…	…	…	…
>100m	…	…	…	…	…	…	…	…	…	…	…

Note: Reproduced from Burq and Chancel [2020].

Table 2 – *Wealth and Tax Composition by Net Wealth Bracket*

Net Wealth Bracket ($)	**Persons**	**Total Wealth**									**Total Income**	
		Currency and deposits	Bonds and loans	Equities and shares	Pension fund and life insurance	Real estate	Businesses and non-fin. assets	Debt	Total domestic assets	Total foreign assets	Capital income	Labour income
< 0	…	…	…	…	…	…	…	…	…	…	…	…
0-10k	…	…	…	…	…	…	…	…	…	…	…	…
10k–100k	…	…	…	…	…	…	…	…	…	…	…	…
100k-1m	…	…	…	…	…	…	…	…	…	…	…	…
1m–10m	…	…	…	…	…	…	…	…	…	…	…	…
10m–100m	…	…	…	…	…	…	…	…	…	…	…	…
100m–1b	…	…	…	…	…	…	…	…	…	…	…	…
1b–5b	…	…	…	…	…	…	…	…	…	…	…	…
5b–10b	…	…	…	…	…	…	…	…	…	…	…	…
>10b	…	…	…	…	…	…	…	…	…	…	…	…

Note: Reproduced from Burq and Chancel [2020].

More generally, this has created a rather unfortunate epistemic situation wherein forming justified beliefs about various key aspects of the economic system—pace of economic growth, progress on poverty eradication, and distributional consequences—is either impossible or very costly. As Jitendranath and Somanchi [2025] argue, this is not merely an inconvenience for researchers, but also an ethical wrong as it prevents citizens from evaluating the legitimacy of the state.

This sorry state of affairs is all the more concerning given that right from the time of its independence, the Indian statistical system provided a role model even to Western economies, particularly in the domain of household surveys. The National Sample Survey Organization (NSSO), set up under the leadership of P. C. Mahalanobis, gained enviable credibility overtime and the data from its nationally representative household surveys formed the bedrock of a wide range of studies on poverty, nutrition, inequality, to name but a few. Unfortunately, the recent years have seen excessive political interference in India's statistical system. The most glaring example must be the case of the 2017–2018 round of the Consumption Expenditure Survey (CES). Conducted roughly on a quinquennial basis by the NSSO, the CES is the primary source for gauging economic progress for a representative sample of the Indian population. While the NSSO collected the data for the 2017–2018 round and even prepared a draft report, the data and report were suppressed by the Indian government, on rather unqualified grounds of data quality concerns. An analysis of the numbers in a leaked summary of the report suggested that the consumption-levels in real-terms may have fallen across the distribution (Subramanian [2019]). On the heels of this decision, 108 eminent economists and social scientists made a public appeal to the government to "… restore access and integrity to public statistics … that would feed into economic policy making and that would make for honest and democratic public discourse" (Kazmin [2019]).

Before concluding, it is worth noting that there are some positive signs of change in the last few years. The Period Labour Force Survey has been conducted annually, *without disruption*, since 2017–2018. Fresh rounds of the CES have been conducted in 2022 and 2023 (albeit with some methodological changes making temporal comparability harder). Moreover, various state governments have undertaken their own initiatives to collect better data capturing India's socio-economic realities. For instance, the governments of Bihar and Telangana have conducted a Socio-Economic Caste Census in their respective states. Other states like Jharkhand were reportedly considering the same. At the national level, there has been a similar demand by the opposition INDIA bloc. These are all positive signs. Access to anonymized micro-data or detailed tabulations from these surveys would provide a wealth of information to study a range of key policy-relevant questions. More generally, there is an urgent need for democratic access to statistical data that is free from political interference.

5. CONCLUSION

The experiences from the World Inequality Lab clearly highlight the essential role that open science and disciplinary openness play, especially in terms of gathering, managing, and sharing data on economic inequality. Transparent, accessible, and high-quality data serve as the backbone for rigorous academic research, informed democratic debate, and the development of effective public policies. While some progress has been made in recent years with regard to availability of inequality statistics, significant barriers remain. Many countries continue to struggle with insufficient data transparency, political interference, and inadequate statistical infrastructure.

Addressing these challenges requires a firm and sustained commitment from governments across the globe. It is crucial that authorities recognize and uphold their responsibility to provide reliable and comprehensive fiscal and socio-economic statistics. By doing so, governments can support meaningful democratic participation, foster informed public discourse, and facilitate targeted policies aimed at not only reducing inequality, but improving lives more generally. Ultimately, enhancing the transparency and availability of socio-economic data is not just a technical requirement but a democratic imperative that empowers societies to better understand—and in turn effectively tackle—the pressing challenges of our times.

REFERENCES

Alstadsæter A., Johannesen N., and Zucman G. [2019], “Tax Evasion and Inequality”, *American Economic Review*, 109 (6), pp. 2073–2103.

Alvaredo F., Atkinson A. B., *et al.* [2016], “Distributional National Accounts Guidelines: Methods and Concepts Used in WID.world”, *WID Working Paper*, 2016/2.

Alvaredo F., Chancel L., *et al.* [2019], “Escaping the Inequality-Data Dark Ages”, *Project Syndicate*, December 23rd, URL: `https://www2.project-syndicate.org/commentary/inequality-data-and-denialism-by-facundo-alvaredo-et-al-2019-12`.

Atkinson A. B. and Piketty T. [2007], *Top Incomes over the Twentieth Century : A Contrast Between Continental European and English-Speaking Countries*, Oxford, Oxford University Press.

– [2010], *Top Incomes: A Global Perspective*, Oxford, Oxford University Press.

Bartels C. and Metzing M. [2018], “An Integrated Approach for a Top-Corrected Income Distribution”, *Journal of Economic Inequality*, 17, pp. 125–143.

Basu D. [2025], “Economic Inequality in India Over the Last Three Decades: A Tale of Two Data Sources”, *UMass Amherst Working Paper*.

Bharti N. K. *et al.* [2024], “Income and Wealth Inequality in India, 1922–2023: The Rise of the Billionaire Raj”, *WIL Working Paper*, 2024/09.

Blanchet T., Flores I., and Morgan M. [2022], “The Weight of the Rich: Improving Surveys Using Tax Data”, *Journal of Economic Inequality*, 20, pp. 119–150.

Blanchet T., Fournier J., and Piketty T. [2022], "Generalized Pareto Interpolation: Theory and Applications", *Review of Income and Wealth*, 68 (1), pp. 262–288.

Bourguignon F. [2018], "Simple Adjustments of Observed Distributions for Missing Income and Missing People", *Journal of Economic Inequality*, 16, pp. 171–188.

Burkhauser R. V., Feng S., *et al.* [2012], "Recent Trends in Top Income Shares in the United States: Reconciling Estimates from March CPS and IRS Tax Return Data", *Review of Economics and Statistics*, 94 (2), pp. 371–388.

Burkhauser R. V., Hérault N., *et al.* [2017], "Top Incomes and Inequality in the UK: Reconciling Estimates from Household Survey and Tax Return Data", *Oxford Economic Papers*, 70 (2), pp. 301–326.

Burq F. and Chancel L. [2020], "Inequality Transparency Index Update", *WIL Technical Note*, 2020/12.

Fisher-Post M. and Gethin A. [2023], "Government Redistribution and Development Global Estimates of Tax-and-Transfer Progressivity, 1980–2019", *WIL Working Paper*, 2023/17.

Jitendranath A. and Somanchi A. [2025], "Hoodwiking the Public: The Ethics and Epistemology of Administrative Data (in the Global South)", *PSE Working Paper (forthcoming)*.

Kazmin A. [2019], "Economists Condemn Politiczation of Modi Government Data", *Financial Times*, March 15th.

Korinek A., Mistiaen J. A., and Ravallion M. [2006], "Survey Non-Response and the Distribution of Income", *Journal of Economic Inequality*, 4, pp. 33–55.

Piketty T. and Saez E. [2003], "Income Inequality in the United States, 1913–1998", *Quarterly Journal of Economics*, 118 (1), pp. 1–39.

Piketty T. [2001], *Les hauts revenus en France au 20^{e} siècle. Inégalités et redistribution, 1901–1998*, Paris, Grasset.

Subramanian S. [2019], "What is Happening to Rural Welfare, Poverty, and Inequality in India?", *The India Forum*, November 27th, URL: https://www.theindiaforum.in/article/what-happened-rural-welfare-poverty-and-inequality-india-between-2011-12-and-2017-18.

Vermuelen P. [2016], "Estimating the Top Tail of the Wealth Distribution", *American Economic Review: Papers & Proceedings*, 106 (5), pp. 646–650.