Using Tax Data to Better Capture Top Incomes in Official UK Inequality Statistics

Martin Shine, Richard Tonkin & Dominic Webber (UK Office for National Statistics)

The Office for National Statistics (ONS) and its predecessors have published statistics on the distribution and redistribution of household income since 1961, beginning with "*The Incidence of Taxes and Social Service Benefits*", which was one of the first publications in the world to give such a complete examination of these issues.

Throughout this time, ONS's statistics on income inequality have been based primarily on household surveys, in common with the majority of official statistics on the distribution of household finances globally. Data are currently derived from the Living Costs and Food Survey (LCFS), a voluntary sample survey of private households in the UK. While household surveys have several important benefits over relying solely on administrative records, there is a well-recognised challenge: they do not fully capture the incomes of the very richest individuals and households, particularly those among the so-called "top 1%". There are several potential reasons for this (see e.g. Lustig, 2018), the relative importance of which vary across countries and across surveys depending on the methods used. These include frame or noncoverage error, unit and item nonresponse error, under-reporting and sparseness. Economic research has employed a variety of methods to address these issues, including those using survey data only (e.g. Ruiz & Woloszko, 2016), those using tax data only (e.g. Alvaredo, 2017; Atkinson & Ooms, 2015), and those that combine survey and administrative data in some way (e.g. Burkhauser et al., 2018a; Jenkins 2017).

This paper presents new research and analysis to develop a methodology for improving measurement of the upper tail of the income distribution that is suitable for use in UK official statistics, in terms of being methodologically sound and based on academic standards; transparent and understandable by users; as well as where adjustments are made on underlying microdata rather than aggregates. The methods presented in the paper build upon the work of both Burkhauser et al. (2018b) and the Department for Work and Pensions (2015) in employing methods in which survey-based mean incomes for quantile groups are replaced by equivalent figures from tax data. The primary data sources used are the LCFS, Household Financial Survey (HFS) and the Survey of Personal Incomes (SPI), the latter being a microdata set containing taxable incomes based on a sample of administrative records from UK taxpayers.

The analysis examines two sets of methods. In the first, the mean gross income for each SPI quantile group is imputed onto individuals in the equivalent quantile groups in the survey data. In the second, means for each SPI quantile group are first imputed onto the survey data based on the monetary value of the boundary for each quantile group, before the dataset is reweighted to reflect control totals for each quantile group and the overall population. Where the primary challenge affecting top incomes is that of under-reporting rather than under-coverage, the effects of the two methods should be largely equivalent in practice. However, where under-coverage also has an important impact, the second 'reweighting' method should prove more effective. In addition, the analysis also examines the effects of using different levels of granularity in the quantile groups, along with those of different thresholds for applying the adjustments.

Finally, full SPI data for any year is not normally available at the time of producing official income distribution statistics, and to wait would introduce an unacceptable time lag for users of these estimates. This paper therefore also assesses the impact of using projected SPI data versus waiting for final outturns on the estimates and the conclusions that might be drawn from them.