Appendix A. Differences in Measurement of Income and Earnings Across Alternative Data Sources

A.1. Top Incomes Database

The Top Incomes Database (TID) (http://g-mond.parisschoolofeconomics.eu/topincomes/) provides data on income shares across a range of countries and time periods. Atkinson (2007), Atkinson and Piketty (2010) and Atkinson et al. (2011) provide extensive discussion of the data.

The UK data we use in this study from TID are derived from HMRC publications based on the Survey of Personal Incomes (see below). The data relate to all forms of taxable income and so in addition to wage and salary income also include self-employment profits, rental income and investment income. HMRC provide income distribution tables for all taxpayers. Two key adjustments are then made to the tax data to generate the denominator of total income. First, the tax data relate to taxpayers rather than to the entire adult population. In 2009, for example there were 30.6 million taxpayers but 51.0 million adults. Second, the tax data do not report all forms of income nor does it cover those adults not liable for tax. There are two methods of dealing with this missing income. First, one can begin with the total taxable income figure (£869 billion in 2009) and add estimates for the non-tax-paying population and any missing form of income. Alternatively, one can begin with the total income of households recorded in the National Accounts (Blue Book 2011, Table 6.1.4, ONS Code QWMJ) and subtract various items. Following the second method gives a total income figure of £955 billion. This is then the denominator. The numerator for a particular top percentile group is then simply the total income accruing to the appropriate number of taxpayers from the HMRC data. Thus, for example to compute the share of income going to the top percentile in 2009, we simply compute the income total for the highest earning 0.51 million taxpayers and divide by the total income figure of £955 billion. As shown in Table 1, this amounted to 13.9% in 2009. This procedure follows Atkinson et al. (2011).

A.2. Annual Survey of Hours and Earnings

Annual Survey of Hours and Earnings (ASHE) is based on a 1% sample of employee jobs taken from HMRC PAYE records. The sample is based on an individual’s National Insurance number and so forms an unbalanced panel of workers. Self-employed workers are excluded (as they are not in the PAYE system), as are workers below the PAYE threshold (although some attempt is made to include low-paid and short-hours workers with additional supplementary samples). The data are provided directly by employers and covers both weekly and annual wages.
In calculating wage shares for this study, it is important to recognise that we are computing a different metric than the income shares reported in the TID. To compute a wage share in ASHE, we take all wages in ASHE to generate the wage bill total. We are therefore explicitly excluding the wages of those who fall below the PAYE threshold and the earnings of the self-employed as neither are covered by ASHE. We make no attempt to estimate the wage bill of these omitted workers. More precisely therefore, the ASHE wage share calculations in the study are shares of the total employee wage bill for those workers earning above the PAYE threshold. In contrast, the TID calculations add in those below the threshold and estimate their income.

A.3. Survey of Personal Incomes

The Survey of Personal Incomes (SPI) is based on information held by HMRC tax offices on individuals who could be liable to UK tax. The survey is conducted annually and provides the basis for the data used in the TID. For the purposes of Table 3 in the study, we use the public-use versions of the SPI data made available by the UK Data Archive from 1996 to 2008. The SPI randomly samples from individual tax records with sampling probabilities rising as income increases. The data contain weights to allow for the computation of statistics covering the entire taxpayer population. The most recent SPI (2008) contains 595,742 valid cases. The data provide the most comprehensive and accurate official source of data on personal incomes for all those taxpayers above the PAYE threshold (£5,225 in 2007/8). However HMRC does not record comprehensive information on those earning below this threshold as they are not liable for tax. Thus, the SPI sample is broadly similar in scope to those captured in ASHE, except that the SPI also includes the self-employed.

The data provide a detailed breakdown of total taxable income, identifying income from employment (both employee and self-employed), rental income and all forms of investment income. It provides basic demographic information including sex, age, industry, retirement status and region. We sample all non-retired adults aged 18–65.

A.3.1. Aggregated records in SPI

To maintain taxpayer confidentiality, individual records are aggregated for those on extremely high incomes. In the 2008 survey, the cut-off point for aggregation was £600,000 in total income. Aggregate records are produced by combining individuals above this threshold according to various personal characteristics. In 2008, there were 104 aggregated records, representing 3,130 cases. For each aggregated record, the mean cell value of each variable is reported. In addition, the documentation gives a breakdown of the distribution of region and industry within the cell. We can use this latter information to approximate the sectoral distribution within the aggregated cell. The following example illustrates our method:

**Aggregate Record No. 41 (SPI 2008):**
- Mean total taxable income: £1,375,200 (range: £1,350,000–£1,400,000)
- Number of underlying cases: 24
- Number of underlying cases excluding pensioners: 20
- Number of underlying cases in particular industry: manufacturing (1), wholesale and retail trade (2), financial intermediation (12), real estate, renting and business (4) and other (1).

We replace the aggregated record with 20 individual records each with a total taxable income equal to the cell mean (and weights equal to 1/20th of the aggregated record). We allocate industry codes to these individual records to match the distribution in the underlying cases. The key assumption is that the distribution of total taxable income within the aggregated cell is independent of industry distribution. If instead, for example the financial intermediation workers were all in the top half of the £1,350,000–£1,400,000 income range, then we would underestimate their share of total income. Given the evidence from ASHE on the distribution of
earnings at the very top, we suspect that we are being conservative on the share of income accruing to financial sector workers by assuming independence.

**References**

Atkinson, A. (2007). ‘The distribution of top incomes in the United Kingdom 1908–2000’ in (A. Atkinson and T. Piketty, eds.), *Top Incomes over the 20th Century: A Contrast Between Continental European and English-speaking Countries*, pp. 82–140, Oxford: Oxford University Press.

Atkinson, A. and Piketty, T. (2010). *Top Incomes: A Global Perspective*, Oxford: Oxford University Press.

Atkinson, A., Piketty, T. and Saez, E. (2011). ‘Top incomes in the long run of history’, *Journal of Economic Literature*, vol. 49(1), pp. 3–71.