I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13359

What Accounts for the Rising Share of Women in the Top 1%?

Richard V. Burkhauser
Nicolas Hérault
Stephen P. Jenkins
Roger Wilkins

JUNE 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13359

What Accounts for the Rising Share of Women in the Top 1%?

Richard V. Burkhauser
Cornell University, IZA and NBER

Nicolas Hérault
University of Melbourne

Stephen P. Jenkins
LSE, ISER, University of Melbourne and IZA

Roger Wilkins
University of Melbourne and IZA

JUNE 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

What Accounts for the Rising Share of Women in the Top 1%?*

The share of women in the top 1% of the UK's income distribution has been growing over the last two decades (as in several other countries). Our first contribution is to account for this secular change using regressions of the probability of being in the top 1%, fitted separately for men and women, in order to contrast between the sexes the role of changes in characteristics and changes in returns to characteristics. We show that the rise of women in the top 1% is primarily accounted for by their greater increases (relative to men) in the number of years spent in full-time education. Although most top income analysis uses tax return data, we derive our findings taking advantage of the much more extensive information about personal characteristics that is available in survey data. Our use of survey data requires justification given survey under-coverage of top incomes. Providing this justification is our second contribution.

JEL Classification: D31, J16, C81

Keywords: Top 1%, top incomes, inequality, gender differences, survey under-coverage

Corresponding author:
Nicolas Hérault
Melbourne Institute
University of Melbourne
Faculty of Business and Economics Building
111 Barry Street
Melbourne VIC 3010
Australia
E-mail: nherault@unimelb.edu.au

* Our initial research on this topic was supported by an Australian Research Council Discovery Grant (award DP150102409). Burkhauser's research was partially supported by a Professorial Research Fellowship at the University of Melbourne. Jenkins's research was also partially supported by core funding of the Research Centre on Micro-Social Change at the Institute for Social and Economic Research by the University of Essex and the UK Economic and Social Research Council (award ES/L009153). We thank Gilles Hérault for developing the routine to code in the SPI composite records.

1. Introduction

'In the recent research on top incomes, there has been little discussion of gender' (Atkinson et al. 2018: 225). This is an important gap to fill since top incomes are the central focus of much recent income distribution analysis, building on the pioneering research of Thomas Piketty and collaborators (see e.g. Atkinson et al. 2011 for a review) and the on-going work centred around the World Inequality Database project (WID, https://wid.world/). Drawing on income tax and related administrative register data sources, the top incomes research field has highlighted how, in many countries, the most significant changes in income distribution have been occurring at the top – it is these that are driving the much-discussed increase in inequality. However, learning more about the gender divide at the top of the income distribution is important not only because it contributes to our knowledge about trends in vertical inequality but also because differences between the sexes are a prominent horizontal inequality and hence interesting in their own right.

In this paper, we build on recent research documenting a rising share of women in the top 1% of the income distribution of several countries and, focusing on the UK, analyse the factors that account for this significant shift using regression-based decompositions. We also demonstrate the validity of survey data for this exercise – we present evidence that the oft-cited issue of survey under-coverage of top incomes does not prejudice our analysis, and so we can exploit the much greater information about personal characteristics that is available in survey data compared to most of the administrative record data sources used in top incomes research to date.

Before Atkinson et al.'s (2018) research, there was 'a strong suspicion that women are under-represented [in the top income group], but there is a shortage of hard evidence' (2018: 226). They provide that evidence for eight countries with income taxation systems in which the income unit is the individual (Australia, Canada, Denmark, Italy, New Zealand, Norway, Spain, and the UK).[1] In all eight countries, the share of women in the top income group of the gross income distribution increased between 2000 and 2013, whether 'top' is defined as the top 10%, top 1%, or top 0.1% (2018: Table 1). For example, in the UK, the share of women

[1] Countries with independent taxation are typically the ones for which one can identify women's incomes separately from men's incomes using administrative record data: hence the focus on these countries. Atkinson et al. (2018: 229–230) present four reasons why 'the presence of women at the top of the income distribution is likely to be overestimated in [their] data, and as a result, … [their] analysis is likely to under-estimate the real extent of the gender divide at the top' (2018: 230).

in the top 1% increased from 14.0% to 18.2%. For the countries for which longer data series are available (Canada, Denmark, Norway, the UK), there has been a secular upward trend since around 1980 (2018: Figure 1). Atkinson et al. (2018) also document that in each of the eight countries, the trend increase in the share of women at the top is smaller, the further one goes towards the very top of the income distribution; thus, the decline in under-representation of women is lower as one goes up the income scale.

Two recent papers show that Atkinson et al.'s findings also describe the situation in Finland and Sweden (two countries also with independent taxation): see Ravaska (2018) and Boschini et al. (2020). There is no comparable evidence for the USA that we are aware of but there is related research indicating that women's incomes have become more important at the top of the income distribution there as well. Yavorsky et al. (2019) consider the top 1% of the US gross pre-tax household income distribution (rather than the individual income distribution as in the studies cited above for countries with independent taxation). They show that, in 1995, the fraction of women whose income was sufficient to put their household's income in the top 1% was 1.7% but the corresponding figure was 4.5% by 2016 (2019: Figure 2).

Why the share of women at the top of the income distribution has increased over time has not been studied in detail. This motivates the focus of this paper – accounting for this trend. Atkinson et al. (2018) examine the composition of top incomes, concluding that 'investment income is a particularly important source of income for women at the top compared to men, with self-employment income playing similar role for men and women, and earned income [of men] compensating the difference in importance of investment income' (2018: 246).[2] They also refer to a number of factors (ageing and mating patterns, tax structure, female labour force participation and wealth holding) that may explain the differences in high-income women's and men's income packages, but they do not explicitly consider how these account for the *trend* in the share of women at the top.

Our paper does account for the trend in the share of women at the top, focusing on the UK. Rather than looking for explanations in terms of changes in the types of income that top-income earners hold, we relate trends to changes between top-income men and women in their characteristics and the returns to those characteristics, employing a regression-based decomposition approach.

[2] As Atkinson et al (2018) emphasize, these results highlight the importance for understanding 'top incomes' of looking at total income, rather than simply employment earnings. For analysis of top earnings in the UK, see e.g. Bell and Van Reenen (2014) and Stewart (2011).

In any given year, women's share of the top *x*% income group is equal to the fraction of women who are in the top *x*% (the 'top income rate' among women) multiplied by the fraction of women in the population, all divided by *x*. But over time, *x* is constant (by construction) and the fraction of women in the population hardly changes at all. Hence to account for trends in the share of women, one can model trends in the top income rate among women. More specifically, we model the probability a woman belongs to the top 1% group in a given year as a function of her characteristics. We then relate changes in the average probability between a pair of years (and hence changes in the aggregate share) to changes in distribution of women's characteristics (e.g. changes in age structure, living arrangements, educational attainment, etc.) and changes in the 'returns' to characteristics. Undertaking an analogous exercise for men as well, we contrast the components accounting for the differential trends over time in men's and women's top income shares.

We are the first to employ this approach to examine top income share trends. Although several papers have looked at the personal characteristics of top-income men and women, most have employed univariate breakdowns and not looked at changes over time. Examples of studies about the UK are Brewer, Sibeta, and Wren-Lewis (2017), Brewer and Sámano-Robles (2019), and Joyce et al. (2019).

Ravaska's (2018) study of top income trends in Finland includes regressions of top income rates on characteristics, for women and for men, fitted to pooled data for 1995–2012 and including year fixed effects (unreported). Bobilev et al. (2020) report estimates from regressions of top income group membership on characteristics, in their case, fitted to LIS data for multiple countries and multiple LIS waves, pooled and including country and wave fixed effects (unreported). The approach undertaken in these two papers is unsuitable for accounting for trends in the share of women at the top; pooling means that there is only one distribution of characteristics and one set of returns to those characteristics.[3]

One reason for the lack of regression-based studies of differential trends in top income rates relates to the data sources used to study top incomes. The top incomes literature is founded on administrative record data from the personal income tax system. By contrast with survey data, tax data have substantial advantages, including greater coverage of the high- and very high-income ranges, very large sample sizes, and very long historical coverage (Atkinson et al. 2011). However, the information on personal and other

[3] Bobilev et al. (2020) also run some pooled data regressions in which covariates for children and education are interacted with survey year, but they do not undertake a full regression-based decomposition.

characteristics of tax-payers that is available in tax record datasets is limited, relating only to information required for the administration of the tax system (educational attainment is not required for instance). Another important limitation is that tax record datasets seldom cover the full population as they are limited to the universe of tax filers. These limitations place severe constraints on the usefulness of regression-based analysis of tax data.[4]

Two notable strengths of income surveys are that (i) they are nationally representative and (ii) they contain very extensive information about not only respondents' incomes but also their personal characteristics and their household context.[5] Hence, surveys are obvious candidate data sources for any regression-based study of the association between personal characteristics and top income group membership. There are two major threats to their usefulness in this context: sample sizes may be too small for top income analysis, and there is under-coverage of top incomes. We address both these threats in this paper.

We deal with sample size issues in our regression analyses by pooling survey data across years. The base year for our study of trends is '1999' with the analysis based on pooled data for 1998/99, 1999/2000, and 2000/01, and our final year '2015' is based on pooled data for 2014/15, 2015/16, and 2016/17. As a result of pooling, our regression sample sizes (unweighted) for base and final years are 65,702 and 47,674 for women and 58,062 and 42,707 for men. The unweighted total size is thus 123,764 in the base year and 90,381 in the final year. We focus on the top 1% in this paper, which corresponds to around 1,230 individuals in the base year and 900 individuals in the final year (ignoring complications from weighting). Women accounted for 13% of the top 1% in the base year and nearly one-fifth in the final year (Atkinson et al. 2018; see also below), so our analysis of top income women is based on at least 180 cases. This is sufficient for deriving statistically reliable estimates. However, it also explains why we focus in this paper on the top 1% and do not

[4] Only for a small number of countries – essentially the Nordic countries, in which income tax registers are combined with many other registers – are there data including many personal and other characteristics. (Even if there is independent taxation of men and women, information about spousal characteristics may be available in these sources.) The other exceptional case is when the survey addresses top-income under-coverage directly by using a top-income over-sample. This is the situation with Yavorsky et al.'s (2019) analysis of the US Survey of Consumer Finances.

[5] A related advantage of surveys is that researchers can construct more comprehensive measures of 'income', including e.g. government transfers and account for deductions such as income tax and social insurance contributions; and they can adjust for differences in household size and composition using equivalence scales. These provide better measures of personal economic well-being or 'living standards' than do the less comprehensive income definitions typically available from tax data. (For more extensive discussion, see e.g. Burkhauser et al. 2018*a*.) We do not pursue the issue of the appropriate definition of 'income' in this paper; throughout we work with the definition that prevails in the top incomes literature. We discuss this gross (pre-tax) individual income definition in greater detail below.

look also at the top 0.1% (say), for which sample sizes would be too small. Bobilev et al. (2020) correctly emphasize the problems of small sample sizes for survey data analysis of top incomes but they are unable to pool data from consecutive years as we do because they use LIS data. In the time series of cross-sections that they use, the survey 'waves' are several years apart, and many LIS countries do not have yearly sample sizes as large as ours in any case (more about our UK survey data below).

How we address the potential problems of survey under-coverage of top incomes is a new contribution. It is well-known that survey under-coverage of top incomes may arise for two reasons (for a recent review, see Lustig 2019, especially Section 3.4). First, the very richest individuals in a population may not participate in a survey at all, because they are not in the sampling frame and hence not at risk of being contacted or because they refuse to participate when contacted (differential unit non-response). Second, the very rich individuals who do respond may under-report their incomes, intentionally or unintentionally.

We find that it is the second factor, under-reporting, that predominates in the UK. We show that the top 1% of individuals in each year of our survey data have remarkably similar (non-income) characteristics as the top 1% of individuals in the same year's tax data, for all the characteristics available in the tax data (which refer to individuals' sex, age, region, and industry – see below). Moreover, in both sources, the characteristics of the top 1% are distinctly different from the characteristics of the next nine percentile groups forming the top 10%. As part of this analysis we develop novel methods for comparing the similarity of a pair of multivariate discrete distributions (defined by the set of categorical variables summarising non-income characteristics in the current application). We propose that 'similarity' be assessed in terms of the 'distance' between the pair of distributions with the distance metric summarizing the extent to which there is overlap between two discrete density functions (which turns out to be equivalent to an L1 distance norm, otherwise known as city block distance).

In sum, the joint distribution of the characteristics cited above leads to income *ranks* that are very similar in the two data sources, and hence, we argue, the top 1% group can be reliably identified in the survey data, since top 1% membership is defined by income rank and not income level (which may be under-reported). In addition, we assume that the joint distribution of the set of characteristics cited above is sufficient to correctly identify the income rank and hence top income group membership in particular. The empirical finding about similarity in characteristics of the top 1% in both types of data source combined with the sufficiency assumption has a powerful and useful implication: having correctly identified

the members of the top-income group, we can describe them in greater detail by drawing on the characteristics in the survey in addition to the ones used for the cross-source alignment exercise and we can compare them to the rest of the population. Neither the full-population coverage nor these additional characteristics are available in the tax data. In sum, we can exploit the survey data to undertake regression-based decomposition of trends in top-income group membership using a large number of relevant characteristics as regressors.

Our alignment assumption is untestable with existing data, though we believe it to be plausible. For it to be invalidated one would need non-negligible non-random unit non-response to the survey that is orthogonal in levels and trend to the characteristics that are shared across data sources (sex, age, region and industry). Also in favour of our case is earlier evidence that it is under-reporting rather than differential unit non-contact and non-response that drives top income under-coverage in the UK. Jenkins (2017) shows that the real income values of the 90th, 95th, and 99th percentiles in the survey data are close to the real income values of the corresponding percentiles in the tax data for the same year, for each year over a 15-year period. If the top 1% were simply absent from the survey data (and there were no under-reporting or other under-representation below the top 1%), we would expect the income values at these percentiles to be lower in the survey data than in the tax data (for example, the 99th percentile in the survey data might correspond to the 95th percentile in the tax data).[6]

The rest of the paper unfolds as follows. In Section 2, we introduce the survey and tax data that we employ, namely (1) the Households Below Average Income subfiles of the Family Resources Survey ('HBAI'), and (2) the Survey of Personal Incomes (SPI) which is a very large stratified sample of UK personal tax returns. Throughout, 'income' is gross taxable (pre-tax pre-transfer) income distributed among individuals aged 15 or older. We explain how we construct an income variable based on this in the survey data so that it corresponds with the tax data definition – which is that used by the majority of the top incomes literature to date). In Section 3, we justify our claim that we can use survey data to reliably identify the

[6] There are several explanations for our findings regarding the nature of survey under-coverage of top incomes. The grossing-up weights in the survey data may be doing a sufficiently good job of representing the population, including top-income earners, despite the number of very-high income respondents being relatively small. Also, there may be problems with the tax data which mean that it is not only survey data that do not capture the very highest income earners (i.e. both data sources may identify the same top 1% once problems with both sources are acknowledged). Income tax data capture only income that is recorded for taxation purposes. Income not captured can legitimately include the income of individuals counted as not domiciled in the UK for tax purposes, income from tax-exempt investments and close company retained profits, and legal tax avoidance schemes. There may also be tax non-compliance (including evasion). For a review of these problems in the UK context, see Summers (2019).

top 1% group corresponding to the tax data benchmark. We explain our distance-based approach to checking alignment across data sources. Section 4 contains our regression-based analysis of differential trends between men and women in top 1% membership. We show that the rise of women in the top 1% is primarily associated with their greater gains (relative to men) in educational attainments. Our summary and conclusions are presented in Section 5. There are two appendices containing additional discussion and estimates.

2. UK tax data (the SPI) and survey data (the HBAI)

In the UK, unit record income tax return data are available from the Survey of Personal Incomes (SPI) since the mid-1990s. Each year's SPI observations are a stratified sample of administrative records for individuals who could be liable to UK tax. The data refer to individuals (as opposed to family units) because the individual is the assessment unit for UK personal income taxes (since 1990) – there is independent taxation. The total number of individuals in the SPI has increased steadily over time, from around 57,000 individuals in 1995/96 to nearly 743,000 in 2015/16 (the latest year for which data are available), representing roughly 49 million people per year. For further details, see the documents accompanying HM Revenue and Customs KAI Data, Policy and Co-ordination (2014) and corresponding documentation for other years.

To derive yearly distributions of income covering the population of all adults aged 15 and over, SPI distributions have to be adjusted because they refer only to taxpayers, and so do not cover individuals with incomes below the tax threshold and their incomes. We make the same adjustments to the SPI data as previous research, using control totals. The numbers of people in the top income groups are based on estimates of the UK population aged 15 or older produced annually by the Office of National Statistics (ONS).

For analysis of the UK income distribution, the most-commonly-used survey data source is the Family Resources Survey (FRS) and the accompanying subfiles of derived income variables called the Households Below Average Income (HBAI) dataset. The Department of Work and Pensions (DWP) administers the FRS, and DWP staff produce the HBAI subfile that they use to derive the UK's official income distribution statistics published

annually.[7] Despite its label, the HBAI provides information about the income distribution as a whole.

It is commonly argued, as we discussed earlier, that household surveys have incomplete coverage of income in the very top income ranges. Atkinson et al. (2011) referring to the USA, state that 'the top percentile plays a major role in the increase in the Gini over the last three decades and [Current Population Survey] data that do not measure top incomes fail to capture about half of this increase in overall inequality' (2011: 32). Similarly, Burkhauser et al. (2018*a*) show that correcting for under-coverage of top incomes leads to a marked increase in both the level and trend in measures of inequality in the UK.

Under-coverage at the top is an issue that has long exercised the producers of the UK's HBAI statistics.[8] Each year since 1992, the derived variables in the HBAI subfile accompanying the basic survey dataset have contained an 'SPI-adjustment' to 'improve the quality of data on very high incomes and combat spurious volatility' (Department of Social Security 1996: 23). However, as we show in Burkhauser et al. (2018*b*), this adjustment does not fully account for under-coverage.

To assign individuals to top income groups in the HBAI data – which involves ranking them in order of income – we use the survey weights and a version of the HBAI data that is not 'SPI-adjusted'. The latter choice is because the adjusted (HBAI-SPI) series includes adjustments to 'very rich' individuals made by the DWP based on data from HMRC that lead to undesirable re-rankings of top income earners (Burkhauser et al. 2018*b*). We find that relying on the HBAI 'SPI-adjusted' series worsens the alignment with tax return data as the original survey information on individual income ranks is largely lost in the cell-mean imputation process.

We use the detailed information about individual income components in the FRS and HBAI to construct an income variable that is the same as the principal one available from the SPI. That is, 'income' is gross income, comprising total taxable income from the market ('market income') plus taxable government transfers, and before the deduction of income tax. The yearly distributions of gross income refer to distributions among all persons aged 15 and over. These are the definitions of income distributions that are conventionally used in the top

[7] See e.g. Department for Work and Pensions (2017) covering fiscal years 1994/95 through 2015/16. (The UK fiscal year runs from April through to the following March.) The Institute for Fiscal Studies, whom the DWP contract to check its HBAI calculations, produce their own annual report based on the data (see e.g. Belfield et al. 2017).

[8] There is also concern about the quality of survey data on the very lowest incomes: see e.g. Brewer, Ethridge, and O'Dea (2017) and references therein. We do not consider this issue.

incomes literature (cf. Atkinson et al. 2011). It is the same definition for the UK as employed by Atkinson et al. (2018) in their study of the gender divide in top incomes.

3. Are the characteristics of the top 1% the same in the survey and tax data?

In this section we investigate the extent to which the non-income characteristics at given top income ranks are the same in the HBAI and SPI data. This exercise treats the SPI microdata as the benchmark on the grounds that they are the most reliable source of information on top incomes in the UK, albeit with relatively little information about the non-income characteristics.

First, we first compare univariate distributions in terms of the characteristics of tax filers that are available in the SPI data (as well as the HBAI data): sex, age, region, and industry. We then show that the HBAI/SPI similarity extends from univariate comparisons to multivariate comparisons by introducing distance measures that are functions of multiple characteristics. In this second-stage analysis, we also examine lower percentile groups – namely, each remaining percentile group in the top 10% – to show that not only are the characteristics of the top 1% very similar in the two data sources, but they are also distinctly different from the characteristics of lower percentiles, making the top 1% a group that stands apart.

In the SPI data, information about age is available from 1997/98 onwards, with individuals classified into seven groups (under 25 years, 25–34, 35–44, 45–54, 55–64, 65–74, and 75 years and over). Region of residence is available for the entire SPI period, with 12 'Government Office' regions distinguished. These age and region categories can also be identified in the HBAI data in every year.

Although the industrial classification used in the HBAI is more detailed than in the SPI, in principle 14 categories of industry of employed people can be compared consistently across both datasets from 1998/99 onwards as they both rely on the Standard Industrial Classification (SIC). However, this comparison is made difficult in practice because additional industry codes are added in the SPI. These additional industry codes do not appear in the SIC as they indicate the presence or absence of income from a specific source (e.g. income from pensions, income from financial investments) instead of the individual's industry. For the individuals with these industry codes, the relevant industry is unknown. These codes refer mostly (but not exclusively) to individuals who are not of working age. We

address this issue by restricting the comparison of industries to the working age population, though we note that comparability is more limited than for the other variables discussed above.

A further difficulty relates to the treatment of composite records in the SPI. In order to preserve anonymity, composite records are created by combining data for individuals with extremely high incomes. For example, in 2010/11, 511 individuals with income over £1m are grouped into 27 composite records. Age and sex are defined for these records, but region and industry are set to missing in the micro-data. However, each year the SPI user manual specifies the (univariate) distributions of each composite record by region and industry. We have compiled this information and we 'split' composite records accordingly to recreate their distribution across region and industry as indicated in each year's user manual. (Brewer and Sámano-Robles 2019 do the same in their analysis of top incomes using SPI data.)

The timing of the survey and tax data collection and the income reference periods are not identical and so, even if the same individuals were covered by both datasets, some differences in the composition of the top 1% might occur.[9] Moreover, it is possible that some individuals live at another address than the one reported on their tax returns, which could generate further discrepancies between the HBAI and the SPI in terms of the distribution by government office regions.

Nonetheless, as we show below, we find a high degree of consistency across both datasets with respect to the distribution of these demographic characteristics in the top 1% income group.

Univariate comparisons

Figure 1 reports the share of women in the top 1% according to both the SPI and the HBAI, together with the 95% confidence intervals associated with these estimates. The comparison covers the period between 1995/96 and 2015/16, the only period for which both the SPI and HBAI were publicly available at the time of writing. However, no SPI data are available for 2008/09 and 2011/12.

[9] Survey interviews may take place in any month during a financial year (the FRS is a continuous survey) and for most income sources respondents are asked the last amount received and the period (week, month, year, etc.) to which it refers. (The data producers convert responses to weekly amounts pro rata. There is no specific reference period such as 'the previous calendar year'.) The SPI data refer to total income received over the whole of a financial year.

The figure shows that the share of women in the top 1% in the HBAI closely tracks the corresponding share in the SPI. This is true both in terms of levels and trends, although the HBAI confidence intervals are larger than the SPI's intervals due to the much smaller HBAI sample size. Consistent with Atkinson et al. (2018), who examine the share of women in top income groups in eight countries including the UK, the clear trend is of a rising female share of the top 1% between 1995/96 and 2015/16. Nonetheless, in 2015/16, the female share of the top 1% was still only 19%.

Table 1 further shows that the HBAI provides a similar picture to the SPI not only for the gender composition of the top 1%, but also for its age, region, and industry composition. Table 1*a* reports the distribution of individuals in the top 1% by age and government office region in both the HBAI and the SPI for the 1995/96 to 2015/16 period. Table 1*b* reports the distribution by industry for those individuals in the top 1% who are of working-age (which we define as 25–54 years).

The first two rows of Tables 1*a* and 1*b* give sample sizes year by year. They show that the SPI is considerably larger than the HBAI. For example, in 2010/11 when 1 percent of the UK adult population was made of slightly more than half a million individuals, this group was represented by 47,287 records in the SPI but 450 in the HBAI.

Tables 1*a* and 1*b* use emboldening and italicisation to highlight the cases where the SPI estimate falls within the 95% confidence interval of the HBAI estimate. The vast majority of cells are highlighted in this way. For example, Table 1*a* indicates that the SPI estimates of the share of women in the top 1% is within the 95% confidence interval of the HBAI estimate for all years (a finding also apparent from Figure 1).

The correspondence for the distributions of industry is not as good as for other characteristics. This is likely because industry is missing for a substantial number of cases, particularly in the SPI, which limits comparability. Between 4% and 12.6% of working-age individuals in the top 1% have missing or no applicable industry in the SPI. In the HBAI, this proportion fluctuates between 0.6% and 5.6%. Even if these industries were missing at random in both datasets, the comparability of the distribution across the remaining industry groups would be affected.

A by-product of our alignment exercise is that we provide new evidence about the age, region, and industry composition of the top 1%, drawing on both SPI and HBAI data. (Brewer and Sámano-Robles 2019 also provide univariate breakdowns.) Table 1 shows that the 35 to 44 and 45 to 54 age groups account for over 60% of the top 1%, but with the older of the two age groups tending to account for a growing share over the two decades from

1995/96. Consistent with this is a broader ageing of the top 1%, with the proportion of the top 1% aged 55 to 64 having increased, and the proportion aged under 35 having decreased.

Comparing across regions, the London and South-East England regions together account for 50% or more of the top 1% in most years. The East of England region also contains a sizeable proportion of the top 1%, accounting for at least 10% of this income group in most years. Also notable is that very few of the top 1% are found in UK regions outside of England.

Comparing across industries, financial intermediation as well as real estate, renting and business dominate the top 1%. In 1997/98, 47% of the top 1% who were of working-age worked in one of these two industries, while in 2015/16, 51% of the top 1% worked in one of these industries. Significant numbers of the top 1% also work in wholesale and retail trade and in health and social work. Significantly, there has been a substantial decline in the proportion of the top 1% working in manufacturing, falling from 11.4% in 1997/98 to 5% in 2015/16.

Multivariate analysis – distance measures

The discussion so far shows the close correspondence in univariate distributions of individual characteristics between the HBAI and the SPI for the top 1% income group. But are the results as good for joint/multivariate distributions? In other words, the HBAI gets the share of men in the top 1% right and the share of those living in London right and the share of those aged 35–44 right. But does it get the share of men aged 35–44 living in London right?

To answer this question, we use a measure of the overall 'distance' between two multivariate distributions. Consider the (weighted) number of individuals who have exactly the same characteristics in the HBAI and SPI in the income group under consideration. This count divided by the total number of individuals is the 'proportional overlap', and one minus this quantity provides a distance measure we term 'overlapping distance', *D*. We show in Appendix A that *D* is directly related to L1 Euclidean distance, also known as the Manhattan or city block distance.

The overlapping distance measure, *D*, is bounded between 0 and 1 and satisfies the Triangle Inequality (the distance between distributions *A* and *C* is no larger than the distance between *A*, *B* plus distance between *B*, *C*), Symmetry, and Identity. Intuitively, the overlapping distance is one minus the common support or the 'overlap' between the two distributions and as such, it has a natural interpretation. For instance, a value of 0.1 means that the overlap (common support) between the HBAI and SPI distributions is 90%. In other

words, the two distributions do not overlap for 10% of individuals. Were we to plot two histograms for the HBAI and SPI distributions respectively, *D* would be the proportion of the area covered by the non-overlapping bars (i.e. one minus the overlapping area).

This approach makes no distinction between observations that are 'close' in characteristics and those that are very different in characteristics. For inherently binary characteristics (such as sex), this is not an issue, but it may be an issue for characteristics that are inherently continuous or that are summarized by an ordered categorical variable. For example, if we use a separate variable for each single year of age, *D* is insensitive to whether individuals in the HBAI and SPI are 30 years apart or one year apart in age.

This means that the choice of the categories is an important step in implementing this measure. It amounts to a decision about when two observations are 'close enough' to be considered similar or, equivalently, when they are 'far enough apart' to be considered different in a meaningful way. For example, returning to our age example, we might consider two individuals to be sufficiently similar in terms of age if they belong to the same five-year age band, in which case we would wish to create a separate variable for each five-year age category (rather than each single-year age category). In our context, the amount of information available in the SPI is limited and we use all that is available, i.e. the two gender groups, seven age groups, and twelve regions.

We show in Appendix A that our findings based on our overlapping distance measure do not change materially if we use any one of four other distance metrics.

Multivariate analysis – implementation

We exclude industry due to the high number of missing values in both datasets and the restricted comparability to working-age adults. We also exclude SPI composite records as there is no information on the joint distributions of their characteristics in the SPI (see above). Composite records represented up to 6.23 per cent (in 2007/08) of the top 1% income group in the SPI, or 0.06% of the adult population. From 2010/11 onwards, however, their share in the top 1% is 0.58% or less. To compare like with like and to ensure consistency across time, we thus exclude the top 0.1% for all multivariate comparisons based on years prior to 2010/11, and do not apply any exclusion thereafter as there are very few composite records from 2010/11 onwards.

We have a limited set of categorical variables present in both datasets: gender (2 groups), age (7 groups) and region (12 groups). If we consider all possible combinations, we have a total of $2 \times 7 \times 12 = 168$ subgroups or cells. To go beyond the univariate comparisons

discussed above and to assess the comparability of the multivariate distributions between the SPI and the HBAI, we first compare the composition of the top 1% in terms of these 168 cells.

Figure 2 reports how individuals in the top 1% are distributed across these 168 cells in the SPI and in the HBAI in 2015/16, the most recent year available. It shows a very similar composition of the top 1% in both datasets. For instance, the HBAI reproduces the two largest spikes corresponding to men of working age living in London with remarkable precision. Results for other years (available upon request) are very similar. In other words, the distribution of the top 1% across the 168 cells is very similar in both SPI and HBAI in all years.

Table 2 shows the distance (summarised by *D*) between the multivariate distribution of individual characteristics of the top 1% in the HBAI and of each of the top 10 percentile income groups in the SPI (Panel A) and the distance between the top 1% in the SPI and each of the top 10 percentile income groups in the HBAI (Panel B).

Panel A of Table 2 indicates that, of the top 10 percentile groups in the SPI data, the closest to the HBAI top 1% is the SPI top 1%; while Panel B of the table indicates that, of the top 10 percentile groups in the HBAI data, the closest to the SPI top 1% is the HBAI top 1%. Although the estimated distance between the two top 1% groups is statistically significant, it is small in absolute terms. The overlapping distance measure indicates that more than 80% of individuals in the top 1% of the HBAI and SPI have precisely the same age, gender and region.

This overlap increases to 90% if we reduce the number of age bands and the number of regions to 4, from 7 and 12 respectively, leading to 32 groups instead of 168 (these additional results are available upon request).

The clear gradient in the distances between the top 1% in one dataset and the other top 10 percentile income groups in the other dataset also suggests clear differences between the top 1% and other top income groups. That is, in both data sources, the further one moves from the top 1% in one data source, the larger the distance with the top 1% in the other data source. The top 1% thus appears to be a group of individuals that stand apart, even within the top 10%.

Implications

The analysis in this section shows that the characteristics of the people in the top 1% of the SPI data are remarkably similar to the characteristics of the people in the top 1% of the HBAI

data. Assuming that the characteristics used in this alignment exercise are sufficient to correctly identify the top 1%, we can exploit the fact that the survey data also contains a large number of additional characteristics for individuals, and these can be used in a regression-based decomposition of trends exercise.

4. The rise of women in the top 1%: a decomposition analysis

The share of women in the top 1% was only 10% in 1995/96 and had nearly doubled to 19% by 2015/16 (Figure 1, SPI series). Based on the evidence presented in Section 3 and showing that the individuals in the top 1% of the survey data correspond to the top 1% in the tax data, we draw on the comparative richness of the HBAI data to examine in greater depth the rise of the share of women in the top 1%.

We analyse changes in the probability of top 1% membership between '1999' and '2015' for men and women respectively, using survey data covering the whole population. As explained earlier, we pool three years of HBAI data together to ensure that sample sizes for each 'year' are sufficiently large (1998/99 to 2000/01 and 2014/15 to 2016/17). These are the earliest and latest sets of years for which the HBAI provides information on occupation and industry. When we repeated the analysis using single years of HBAI data rather than pooling three years, the results were essentially the same.

We use an Oaxaca-Blinder style decomposition approach extended to non-linear regressions to account for the rise in the probability of being in the top 1% for women and the corresponding fall for men. What we are looking to see is the extent to which changes over time in the probability of top 1% membership are due to differential trends in the distributions of men's and women's characteristics (their education, employment status, where they live, etc.) or to differentials trends in the extent to which men's and women's characteristics translate into higher probabilities of top-income group membership.

We suppose that each individual *i*'s probability of top 1% membership in any year *t*, P_i^t, is a non-linear function of individual *i*'s characteristics, i.e.

$$P_i^t = F(X_i^t \hat{\beta}^t) \quad (1)$$

where X_i^t is a vector of characteristics for *i* in year *t*, and $\hat{\beta}^t$ is the vector of estimated returns associated with each of those characteristics. Taking averages over individuals, we write the

change in the top-income membership rate between years *s* and *t* as $\Delta\bar{P} = \bar{P}^t - \bar{P}^s$. From eq. (1), this change can be written as:

$$\bar{P}^t - \bar{P}^s = \left[\sum_{i=1}^{N^t} \frac{F\left(X_i^t \hat{\beta}^t\right)}{N^t} - \sum_{i=1}^{N^s} \frac{F\left(X_i^s \hat{\beta}^t\right)}{N^s}\right] + \left[\sum_{i=1}^{N^s} \frac{F\left(X_i^s \hat{\beta}^t\right)}{N^s} - \sum_{i=1}^{N^s} \frac{F\left(X_i^s \hat{\beta}^s\right)}{N^s}\right]. \qquad (2)$$

$$\Delta\bar{P} = \quad \text{changes in characteristics} \quad + \quad \text{changes in returns to characteristics}$$

Thus, the change in the top-income membership rate between years *s* and *t* is decomposed into two components: the first term on the right-hand side of eq. (2) reflects changes in characteristics; the second term reflects changes in the returns to characteristics.

We fit regression models separately for men and women and hence derive separate trend decompositions for each sex. In our application, years *s* and *t* are 1999 and 2015 (in the pooled data sense explained earlier). Because this type of decomposition is potentially sensitive to which year is used as the reference point, we undertake the decomposition analysis twice, once for each the reference year. The specific method we use to implement the decomposition is that of Fairlie (2005) because it straightforwardly allows us to assess not only the total contributions to trends of characteristics and of returns to characteristics but also the separate contributions of subsets of characteristics. We take *F*(.) to be the logit function, randomise the order of the variables in the detailed decompositions, and use 1000 replications. We implement the decomposition calculations using the software module of Jann (2006).[10]

Our choice of characteristics (*X*) is informed by conventional models of the determinants of employment earnings as this is the primary income source for most working-age individuals. However, there are complications because 'income' in this context also includes income from self-employment, investment income, and so on. Relatedly, income from savings and investments is particularly important for individuals older than commonly-defined upper bounds to 'working age'. It is also well-known that the chances of being a top earner in the UK are closely related to where one lives (highest in London and the South-East) and occupation and industry (highest for higher-tier jobs in finance and related industries, for instance).[11] Partnership status and the presence of children are also relevant.

[10] Application of the methods of Yun (2004) as implemented by Jann (2008) provided very similar results for the decomposition into characteristics and returns components.
[11] See e.g. Bell and Van Reenen (2014) and Stewart (2011).

Although income taxation in the UK is assessed on an individual basis, couples may choose to split assets (and hence investment income) to gain tax advantages, and the labour supply of husbands and especially wives depends on the presence of dependent children.

Taking these various factors into account, the characteristics (X) we use are: age and age squared, age completed full time education and its square, a binary indicator for membership of a 'non-white' ethnic group, family type (six categories defined by whether the family head is of pension age, whether it is a single-adult or couple family and, for families in which the head is not of pension age, whether there are dependent children present), region of residence (London, South East, Rest of the UK), employment status (employee, self-employed, not in the labour force (NILF), unemployed), whether in part-time or full-time work (if working), occupation (five categories, itemised in Appendix B) and industry (nine categories, also itemised in Appendix B). In addition, we include binary indicators of whether the respondent's partner (if present) is a member of the top 1% income group or whether a member of the top 10% income group. These are intended to capture potential within-family portfolio choices that are particularly relevant at the top of income distribution. Finding both members of a couple as members of a top income group may reflect marital homogamy, but we cannot identify that pathway specifically.

More generally, and as is clear from the discussion above, our specification is a reduced-form one, providing a descriptive model rather than a causal one. We are summarizing associations and how these have changed over time in tandem with the rise in the top income rate among women. Nonetheless, our set of explanatory variables is similar to but somewhat more extensive than those used in the pooled-data regressions of Bobilev et al. (2020) and Yavorsky et al. (2019). Our specification is also similar to that used in the pooled-data regressions of Ravaska (2018), which is the only paper to date that includes covariates summarizing partner's income group membership.

The distributions of characteristics in 1999 and 2015 are summarised in Appendix Tables B1 and B2 for women and men separately and we also contrast the distributions for the top 1% and the bottom 99% income groups. There are some changes over the period that are common to both men and women; for example, there is population ageing and a doubling of the share of individuals who are non-white. Individuals became more likely to leave full-time education at a later age and – especially relevant for the decomposition reported shortly – this change is greater for women than for men. For example, the fraction of women in the top 1% leaving full-time education at age 22 or older increased from 34% to 44%. Among

women in the bottom 99%, the fraction increased from 6.5% to 13.5%. For men, the corresponding fractions are 36% to 42%, and 9 % to 14%.

Looking at labour force participation, occupation and industry, there are some well-known gender differences among the poorest 99%, though the contrasts for the top 1% members are less well-known because survey data cannot usually be used to look at this group reliably (though we can, for the reasons discussed earlier). The biggest change for both women and men in the top 1% is a large increase in the proportion who are employees rather than self-employed. (For women the employee fraction increased from 62% to 77%; for men, from 67% to 80%.) For both women and men, there are increases between 1999 and 2015 in the proportions working in professional and higher occupations, and in service industries. In both years, women in the top 1% are much more likely than men in the top 1% to be living in London but there is little trend over time. Among both women and men in the top 1%, the majority belong to a non-pensioner couple family (with or without children) but, for women, the fraction in this group increased from 64% to 70% (representing a shift from being single), whereas for men the fraction remained constant at around 83%. Although top-income women became more likely to have a partner, the chances of having a top-income partner did not change. In both 1999 and 2015, around one half of all women in the top 1% had a partner in the top 10% but only around 15%–20% of men did; around one-quarter of women in the top 1% had a partner also in the top 1% but only around 5% of men did.[12]

The logistic regression estimates underpinning our decomposition analysis are summarized in Tables B3–B6. Tables B3 and B4 show regression coefficients for women and men respectively; Tables B5 and B6 show the average partial effects (APEs) implied by the coefficients. To examine 'changes in returns' below we focus on the APEs rather than the raw coefficients. The numerator terms in decomposition equation (2) refer to counterfactual probabilities. Because we have a non-linear model, the raw coefficients cannot be straightforwardly compared with each other, but APEs can be when we calculate them using a fixed (common) year's distribution of characteristics (i.e. to be consistent with eq. (2).)

Tables B3 and B4 show that almost all groups of characteristics have statistically significant associations with the probability of top 1% group membership, for both years and

[12] The top-income gender differences reported here for education, partnered rather than single, and having a top-income partner, are broadly similar to those reported by Boschini et al. (2020) for Sweden (using register data) when we focus on the same time period. Also consistent with our UK estimates, Bobilev et al. state that '[t]o the extent that top women have a partner, they are more likely than men to have a high earning partner. These patterns are relatively stable over time' (2020: 86). Their finding is based on LIS data for Canada, Denmark, Finland, Norway and the USA – the only countries for which sample sizes are sufficiently large.

for women and men. For both men and women, the probability of belonging to the top 1% is higher for older people, for individuals who stayed in full-time education longer, who are working (and working full-time rather than part-time), working in higher level occupations and industries such as financial intermediation, living in London, and with a partner who belongs to the top 1%. Women belonging to a non-pensioner couple are less likely than childless single women to be in the top 1% whereas men belonging to a non-pensioner couple are more likely than single men to belong to the top 1%.[13] But what is important for the decomposition is the *changes* in the magnitude of returns (and how these differ for men and women), and we return to these shortly.

Table 3 presents the results of the components of our regression-based decomposition. Estimates for women are shown in panel A and for men in panel B. The left-hand side of the table shows the decomposition estimates for the case in which 1999 is used as the reference year; the right-hand side shows estimates for the case in which 2015 is used instead.

For *women*, the probability of being in the top 1% increased from 0.27% to 0.43% between 1999 and 2015. Of this 0.16 percentage point increase, virtually all of it – 0.15 percentage points – is accounted for by changes in women's observable characteristics, which leaves a contribution of 0.01 (0.16 – 0.15) percentage points attributable to changes in estimated returns to characteristics. These estimates are shown in italics at the foot of Panel A under the reference year 1999 heading. If 2015 is used as the reference year instead, the sizes of the decomposition components change hardly at all.

<Table 3 near here>

[13] Bobilev et al. (2020, Table 1) run pooled LIS-country-year linear-probability regressions, separately for men and women, for the probability of belonging to the top 1% (and of the probability of belong to the top 10%) with age, age-squared, education level, marital status, and number of children as the explanatory variables. They find, as we do, that being older and having more education each raise top-income group membership chances for both men and women, and that belonging to a couple and having children are each associated with lower chances for women but larger ones for men. Ravaska (2018, Table 2) runs pooled-data logit regressions for the probability of belonging to the top 10% in Finland separately for men and women and separately by partnership status for each sex. She finds, as we do, that for both sexes the chances of top income group membership are higher for those with more education, and those who work in finance, or in higher-level occupations. She also finds as we do that having a top-income partner is associated with higher chances of belonging to the top 1%, with the effect larger for men than for women – but our results are not fully comparable because she also controls for spousal occupation. Yavorsky et al. (2019, Table 3) run pooled-data logit regressions for the probability of an individual earning sufficient income on their own to be a member of the top 1% defined in terms of *household* income. Despite the difference in outcome variable definition, they find, like us, that higher chances of top-income group membership are associated with having more education, being self-employed, being older, and being white rather than non-white. Being married is associated with higher chances for men (as we find), but non-significant for women. Differently from us, Yavorsky et al. report that having children is positively associated with higher chances of top-income group membership for women as well as men. Boschini et al. (2020) do not report any regression estimates.

For *men*, the probability of being in the top 1% declined by 0.111 percentage points from 1.96% to 1.85% between 1999 and 2015. Changes in observable characteristics contributed to increase this probability by between 0.353 percentage points (if the reference year is 1999) and 0.527 percentage points (reference year 2015). In other words, given the changes in their characteristics, men should have seen a large increase in their probability of being in the top 1%. Instead, there was a decline because these changes were more than offset by the reduced returns to their characteristics, ranging between 0.638 and 0.464 percentage points depending on the base year. Clearly the factors accounting for the trend in top income group membership for men differ from those for women.

There is an important feature of the trend that is common to both men and women, however. Table 3 shows that for both sexes, changes in the distribution of education account for by far the largest proportion of the increase in top 1% group membership that is attributed to changes in observable characteristics as a whole. For women, it explains more than two-thirds of the increase (0.111/0.150 if 1999 is the reference year; 0.108/0.151 if it is 2015. For men, it explains nearly all the increase (0.422/0.527 versus 0.340/0.353). For women, it is only the estimate of the education variables' contribution to the total characteristics component that differs significantly from zero (estimate around 4 times larger than standard error). For men as well, the education variables' contribution is statistically significant (ratio of estimate to SE of at least 7.7) but so too is the contribution of the partner variables (ratio over 5).[14]

Propensities to remain longer in full-time education increased more for women than for men, as we pointed out earlier. Although the return to staying longer in education in terms of chances to get into the top 1% is much larger for men than for women in both years, this return hardly changed for men between 1999 and 2015 whereas it increased for women more noticeably. For women, the return to an extra year of education increased from 0.057 to 0.072 percentage points if evaluated using 1999 sample characteristics (0.055 to 0.067 percentage points with 2015 sample characteristics: see Appendix Table B5). In contrast, the return for men of an additional year of full-time education increased from 0.342 percentage points to 0.373 percentage points if evaluated using 1999 sample characteristics (0.280 to 0.301 percentage points with 2015 sample characteristics: see Appendix Table B6).

[14] Depending on the reference year, some other characteristic variable sets are statistically significant for men as well.

The large negative 'change in returns' component in the decomposition for men reflects several factors. For instance, Table B6 shows that for men the penalties – lower chances of top 1% group membership – grew for non-white people, for individuals not in paid work, and for individuals living in the Rest of UK rather than London. The same penalties exist for women and also increased but, in each case, the penalty for women is much smaller than the corresponding one for men in absolute magnitude (Table B5), and so too is the contribution to 'changes in returns'.

5. Summary and conclusions

The substantive contribution of this paper is the demonstration that the rising share of women in the top 1% of the UK income distribution is largely accounted for by women having increased the time they spend in full-time education by more than the increase for men. A minor supporting role is played by an increase for women in the return in terms of securing top income group membership from having longer education that is larger than the increase for men.

Rather than explaining the rise in the top income group membership rate for women by focusing on changes in the types of income that women hold – the approach taken by Atkinson et al. (2018) and Boschini et al. (2020), for example – we have examined the roles of more fundamental factors: individuals' characteristics and the returns to their characteristics in terms of chances of top-income group membership.

We are able to implement our decomposition approach because we use survey data which – by their very nature and in sharp contrast to most administrative record data sets outside the Nordic countries – include an extensive range of characteristics and provide full-population coverage. Although it is commonly argued that survey data cannot be used to reliably examine top incomes because of under-coverage problems, we have also shown that our UK survey data (the HBAI subfile of the Family Resources Survey) can reliably identify who belongs to the top 1% of the individual income distribution. Whether this finding carries over to other UK survey datasets or to survey data for other countries is a topic for further research.

6. References

Atkinson, A. B., Casarico, A., and Voitchovsky, S. (2018). 'Top incomes and the gender divide', *Journal of Economic Inequality* 16 (2): 225–256.

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

Belfield, C., Blundell, R., Cribb, J., Hood, A. and Joyce, R. (2017), 'Two decades of income inequality in Britain: the role of wages, household earnings and redistribution', *Economica*, 84: 157–79.

Bell, B. and Van Reenen, J. (2014). 'Bankers and their bonuses', *Economic Journal,* 124: 20–21.

Bobilev, R., Boschini, A., and Roine, J. (2020). 'Women at the top of the income distribution: what can we learn from LIS-data?', *Italian Economic Journal*, 6: 63–107.

Boschini, A., Gunnarsson, K., and Roine, J. (2020). 'Women in top incomes: evidence from Sweden 1974–2013', *Journal of Public Economics*, online ahead of print.

Brewer, M., Etheridge, B., and O'Dea, C. (2017). 'Why are households that report the lowest incomes so well-off?', *Economic Journal*, 127: F24–F49.

Brewer, M., Sibieta, L., and Wren-Lewis, L. (2017). 'Racing away? Income inequality and the evolution of high incomes', Briefing Note 76. London: Institute for Fiscal Studies.

Brewer, M. and Sámano-Robles, C. (2019). 'Top incomes in the UK: analysis of the 2015-16 Survey of Personal Incomes', ISER Working Paper 2019–06. Colchester: ISER, University of Essex.

Burkhauser, R. V., Hérault, N., Jenkins, S. P. and Wilkins, R. (2018*a*). 'Top incomes and inequality in the UK: reconciling estimates from household survey and tax return data', *Oxford Economic Papers*, 70 (2): 301–326.

Burkhauser, R. V., Hérault, N., Jenkins, S. P. and Wilkins, R. (2018*b*). 'Survey under-coverage of top incomes and estimation of inequality: what is the role of the UK's SPI adjustment?', *Fiscal Studies*, 39 (2): 213–240.

Department of Social Security (1996). *Households Below Average Income. Methodological Review. Report of a Working Group*. London: Department of Social Security.

Department for Work and Pensions (2017). *Households Below Average Income An Analysis of the UK Income Distribution 1994/95 – 2015/16*. London: Department for Work and Pensions.

Fairlie, R. W. (2005). 'An extension of the Blinder-Oaxaca decomposition technique to logit and probit models', *Journal of Economic and Social Measurement*, 30 (873): 305–316.

HM Revenue and Customs KAI Data, Policy and Co-ordination (2014), *Survey of Personal Incomes, 2010-2011: Public Use Tape* [computer file]. Colchester, Essex: UK Data Archive [distributor], November 2014. SN: 7569.

Jann, B. (2006). 'FAIRLIE: Stata module to generate nonlinear decomposition of binary outcome differentials', Statistical Software Components S456727, Boston College Department of Economics, revised 26 May 2008.

Jann, B. (2008). 'OAXACA: Stata module to compute the Blinder-Oaxaca decomposition', Statistical Software Components S456936, Boston College Department of Economics, revised 25 August 2011.

Jenkins, S. P. (2017), 'Pareto models, top incomes and recent trends in UK income inequality', *Economica*, 84: 261–89.

Joyce, R., Pope, T., and Roantree, B. (2019). 'The characteristics and incomes of the top 1%', Briefing Note BN 254. London: Institute for Fiscal Studies.

Lustig, N. (2019). 'Measuring the distribution of household income, consumption and wealth', Chapter 3 in: J. E. Stiglitz, J.-P. Fitoussi, and M. Durand (eds), *For Good Measure. Advancing Research on Well-Being Metrics beyond GDP*. Paris: OECD Publishing, 49–83.

McCune, B., Grace, J. B., and Urban, D. L. (2002). *Analysis of Ecological Communities*. Gleneden Beach, OR: MjM Software Design.

McLachlan, G. J. (1999). 'Mahalanobis distance', *Resonance*, June, 20–26.

Ravaska, T. (2018). 'Top incomes and income dynamics from a gender perspective: evidence from Finland 1995–2012'. ECINEQ Working Paper 2018–469.

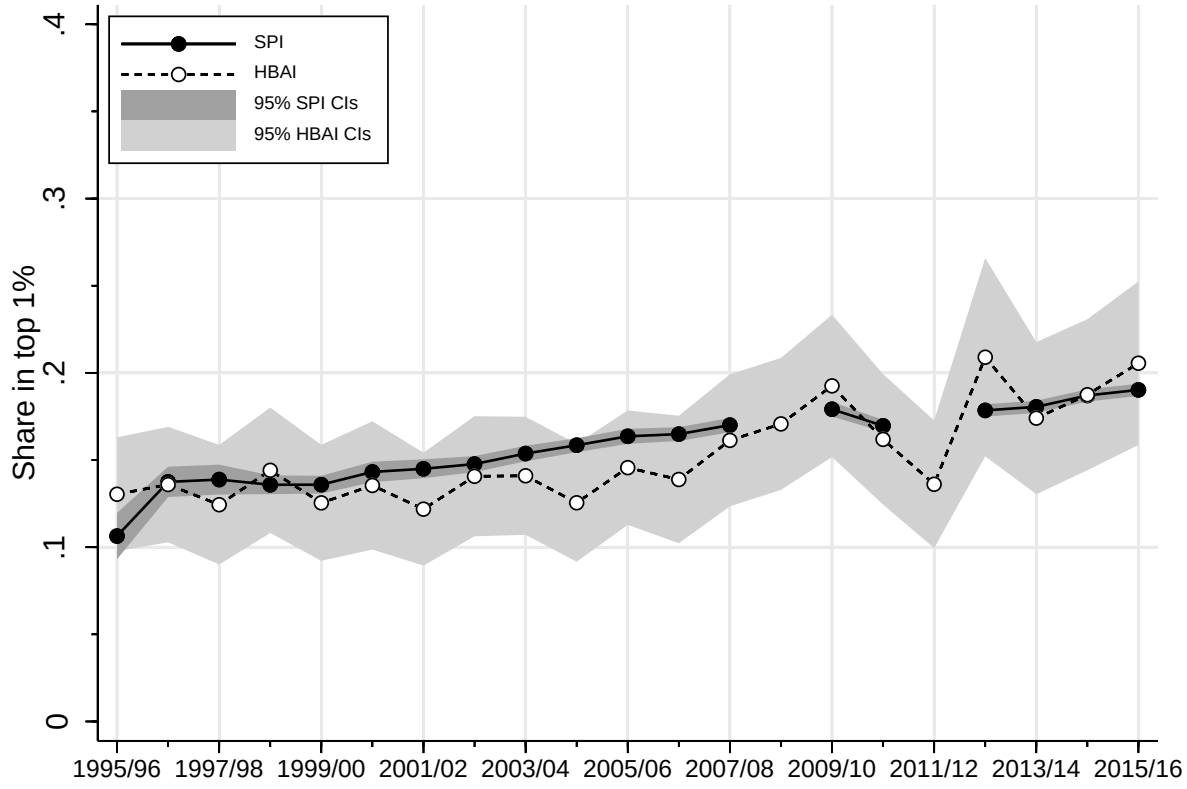
Stewart, M. B. (2011). 'The changing picture of earnings inequality in Britain and the role of regional and sectoral differences', *National Institute Economic Review* 218(1), 20–32.

Summers, A. (2019). 'The missing billions: measuring top incomes in the UK, Seminar presentation, 5 February 2019. London: LSE International Inequalities Institute. http://www.lse.ac.uk/International-Inequalities/Videos-Podcasts/Inequalities-Seminar-The-Missing-Billions-Measuring-Top-Incomes-in-the-UK

Yavorsky, J. E., Keister, L. A., Qian, Y., and Nau, M. (2019). 'Women in the one percent: gender dynamics in top income positions', *American Sociological Review*, 84(1): 54–81.

Yun, M.-S. (2004). ‘Decomposing differences in the first moment’, *Economics Letters*, 82: 275–280.

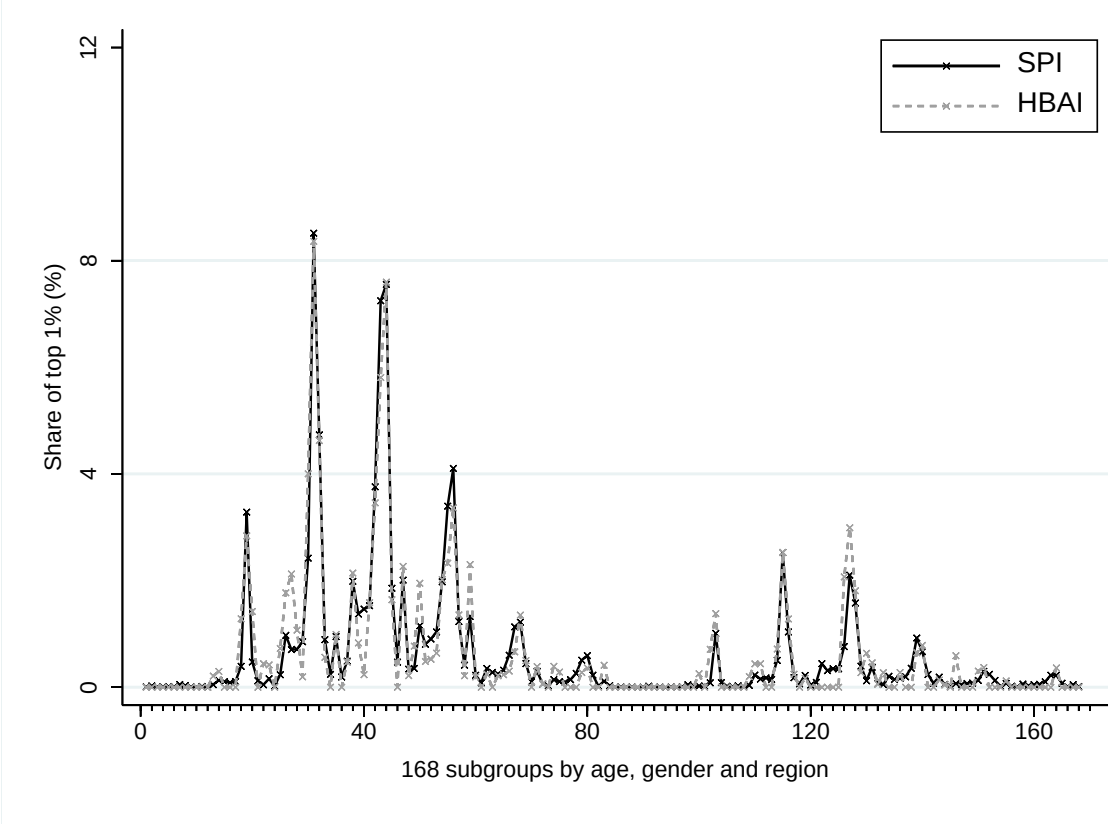
Figure 1. Share of women in the top 1% gross income group in survey and tax return data (1995/96 to 2015/16)



Note: The individual is the unit of analysis. Estimates are based on the adult population (aged 15 or above). The acronyms refer to the following data sources and series: (a) SPI, the Survey of Personal Incomes (income tax return data), not available in 2008/09 and 2011/12; (b) HBAI: the DWP's cleaned-up FRS.
Source: Authors' calculations based on FRS, HBAI, SPI and ONS data.

(*Stata graph sex_2_top100_SPIvsHBAI*)

Figure 2. Composition of the top 1% income group in the SPI and HBAI across the 168 possible combinations of age, gender and government office region categories (2015/16)



Notes: As for Figure 1.

(*Stata graph Top1_HBAIvsSPI_2015_cmean*)

Table 1*a*. Distribution of age and government office region in the top 1% in the survey and tax return data (1995/96 to 2015/16)

	1995		1996		1997		1998		1999		2000		2001		2002		2003	
	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	5,055	481	6,986	477	18,976	432	25,698	421	31,879	464	31,710	416	36,299	469	36,846	514	39,633	514
No. of weighted obs.	454,205	453,896	455,596	455,963	456,860	456,998	458,444	458,338	460,577	460,022	463,622	463,444	466,498	466,772	483,103	482,734	486,526	486,276
Age																		
Under 25	0.0	0.0	0.0	0.3	***0.2***	***0.8***	***0.2***	***0.7***	***0.2***	***0.9***	0.4	2.6	0.5	0.0	0.3	0.0	***0.3***	***0.3***
25 – 34	0.0	15.3	0.0	14.4	9.4	18.4	***10.9***	***10.0***	***10.4***	***11.5***	12.6	19.6	***11.5***	***14.4***	***10.4***	***11.9***	***9.0***	***11.1***
35 – 44	0.0	32.8	0.0	29.9	***30.8***	***28.0***	***31.5***	***31.7***	***31.2***	***34.6***	***33.7***	***32.8***	***34.7***	***38.6***	34.0	40.0	***33.1***	***35.7***
45 – 54	0.0	34.3	0.0	34.7	38.2	32.7	***35.6***	***35.9***	***36.1***	***34.5***	***33.0***	***30.9***	32.6	26.2	***33.0***	***30.1***	33.9	29.4
55 – 64	0.0	11.4	0.0	14.0	***14.8***	***14.5***	***15.2***	***14.8***	***15.3***	***13.8***	14.4	10.9	***15.1***	***14.6***	***16.4***	***14.1***	***17.4***	***18.3***
65 – 74	0.0	4.6	0.0	5.9	***4.3***	***4.8***	***4.7***	***5.5***	***4.8***	***3.6***	4.2	1.7	***3.9***	***4.7***	***4.2***	***3.3***	***4.4***	***4.8***
75 and over	0.0	1.6	0.0	0.7	2.4	0.7	***2.0***	***1.3***	2.1	1.0	***1.7***	***1.6***	***1.7***	***1.5***	1.8	0.5	1.9	0.3
Missing age	100.0		100.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Sex																		
Male	***86.5***	***86.7***	***86.8***	***87.3***	***86.4***	***87.7***	***86.7***	***86.4***	***86.8***	***88.0***	***86.1***	***86.4***	***85.8***	***86.8***	***85.8***	***86.9***	***85.2***	***86.0***
Female	***10.4***	***13.3***	***13.2***	***12.7***	***13.6***	***12.3***	***13.3***	***13.6***	***13.2***	***12.0***	***13.9***	***13.6***	***14.1***	***13.2***	***14.2***	***13.1***	***14.8***	***14.0***
Missing sex	3.2		0.0		0.0		0.0		0.0		0.0		0.1		0.0		0.0	
Government office region																		
North East	***2.2***	***1.8***	1.6	3.3	***1.8***	***2.1***	***1.7***	***2.0***	***1.6***	***1.8***	***1.4***	***1.3***	***1.5***	***1.7***	***1.6***	***1.2***	***1.8***	***1.1***
North West	***6.8***	***6.6***	***6.9***	***6.5***	***6.7***	***6.4***	6.8	4.7	***7.1***	***7.4***	***6.7***	***5.6***	***6.4***	***5.7***	***6.8***	***6.6***	***7.1***	***6.8***
Yorkshire and the Humber	***5.3***	***7.0***	***5.2***	***5.4***	***5.1***	***4.4***	***4.7***	***5.4***	***4.8***	***5.0***	***4.3***	***5.4***	***4.5***	***4.1***	***5.0***	***4.9***	5.1	3.4
East Midlands	***4.5***	***5.3***	***5.0***	***3.6***	***5.4***	***4.6***	***4.8***	***5.8***	***5.2***	***4.0***	***4.9***	***4.4***	4.8	2.9	5.2	3.3	***5.2***	***3.8***
West Midlands	***6.5***	***6.4***	***6.2***	***6.0***	5.1	8.8	***5.7***	***6.0***	5.7	8.9	5.7	3.4	5.5	3.1	***5.7***	***6.9***	***5.6***	***4.9***
East of England	***12.4***	***13.9***	***11.5***	***11.0***	***12.9***	***14.7***	***12.4***	***13.6***	***12.3***	***10.0***	***12.4***	***12.7***	***12.1***	***12.3***	***11.8***	***12.7***	***11.9***	***10.2***
London	***21.0***	***21.7***	***20.6***	***23.3***	***24.1***	***23.2***	***25.5***	***25.8***	***24.8***	***25.8***	***27.5***	***30.6***	***27.1***	***30.2***	***25.3***	***26.4***	***24.7***	***28.2***
South East	***22.5***	***24.0***	22.3	28.5	***24.8***	***22.9***	***24.7***	***24.2***	***25.1***	***22.4***	***24.4***	***24.4***	***25.0***	***28.4***	***24.2***	***24.3***	***23.5***	***26.3***
South West	***5.5***	***5.8***	***6.6***	***6.0***	***6.2***	***5.2***	***6.1***	***5.9***	***6.5***	***5.9***	5.9	4.0	***6.3***	***5.9***	***6.3***	***5.7***	***6.5***	***5.6***
Wales	***2.2***	***2.1***	***2.0***	***1.9***	***1.9***	***2.4***	***1.6***	***1.3***	***1.7***	***2.2***	***1.6***	***1.9***	***1.7***	***1.2***	***1.8***	***1.3***	***1.9***	***1.7***
Scotland	***5.7***	***5.5***	***5.5***	***4.4***	***5.3***	***5.2***	***5.4***	***5.4***	***5.3***	***6.8***	***5.0***	***6.4***	***5.1***	***4.4***	***5.2***	***6.0***	***5.5***	***6.8***
Northern Ireland	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0***	***1.2***	***0.8***	***1.4***	***1.0***
Pub. dpt/unknown	0.8		0.3		0.7		0.4		0.0		0.1		0.0		0.0		0.0	
Region missing	4.5		6.3		0.0		0.1		0.0		0.0		0.0		0.0		0.0	

Table 1*a*. continued

	2004		2005		2006		2007		2009		2010		2012		2013	
	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	42,701	483	40,044	526	41,398	438	42,168	431	43,060	443	47,287	450	47,097	333	47,043	357
No. of weighted obs.	489,747	489,962	494,898	494,945	499,431	499,159	503,975	502,784	512,565	511,863	516,998	516,428	524,866	524,617	527,119	527,514
Age																
Under 25	***0.5***	***0.4***	***0.3***	***0.5***	0.2	0.0	***0.2***	***1.8***	***0.2***	***0.6***	***0.2***	***0.7***	***0.2***	***0.7***	0.2	0.0
25 – 34	8.6	13.2	8.5	12.5	***8.5***	***11.1***	***8.6***	***10.3***	***7.5***	***9.8***	***7.3***	***8.5***	***6.8***	***7.2***	***6.4***	***8.7***
35 – 44	***33.7***	***33.3***	***32.6***	***36.0***	31.3	36.4	***31.0***	***35.5***	***29.6***	***32.2***	***29.7***	***30.6***	***28.3***	***34.1***	***27.8***	***25.1***
45 – 54	***33.7***	***32.7***	35.2	29.2	36.4	29.6	36.7	30.1	***37.2***	***37.3***	***35.9***	***34.4***	***36.6***	***35.9***	***36.9***	***40.0***
55 – 64	***17.4***	***16.0***	***17.3***	***18.1***	***17.5***	***17.0***	***17.0***	***17.0***	18.0	14.2	***18.6***	***20.3***	***18.9***	***15.0***	***19.3***	***19.2***
65 – 74	***4.3***	***3.0***	***4.2***	***3.0***	***4.2***	***3.8***	***4.5***	***4.0***	***5.3***	***3.7***	5.7	3.8	***6.2***	***5.5***	6.3	4.2
75 and over	***1.9***	***1.3***	1.9	0.7	***2.0***	***2.1***	***2.1***	***1.4***	***2.3***	***2.1***	***2.6***	***1.8***	3.0	1.6	***3.0***	***2.8***
Missing age	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Sex																
Male	***84.8***	***87.7***	***84.3***	***85.1***	***84.2***	***85.8***	***83.6***	***84.1***	***82.8***	***81.9***	***83.0***	***83.8***	***82.1***	***79.1***	***81.9***	***82.6***
Female	***15.2***	***12.3***	***15.7***	***14.9***	***15.8***	***14.2***	***16.4***	***15.9***	***17.2***	***18.1***	***17.0***	***16.2***	***17.8***	***20.9***	***18.1***	***17.4***
Missing sex	0.0		0.0		0.0		0.0		0.0		0.0		0.0		0.0	
Government office region																
North East	***1.8***	***2.6***	***1.9***	***1.1***	***1.9***	***1.5***	***1.7***	***1.4***	***1.7***	***1.6***	***1.7***	***1.1***	***1.6***	***2.5***	***1.5***	***1.1***
North West	***7.1***	***6.6***	***7.0***	***5.6***	***6.8***	***8.0***	***6.6***	***7.1***	***6.6***	***5.2***	***6.3***	***5.7***	***6.1***	***4.9***	***5.8***	***4.2***
Yorkshire and the Humber	***5.1***	***4.4***	5.0	3.2	***5.1***	***5.1***	***4.7***	***4.0***	***4.8***	***4.7***	***4.5***	***3.6***	***4.3***	***3.4***	***4.1***	***4.0***
East Midlands	5.1	3.6	***4.8***	***5.2***	4.7	2.8	***4.5***	***3.2***	***4.7***	***5.1***	***4.6***	***3.4***	***4.5***	***2.9***	***4.4***	***3.9***
West Midlands	***5.6***	***4.3***	***5.5***	***6.0***	***5.6***	***5.2***	***5.4***	***3.8***	***5.3***	***4.2***	***5.1***	***3.6***	***4.9***	***3.7***	4.7	2.8
East of England	***11.6***	***11.8***	***11.7***	***9.8***	***11.8***	***10.0***	***11.4***	***9.9***	***11.4***	***13.5***	***11.6***	***13.8***	***11.4***	***12.0***	***11.2***	***11.8***
London	25.8	32.2	25.7	30.3	***26.1***	***30.6***	***27.7***	***29.8***	***27.8***	***25.7***	***28.4***	***30.5***	***29.1***	***34.5***	***29.3***	***27.3***
South East	***22.7***	***22.0***	***22.4***	***21.9***	***22.4***	***21.1***	***22.3***	***23.6***	***22.5***	***24.3***	***22.6***	***24.4***	***23.1***	***23.1***	22.2	27.3
South West	***6.2***	***5.6***	***6.3***	***6.8***	***6.2***	***6.7***	***6.2***	***5.8***	***6.0***	***4.3***	***6.0***	***4.6***	***5.9***	***4.1***	***5.8***	***7.4***
Wales	***2.0***	***1.5***	***2.1***	***1.7***	***1.9***	***3.3***	***1.8***	***3.1***	***2.0***	***2.3***	***1.9***	***1.7***	***1.8***	***1.6***	***1.7***	***2.2***
Scotland	***5.6***	***4.4***	***6.0***	***7.1***	***6.0***	***4.7***	***5.9***	***7.2***	***5.9***	***7.8***	***6.1***	***6.4***	***6.2***	***6.5***	***6.2***	***6.8***
Northern Ireland	***1.4***	***0.9***	***1.5***	***1.2***	1.5	1.0	***1.6***	***1.2***	***1.3***	***1.2***	***1.3***	***1.4***	***1.2***	***0.8***	***1.1***	***1.0***
Pub. dpt/unknown	0.0		0.0		0.0		0.0		0.0		0.0		0.1		2.0	
Region missing	0.0		0.0		0.0		0.2		0.0		0.0		0.0		0.0	

Table 1*a*. continued

	2014		2015	
	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	48,917	347	49,077	335
No. of weighted obs.	531,859	531,438	535,590	534,202
Age				
Under 25	***0.2***	***1.0***	0.2	0.0
25 – 34	***6.4***	***6.6***	***6.3***	***9.3***
35 – 44	***27.3***	***32.4***	***26.7***	***30.5***
45 – 54	***37.1***	***35.9***	***37.0***	***34.6***
55 – 64	***19.6***	***16.9***	***20.2***	***18.3***
65 – 74	***6.3***	***5.9***	***6.3***	***5.3***
75 and over	3.1	1.3	***3.2***	***2.1***
Missing age	0.0		0.0	
Sex				
Male	***81.3***	***81.3***	***81.0***	***79.4***
Female	***18.7***	***18.7***	***19.0***	***20.6***
Missing sex	0.0		0.0	
Government office region				
North East	***1.5***	***2.3***	***1.4***	***2.7***
North West	***6.0***	***7.0***	***5.8***	***7.8***
Yorkshire and the Humber	***4.2***	***3.8***	***4.1***	***4.2***
East Midlands	4.5	2.6	4.3	2.6
West Midlands	***4.8***	***6.4***	4.8	2.7
East of England	***11.4***	***11.0***	***11.3***	***14.9***
London	30.1	21.3	***31.1***	***28.2***
South East	***22.4***	***27.0***	***22.4***	***23.0***
South West	***6.0***	***7.6***	***5.8***	***4.6***
Wales	***1.6***	***2.9***	***1.6***	***1.3***
Scotland	***6.4***	***7.0***	***5.8***	***7.7***
Northern Ireland	***1.1***	***1.2***	1.1	0.6
Pub. dpt/unknown	0.0		0.4	
Missing gor code	0.0		0.1	

Note: Bold italicised font indicates that the SPI estimate falls within the 95% confidence interval of the HBAI estimate. Northern Ireland is included in the survey-based series only from 2002/03 onwards. The individual is the unit of analysis. Estimates are based on the adult population (aged 15 or above). The acronyms refer to the following data sources and series:

- SPI: the Survey of Personal Incomes (income tax return data), not available in 2008/09 and 2011/12.
- HBAI: the DWP's cleaned-up FRS.

Source: Authors' calculations based on FRS, HBAI, SPI and ONS data.

Table 1*b*. Distribution of industry in the top 1% in the survey and tax return data (1995/96 to 2015/16)

	1995		1996		1997		1998		1999		2000		2001		2002		2003	
	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	0	394	0	377	14,717	342	19,923	324	24,804	374	25,105	345	28,450	369	28,344	417	30,073	391
No. of weighted obs.	0	374,409	0	360,218	358,047	361,650	357,181	356,167	357,728	371,074	367,683	385,723	367,504	369,500	373,649	396,146	369,893	370,834
Industry																		
Agriculture, forestry and fishing	0.0	1.6	0.0	1.1	1.5	0.0	***1.1***	***0.6***	***0.7***	***1.0***	***0.5***	***1.2***	***0.7***	***0.6***	***0.6***	***1.3***	***0.7***	***0.4***
Mining & quarrying	0.0	2.4	0.0	0.9	1.0	0.0	***0.7***	***1.8***	***0.7***	***1.4***	***0.8***	***2.0***	***0.7***	***1.4***	***0.6***	***1.5***	0.6	2.0
Manufacturing	0.0	13.6	0.0	18.4	11.4	0.0	11.0	17.1	10.9	18.2	9.7	16.3	9.1	12.9	8.2	14.7	8.2	11.6
Electricity, Gas and Water supply	0.0	0.7	0.0	1.0	0.6	0.0	***0.5***	***0.6***	***0.5***	***0.6***	***0.5***	***0.3***	***0.4***	***1.2***	0.3	1.5	0.2	1.6
Construction	0.0	6.5	0.0	3.7	2.8	0.0	***3.2***	***3.2***	***3.2***	***4.3***	3.1	6.3	***3.6***	***2.6***	***3.8***	***3.8***	***4.3***	***6.2***
Wholesale and retail trade	0.0	11.6	0.0	10.0	12.3	0.0	***11.1***	***9.7***	10.7	4.0	10.0	4.7	10.1	5.8	9.5	5.7	9.6	4.9
Hotels and restaurants	0.0	1.2	0.0	2.0	1.5	0.0	***1.0***	***1.4***	***1.0***	***0.9***	***0.9***	***0.5***	***1.0***	***1.1***	1.1	0.5	1.0	0.0
Transport, storage and communications	0.0	5.5	0.0	5.4	4.1	0.0	***3.9***	***4.3***	***4.1***	***4.7***	***3.6***	***5.6***	***4.1***	***5.3***	***3.5***	***4.7***	3.5	7.1
Financial intermediation	0.0	13.2	0.0	14.9	15.2	0.0	16.9	11.4	17.0	11.6	***18.0***	***18.7***	***18.0***	***18.0***	***17.0***	***18.0***	16.6	22.2
Real estate, renting & business	0.0	22.5	0.0	22.5	31.5	0.0	***30.0***	***31.0***	30.6	36.1	***32.2***	***30.4***	***30.8***	***34.6***	***29.3***	***33.9***	28.8	22.8
Public administration and defence	0.0	4.8	0.0	5.4	1.0	0.0	0.9	4.1	***0.6***	***1.7***	0.5	2.2	0.4	4.1	0.5	2.0	0.4	3.2
Education	0.0	1.3	0.0	2.9	1.2	0.0	***1.1***	***2.3***	***1.2***	***1.7***	***1.2***	***1.1***	***1.2***	***2.2***	***1.0***	***1.0***	***1.2***	***1.6***
Health and social work	0.0	9.8	0.0	9.0	7.5	0.0	***6.4***	***7.4***	***7.4***	***9.5***	***7.0***	***5.4***	***7.4***	***6.5***	***8.4***	***7.6***	***9.7***	***10.0***
Other services (community, personal)	0.0	4.3	0.0	2.1	2.9	0.0	***3.1***	***4.3***	***3.2***	***3.4***	***3.0***	***3.9***	***3.0***	***2.8***	***3.1***	***2.5***	***3.0***	***2.1***
Others	0.0	0.0	0.0	0.3	0.6	0.0	0.7	0.0	1.9	0.3	0.3	0.0	0.4	0.0	0.5	0.0	0.5	0.0
Missing/Not applicable	100.0	1.0	100.0	0.6	5.1	100.0	8.3	0.7	6.3	0.8	8.4	1.6	9.3	1.1	12.5	1.4	11.7	4.3

Table 1*b*, continued

	2004		2005		2006		2007		2009		2010		2012		2013	
	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	32,282	376	30,359	401	31,324	328	31,876	325	31,585	346	34,618	321	33,870	246	33,620	257
No. of weighted obs.	372,391	388,361	377,419	384,374	380,361	385,135	384,130	381,795	380,411	406,253	377,006	379,701	376,257	405,125	375,032	389,319
Industry																
Agriculture, forestry and fishing	***0.8***	***0.3***	0.7	0.1	0.6	0.0	***0.7***	***0.3***	***0.5***	***0.7***	0.5	0.0	0.5	0.0	***0.4***	***1.5***
Mining & quarrying	***0.6***	***0.7***	***0.6***	***1.1***	***0.7***	***1.5***	0.6	2.4	0.8	2.6	0.8	2.6	***1.0***	***2.5***	***1.1***	***2.9***
Manufacturing	8.3	14.6	7.4	12.6	***7.2***	***9.4***	***7.2***	***8.7***	***5.9***	***4.6***	***5.6***	***7.0***	***5.7***	***5.5***	***5.6***	***7.9***
Electricity, Gas and Water supply	***0.2***	***0.3***	***0.3***	***0.8***	***0.4***	***1.1***	***0.4***	***0.6***	***0.8***	***0.4***	***0.7***	***0.8***	0.7	0.0	***0.7***	***0.9***
Construction	***4.0***	***5.3***	***4.0***	***5.3***	***3.8***	***3.2***	***4.1***	***5.5***	***3.9***	***6.5***	***3.1***	***5.6***	2.9	9.2	***2.9***	***3.8***
Wholesale and retail trade	8.7	5.0	8.6	3.8	9.0	4.8	***8.7***	***8.3***	9.4	5.4	***8.8***	***10.2***	8.6	3.4	***9.0***	***6.5***
Hotels and restaurants	***1.0***	***0.9***	***1.0***	***0.7***	***0.9***	***2.0***	***0.9***	***0.6***	***0.9***	***1.2***	***0.8***	***1.5***	***0.8***	***1.4***	***0.7***	***0.8***
Transport, storage and communications	3.3	6.2	***3.7***	***4.4***	***3.5***	***2.5***	***3.5***	***5.7***	***10.2***	***9.5***	***10.8***	***7.8***	***10.8***	***10.5***	***11.0***	***13.3***
Financial intermediation	***17.0***	***20.3***	***19.0***	***22.5***	***21.6***	***25.2***	***22.0***	***17.5***	***22.1***	***20.6***	***23.2***	***20.2***	***23.9***	***19.1***	24.4	16.7
Real estate, renting & business	***27.4***	***26.0***	***28.2***	***26.4***	***30.9***	***29.7***	***31.3***	***31.2***	***23.3***	***22.6***	***23.7***	***23.2***	***23.7***	***23.2***	***25.1***	***23.3***
Public administration and defence	***0.7***	***1.5***	***0.9***	***2.5***	0.9	3.7	0.8	2.3	1.0	4.3	0.9	5.5	0.7	5.3	0.6	3.6
Education	***1.3***	***1.3***	***1.5***	***1.7***	***1.4***	***1.0***	***1.4***	***1.5***	***1.5***	***2.8***	***1.5***	***1.0***	***1.4***	***1.1***	***1.3***	***4.1***
Health and social work	***11.9***	***10.3***	***12.2***	***11.6***	***11.3***	***10.7***	***10.2***	***9.5***	***11.2***	***11.0***	11.8	7.9	***10.0***	***11.0***	***9.4***	***7.3***
Other services (community, personal)	***3.3***	***1.9***	***3.0***	***2.5***	***3.0***	***3.6***	***3.4***	***4.0***	***2.2***	***3.1***	***2.1***	***1.6***	2.0	0.5	***2.1***	***3.1***
Others	0.5	0.0	0.6	0.0	***0.6***	***0.2***	0.6	0.0	0.0	0.0	***0.0***	***0.0***	0.0	0.0	0.0	0.0
Missing/Not applicable	11.1	5.6	8.3	4.1	4.5	1.5	4.0	1.9	***6.5***	***4.6***	***5.8***	***5.0***	***7.3***	***7.2***	***5.6***	***4.0***

Table 1*b*, continued

	2014		2015	
	SPI	HBAI	SPI	HBAI
No. of unweighted obs.	34,760	254	34,523	237
No. of weighted obs.	376,320	398,297	375,229	396,743
Industry				
Agriculture, forestry and fishing	***0.3***	***0.6***	0.3	0.0
Mining & quarrying	***1.1***	***1.7***	***0.9***	***2.6***
Manufacturing	***5.3***	***9.0***	***5.0***	***8.2***
Electricity, Gas and Water supply	***0.7***	***1.2***	0.6	0.0
Construction	***3.1***	***4.8***	***3.5***	***6.4***
Wholesale and retail trade	***8.9***	***6.2***	***8.4***	***7.7***
Hotels and restaurants	***0.9***	***0.9***	***0.9***	***0.9***
Transport, storage and communications	***11.3***	***12.8***	***11.4***	***11.9***
Financial intermediation	***24.3***	***22.5***	***24.3***	***24.3***
Real estate, renting & business	***25.7***	***21.9***	***27.1***	***25.2***
Public administration and defence	0.6	3.1	***0.5***	***1.7***
Education	***1.3***	***2.7***	***1.1***	***2.3***
Health and social work	***9.2***	***8.2***	***8.3***	***6.5***
Other services (community, personal)	***2.0***	***1.9***	***2.1***	***1.4***
Others	0.0	0.0	0.0	0.0
Missing/Not applicable	5.2	2.6	5.5	1.0

Note: Bold italicised font indicates that the SPI estimate falls within the 95% confidence interval of the HBAI estimate. The individual is the unit of analysis. Estimates are based on the working-age population (25 to 54). The acronyms refer to the following data sources and series:

- SPI: the Survey of Personal Incomes (income tax return data), not available in 2008/09 and 2011/12 and no industry information prior to 1997/98.
- HBAI: the DWP's cleaned-up FRS.

Source: Authors' calculations based on FRS, HBAI, SPI and ONS data.

Table 2. Distance between the HBAI top 1% and the SPI top 10 percentile income groups (Panel A) and between the SPI top 1% and the HBAI top 10 percentile income groups (Panel B): 1997/98–2015/16 averages and bootstrap standard errors

Panel A	Distance between HBAI top 1% and SPI percentile income group									
	91	92	93	94	95	96	97	98	99	100
Overlapping distance	36.5	35.4	34.3	32.5	31.1	29.6	27.1	25.2	22.6	19.0
	(2.5)	(2.2)	(2.4)	(2.4)	(2.3)	(2.3)	(2.4)	(2.3)	(2.2)	(2.3)
Panel B	Distance between SPI top 1% and HBAI percentile income group									
	91	92	93	94	95	96	97	98	99	100
Overlapping distance	36.6	35.1	33.0	31.9	30.9	28.7	26.4	23.8	22.1	19.0
	(3.6)	(3.8)	(3.5)	(3.2)	(4.3)	(3.8)	(3.1)	(3.0)	(3.2)	(2.3)

Notes: 1997/98 to 2015/16 averages. Bootstrap standard errors in parentheses (100 bootstraps). All distances are multiplied by 100. The top percentile group (100) excludes the top 0.1% prior to 2010/11 (due to the presence of composite records in the SPI). The distances are based on all possible combinations of gender (2 categories), age (7 categories) and region (12 categories), resulting in 168 dummy variables.
Source: authors' calculations based on FRS, HBAI, SPI and ONS data.

Table 3. Decomposition of the change in the probability of being in the top 1% (in %)

	Base year 1999		Base year 2015	
Panel A – Women				
Probability of being in top 1%	0.27		0.43	
	Contribution (ppt)	(SE)	Contribution (ppt)	SE
Demographics	0.015	(0.032)	0.030	(0.025)
Family	–0.029	(0.041)	0.003	(0.017)
Education	0.111	(0.029)	0.108	(0.025)
Region	0.002	(0.011)	0.001	(0.009)
Employment	–0.004	(0.030)	–0.009	(0.022)
Occupation	0.054	(0.037)	–0.001	(0.030)
Industry	0.005	(0.033)	0.021	(0.028)
Partner	–0.005	(0.029)	–0.003	(0.023)
Changes in characteristics (all)	*0.150*		*0.151*	
Changes in returns	*0.010*		*0.009*	
Total change (ppt)	*0.160*		*0.160*	
Panel B – Men				
Probability of being in top 1%	1.96		1.85	
	Contribution (ppt)	SE	Contribution (ppt)	SE
Demographics	0.043	(0.044)	0.072	(0.044)
Family	0.062	(0.027)	0.048	(0.027)
Education	0.422	(0.042)	0.340	(0.044)
Region	0.043	(0.018)	0.023	(0.015)
Employment	–0.083	(0.030)	–0.044	(0.028)
Occupation	–0.070	(0.041)	–0.229	(0.043)
Industry	0.056	(0.034)	0.085	(0.031)
Partner	0.054	(0.010)	0.059	(0.011)
Changes in characteristics (all)	*0.527*		*0.353*	
Changes in returns	*–0.638*		*–0.464*	
Total change (ppt)	*–0.111*		*–0.111*	

Notes: Oaxaca-Blinder decomposition extended to a logit model by Fairlie (2005): see main text. Based on 3-year pooled samples ('1999' refers to 1998/99–2000/01 and '2015' to 2014/15–2016/17). Demographics includes age and age squared and a dummy for non-white; Family is family type (6 categories depending on whether head is of pension age, number of adults (one or two), and whether children present); Education is age completed full time education (and its square); Region is the region of residence (London, South East, Rest of the UK); Employment includes employment status (employee, self-employed, NILF or unemployed), whether working part-time, job occupation (5 categories) and job industry (9 categories); Partner includes binary indicators for whether respondent's partner belongs to top 10% or to top 1% income group. See Appendix Tables B1 and B2 for sample distributions and Appendix Tables B3–B6 for coefficient estimates and average partial effects. *Source*: Authors' calculations based on FRS, HBAI and ONS data.

Appendices

A. Measures of distance between a pair of discrete multivariate distributions

B. Decomposition: sample distributions, estimated coefficients and average partial effects

Appendix A. Measures of distance between a pair of discrete multivariate distributions

We observe *k* variables $(P_1, \dots, P_k)$ and $(Q_1, \dots, Q_k)$ in datasets *P* and *Q*, respectively.[15] We are interested in producing a scalar that summarises the distance between these two datasets or the populations they represent. Given the nature of the data on non-income characteristics available, we restrict discussion to the case where all variables in *P* and *Q* are discrete rather than continuous.[16] They can include binary and categorical variables. Assume also that we are interested in comparing two population groups of equal (weighted) size. For instance, with *P* for HBAI data and *Q* for SPI data, we can use sample weights (or grossing-up factors) and a common population control total to define the top income 1 per cent income group, which will be of the same size in both datasets. Denote *N* the number of (weighted) individuals in both *P* and *Q*.

We now have two matrices *P* and *Q* of the same size (*N x k*), with each row representing one individual. One possible distance measure is the (weighted) number of individuals that are observationally the same in *P* and *Q*. This count divided by *N* is the 'proportional overlap', where one minus this quantity provides a distance measure we term 'overlapping distance', *D(P,Q)*, also known as the simple matching coefficient. We show below that it is directly related to L1 Euclidean distance, also known as the Manhattan or city block distance.

With no loss of generality assume *P* and *Q* only includes dummy variables (categorical variables can always be broken down into dummy variables). We can transform *P* and *Q* into vectors *p* and *q* of size *2k*, with each element i being the proportion of individuals with unique possible combination i of variables *X*.

$p = (p_1, \dots, p_{2k})$ and $q = (q_1, \dots, q_{2k})$ with $\sum_{i=1}^{2k} q_i = 1$ and $\sum_{i=1}^{2k} p_i = 1$

The overlapping distance is then:

$$D(P,Q) = D(p,q) = 1 - \sum_{i=1}^{2k} \min\,(p_i, q_i) = \frac{1}{2}\sum_{i=1}^{2k} |p_i - q_i|$$

We can now compare population groups of different size because each group has in effect be rescaled to one, or we can compare one group to the population average.

$D(p,q)$ is bounded between 0 and 1. $D(p,q)$ increases with *k* and $D \to 1$ as $k \to \infty$. Importantly, $D(p,q)$ cannot increase as *k* decreases. $D(p,q)$ is one minus the relative

[15] With no loss of generality, *P* and *Q* can also be subsets of the same dataset.
[16] Continuous variables can easily be turned into categorical variables.

Sørensen (or Bray-Curtis) coefficient (McCune et al. 2002) and it can be shown that $D(p,q)$ is also half the Manhattan distance, which is defined as $L1(p,q) = \sum_{i=1}^{2k}|p_i - q_i|$.[17] The overlapping distance measure satisfies the Triangle Inequality (the distance between distributions *A* and *C* is no larger than the distance between *A*, *B* plus distance between *B*, *C*), Symmetry (the distance is the same if you treat HBAI as *P* and SPI as *Q* or vice versa) and Identity (distance equals 0 if $p = q$).

All elements of *p* and *q* are on the same scale. There are other distance metrics, including the Euclidean and Hellinger metrics discussed next. They also satisfy the Triangle Inequality, Symmetry, and Identity properties.

Euclidean distance

Euclidean distance *E*(*p*, *q*) is the square root of the sum of squared differences in discrete density:

$$E(p,q) = \sqrt{\sum_{i=1}^{2k}\left(p_{i,} - q_i\right)^2}\,.$$

Hellinger distance

The Hellinger distance *H*(*p*, *q*) is:

$$H(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{2k}\left(\sqrt{p_{i,}} - \sqrt{q_i}\right)^2}\,.$$

H and *E* differ from *D* because they assign a different weight to each difference $\left(p_{i,} - q_i\right)$ using a non-linear transformation. The Euclidean distance assigns more weight to large differences by taking the square of $\left(p_{i,} - q_i\right)$. The Hellinger distance first takes the square root of each element of *p* and *q*, thereby giving relatively more weight to combinations of variables with small probabilities.

Two further measures – which, furthermore, can also be applied when *P* and *Q* contain continuous variables – are the standardised Euclidean distance and Mahalanobis distance.

[17] Replace $\min(p_i, q_i)$ by $(p_i + q_i - |p_i - q_i|)/2$ in $D(p,q)$ and use $\sum_{i=1}^{2k} q_i = \sum_{i=1}^{2k} p_i = 1$.

These measures are derived by taking the sample mean of each variable: $\bar{P} = (\bar{P}_1, \ldots, \bar{P}_k)$ and $\bar{Q} = (\bar{Q}_1, \ldots, \bar{Q}_k)$.[18]

To address the loss of information on the relationship between variables (i.e., the covariance matrix) that this process entails, and to address the issue that not all elements of $\bar{P}$ and $\bar{Q}$ are necessarily on the same scale, both measures 'standardise' the differences between each element of $\bar{P}$ and $\bar{Q}$, but in different ways.

Standardised Euclidean distance

This is an adaptation of the Euclidean distance that adjusts for the variance of each variable. The intuition is that differences in variables with high variance should count less than differences in variables with low variance. The standardised Euclidean distance is:

$$SE(\bar{P}, \bar{Q}) = \sqrt{\sum_{i=1}^{k} \frac{(\bar{P}_i - \bar{Q}_i)^2}{s_i^2}}$$

where s_i is the standard deviation of the p_i and q_i over the sample set. $SE(\bar{P}, \bar{Q})$ is positive but unbounded. It is zero for a pair of identical distributions of characteristics.

Mahalanobis distance

The Mahalanobis distance is:

$$M(\bar{P}, \bar{Q}) = \sqrt{(\bar{P} - \bar{Q})' \Sigma^{-1} (\bar{P} - \bar{Q})}$$

where Σ is the common covariance matrix (McLachlan 1999: 24). The Mahalanobis distance accounts for each variable's variance (as the standardised Euclidean distance) and for the correlation between variables. The Mahalanobis distance limits the extent to which the presence of two highly correlated variables leads to counting the same information twice. Hence, the Mahalanobis distance is less sensitive to the addition of a new variable because it accounts for the information from the new variables that is already embedded in other (pre-included) variables.

The Mahalanobis distance is bounded below by 0 but does not have an upper bound. Replacing Σ by the identity matrix leads to the Euclidean distance. The standardised

[18] A correspondence with the data transformation described above is obtained when *P* and *Q* are only made of discrete variables. *P* and *Q* can then be expanded so that each element is an indicator of a unique combination of variables. In this case, $\bar{P}$ and $\bar{Q}$ contain *2k* elements and $\bar{P} = p$ and $\bar{Q} = q$.

Euclidean distance is obtained by using only the diagonal of Σ (and setting other elements to 0).

The Mahalanobis and standardised Euclidean distance measures, which use the inverse of the variance of elements of p and q, may be inappropriate to use with binary variables. With binary variables the variance tends to 0 as the probability of a positive outcome tends to 0. As the variance appears in the denominator in these distance measures, this implies that the combinations of variables with the smallest probabilities tend to have disproportionately large weights.

Another way to reduce the loss of information due to the use of averages is to first expand the matrices P and Q to add interactions between variables. At one extreme, all variables can be fully interacted. Because of the presence of continuous variable(s), a scaling issue will remain so that a 'standardisation' distance may still be desirable. But having more interactions implies having a greater number of dummy variables, thereby potentially making analysis infeasible.

Table A1. Distance between the HBAI top 1% and the SPI top 10 percentile income groups (Panel A) and between the SPI top 1% and the HBAI top 10 percentile income groups (Panel B) (1997/98-2015/16 averages and bootstrapped standard errors)

Panel A	Distance between HBAI top 1% and SPI percentile income group									
	91	92	93	94	95	96	97	98	99	100
Overlapping distance	36.5	35.4	34.3	32.5	31.1	29.6	27.1	25.2	22.6	19
	(2.5)	(2.2)	(2.4)	(2.4)	(2.3)	(2.3)	(2.4)	(2.3)	(2.2)	(2.3)
Euclidean distance	12	11.6	11.4	10.9	10.4	9.9	9.0	8.2	7.1	5.6
	(1.2)	(1.2)	(1.3)	(1.1)	(1.1)	(1.1)	(1.2)	(1.2)	(1.0)	(1.0)
Standardised Euclidean distance	23.6	21.5	21.3	19.1	17.7	15.8	13.2	11.2	9.2	8.5
	(5.1)	(4.1)	(4.9)	(4.2)	(4.8)	(3.9)	(4.2)	(3.9)	(3.5)	(3.4)
Hellinger distance	21.7	20.7	19.5	18.2	17	15.7	14.1	12.7	11.0	8.8
	(2.4)	(2.2)	(2.4)	(2.3)	(2.2)	(2.3)	(2.1)	(2.0)	(1.7)	(1.7)
Mahalanobis distance	23.6	21.5	21.3	19.1	17.7	15.8	13.2	11.2	9.2	8.5
	(5.1)	(4.1)	(4.9)	(4.1)	(4.8)	(3.9)	(4.2)	(3.9)	(3.5)	(3.4)
Panel B	Distance between SPI top 1% and HBAI percentile income group									
	91	92	93	94	95	96	97	98	99	100
Overlapping distance	36.6	35.1	33.0	31.9	30.9	28.7	26.4	23.8	22.1	19
	(3.6)	(3.8)	(3.5)	(3.2)	(4.3)	(3.8)	(3.1)	(3.0)	(3.2)	(2.3)
Euclidean distance	11.2	10.7	10.0	9.7	9.3	8.6	7.9	7.0	6.7	5.6
	(1.2)	(1.1)	(1.1)	(1.2)	(1.4)	(1.3)	(1.0)	(1.0)	(1.3)	(1.0)
Standardised Euclidean distance	24.2	23	20.8	19.5	18.4	16.6	15.6	13.4	11.0	8.5
	(4.1)	(4.2)	(4.0)	(4.2)	(4.5)	(3.0)	(4.4)	(3.7)	(3.6)	(3.4)
Hellinger distance	16.8	15.4	13.9	13.9	12.9	11.9	10.7	9.5	8.9	8.8
	(3.2)	(3.4)	(2.9)	(3.2)	(3.4)	(2.8)	(2.5)	(2.0)	(2.5)	(1.7)
Mahalanobis distance	24.2	23.0	20.8	19.5	18.4	16.6	15.6	13.4	11.0	8.5
	(4.1)	(4.3)	(4.0)	(4.2)	(4.5)	(3.1)	(4.4)	(3.7)	(3.6)	(3.4)

Notes: 1997/98 to 2015/16 averages. Bootstrap standard errors in parentheses (100 bootstraps). All distances are multiplied by 100. The top percentile group (100) excludes the top 0.1% prior to 2010/11 (due the presence of composite records in the SPI). Within each row, the cell shading goes from dark grey for the largest value to light grey for the smallest value. The distances are based on gender (2 categories), age (7 categories) and region (12 categories). The Euclidean, Hellinger and overlapping distances are based on the interactions of all variables (resulting in 168 dummy variables). The Mahalanobis and standardised Euclidean distances are based on the sample means of each variable and no interaction is included.
Source: authors' calculations based on FRS, HBAI, SPI and ONS data.

Appendix B

Decomposition materials:

sample distributions, coefficient estimates and average partial effects

Table B1. Women: sample distribution (in per cent), 1999 and 2015

	1999				2015			
	Top 1%		Bottom 99%		Top 1%		Bottom 99%	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Age (years)								
Under 25	1.08	(0.76)	8.89	(0.12)	0.00	(0.00)	7.83	(0.16)
25–34	16.14	(3.22)	19.28	(0.16)	10.08	(2.45)	17.36	(0.20)
35–44	39.28	(3.95)	18.67	(0.15)	31.93	(3.59)	16.69	(0.19)
45–54	28.82	(3.66)	17.20	(0.15)	40.68	(3.84)	18.43	(0.20)
55–59	5.02	(1.69)	6.91	(0.10)	8.90	(2.10)	7.99	(0.15)
60–64	4.63	(1.62)	6.50	(0.10)	3.28	(1.27)	7.11	(0.12)
65–74	3.50	(1.41)	11.57	(0.12)	4.51	(1.47)	13.20	(0.16)
75 and over	1.54	(0.89)	10.98	(0.13)	0.62	(0.62)	11.39	(0.16)
Non-white	5.06	(1.77)	5.36	(0.09)	11.42	(2.74)	11.02	(0.17)
Family type								
Pensioner couple	2.56	(1.14)	11.86	(0.12)	7.40	(1.78)	15.23	(0.17)
Single pensioner	5.79	(1.80)	15.09	(0.14)	1.62	(0.94)	12.74	(0.16)
Couple with children	30.27	(3.58)	23.56	(0.17)	36.05	(3.71)	23.72	(0.22)
Couple without children	33.34	(3.92)	27.73	(0.19)	33.60	(3.75)	25.23	(0.23)
Single with children	2.48	(1.10)	7.32	(0.10)	4.84	(1.84)	6.95	(0.12)
Single without children	25.57	(3.66)	14.45	(0.14)	16.48	(2.76)	16.12	(0.20)
Age completed full-time education								
16 or younger	14.52	(2.77)	65.00	(0.19)	9.35	(2.18)	47.23	(0.25)
17 to 18	18.86	(3.06)	19.16	(0.16)	18.68	(3.08)	23.21	(0.22)
19 to 21	32.97	(3.89)	9.36	(0.12)	27.59	(3.41)	16.06	(0.19)
22	12.40	(2.62)	3.51	(0.08)	22.30	(3.37)	6.54	(0.13)
23 or older	21.26	(3.38)	2.98	(0.07)	22.08	(3.12)	6.98	(0.14)
Employment Status – ILO definition								
Employee	62.09	(3.86)	48.61	(0.20)	77.43	(3.18)	50.94	(0.25)
Self-employed	31.45	(3.67)	3.98	(0.08)	19.14	(3.02)	5.22	(0.12)
NILF or unemployed	6.46	(1.90)	47.41	(0.20)	3.43	(1.29)	43.84	(0.25)
Part-time worker	15.69	(2.92)	21.76	(0.16)	18.58	(2.86)	24.64	(0.22)
Occupation								
Managers, directors & senior officials	39.72	(3.98)	6.17	(0.10)	37.10	(3.77)	4.05	(0.10)
Professional occupations	25.19	(3.47)	5.20	(0.09)	36.88	(3.74)	11.78	(0.17)
Associate prof. & technical occupations	18.18	(3.22)	6.23	(0.10)	16.79	(2.88)	7.77	(0.14)
Other	10.46	(2.50)	34.26	(0.19)	4.70	(1.58)	32.41	(0.24)
Missing/not applicable	6.46	(1.90)	48.15	(0.20)	4.53	(1.53)	43.99	(0.25)
Industry								
Manufacturing	13.38	(2.79)	6.47	(0.10)	3.47	(1.23)	2.01	(0.07)
Transport, storage and communications	3.46	(1.42)	2.08	(0.06)	11.93	(2.66)	2.59	(0.08)
Financial intermediation	10.72	(2.66)	2.76	(0.07)	17.15	(2.98)	2.23	(0.08)
Real estate, renting & business	31.25	(3.82)	5.47	(0.09)	24.13	(3.34)	7.75	(0.14)
Public administration and defence	2.47	(1.12)	3.24	(0.07)	2.82	(1.18)	3.41	(0.09)
Education	3.32	(1.37)	6.41	(0.10)	5.84	(1.79)	8.61	(0.15)
Health and social work	12.04	(2.52)	10.17	(0.12)	15.78	(2.67)	13.41	(0.17)
Others	16.90	(3.02)	15.10	(0.15)	13.75	(2.76)	15.34	(0.19)
Missing/not applicable	6.46	(1.90)	48.31	(0.20)	5.13	(1.64)	44.65	(0.25)
Government office region								
London	40.29	(4.09)	12.27	(0.10)	34.09	(3.88)	13.04	(0.16)
South East	17.22	(2.90)	13.99	(0.10)	20.11	(3.01)	14.15	(0.13)
Rest of the UK	42.49	(3.95)	73.74	(0.13)	45.80	(3.81)	72.81	(0.18)
Partner in top 10%	49.00	(4.06)	13.48	(0.14)	49.83	(3.88)	13.12	(0.17)
Partner in top 1%	25.67	(3.55)	1.53	(0.05)	23.78	(3.32)	1.47	(0.06)
N (weighted, in millions)	0.177		65.719		0.311		72.177	

Notes: Based on 3-year pooled samples (1998/99–2000/01 and 2014/15–2016/17).
Source: authors' calculations based on FRS, HBAI and ONS data.

Table B2. Men: sample distribution (in per cent), 1999 and 2015

	1999				2015			
	Top 1%		Bottom 99%		Top 1%		Bottom 99%	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Age (years)								
Under 25	0.75	(0.31)	9.76	(0.14)	0.36	(0.28)	8.89	(0.18)
25–34	13.45	(1.12)	20.46	(0.18)	7.79	(1.16)	18.26	(0.24)
35–44	32.35	(1.43)	19.60	(0.17)	28.50	(1.74)	17.06	(0.21)
45–54	34.75	(1.44)	17.77	(0.16)	36.04	(1.84)	18.19	(0.22)
55–59	9.81	(0.91)	7.50	(0.11)	12.90	(1.30)	8.42	(0.15)
60–64	3.98	(0.58)	6.74	(0.10)	6.48	(0.88)	7.20	(0.14)
65–74	3.61	(0.53)	10.99	(0.13)	6.51	(0.85)	12.89	(0.16)
75 and over	1.29	(0.33)	7.19	(0.11)	1.44	(0.38)	9.09	(0.14)
Non-white	5.03	(0.68)	5.47	(0.10)	8.81	(1.11)	10.59	(0.19)
Family type								
Pensioner couple	3.85	(0.55)	12.98	(0.14)	7.47	(0.90)	16.17	(0.18)
Single pensioner	1.05	(0.29)	5.20	(0.09)	0.48	(0.20)	5.82	(0.12)
Couple with children	46.58	(1.51)	25.12	(0.18)	45.77	(1.91)	24.57	(0.23)
Couple without children	37.22	(1.51)	29.98	(0.20)	37.40	(1.90)	27.03	(0.25)
Single with children	0.59	(0.20)	0.67	(0.03)	0.57	(0.24)	0.75	(0.04)
Single without children	10.72	(1.03)	26.07	(0.20)	8.33	(1.08)	25.67	(0.27)
Age completed full-time education								
16 or younger	21.25	(1.23)	66.84	(0.21)	15.74	(1.39)	50.07	(0.28)
17 to 18	19.17	(1.19)	15.86	(0.16)	15.89	(1.39)	20.65	(0.23)
19 to 21	23.91	(1.31)	8.56	(0.13)	26.73	(1.70)	14.92	(0.21)
22	12.93	(1.05)	3.74	(0.09)	16.63	(1.46)	6.09	(0.13)
23 or older	22.74	(1.28)	5.01	(0.10)	25.02	(1.66)	8.29	(0.17)
Employment Status – ILO definition								
Employee	66.63	(1.43)	56.89	(0.21)	79.59	(1.56)	55.27	(0.28)
Self-employed	29.68	(1.39)	10.19	(0.13)	16.85	(1.47)	11.31	(0.18)
NILF or unemployed	3.70	(0.55)	32.92	(0.20)	3.56	(0.65)	33.43	(0.25)
Part-time worker	3.67	(0.56)	5.14	(0.10)	5.17	(0.82)	9.29	(0.18)
Occupation								
Managers directors & senior officials	46.30	(1.52)	12.52	(0.14)	40.56	(1.89)	7.75	(0.14)
Professional occupations	28.18	(1.38)	6.90	(0.11)	33.42	(1.79)	11.81	(0.18)
Associate prof. & technical occupations	13.46	(1.07)	6.57	(0.11)	15.45	(1.42)	10.06	(0.18)
Other	8.37	(0.84)	40.26	(0.21)	6.34	(0.99)	36.66	(0.27)
Missing/not applicable	3.70	(0.55)	33.74	(0.20)	4.23	(0.72)	33.71	(0.26)
Industry								
Manufacturing	17.16	(1.14)	17.10	(0.16)	8.37	(1.09)	8.08	(0.15)
Transport, storage and communications	5.09	(0.66)	6.29	(0.11)	11.10	(1.23)	9.06	(0.16)
Financial intermediation	12.77	(1.03)	2.29	(0.07)	20.66	(1.58)	2.39	(0.09)
Real estate, renting & business	29.63	(1.41)	7.57	(0.12)	22.42	(1.59)	9.20	(0.16)
Public administration and defence	3.34	(0.55)	3.83	(0.08)	2.32	(0.56)	3.69	(0.10)
Education	1.79	(0.39)	2.89	(0.07)	2.16	(0.52)	3.69	(0.10)
Health and social work	7.48	(0.79)	2.52	(0.07)	6.86	(0.89)	3.75	(0.12)
Others	18.45	(1.18)	23.48	(0.19)	20.72	(1.59)	25.74	(0.25)
Missing/not applicable	4.28	(0.59)	34.04	(0.20)	5.39	(0.83)	34.40	(0.26)
Government office region								
London	25.51	(1.40)	12.03	(0.12)	25.42	(1.80)	13.15	(0.20)
South East	24.79	(1.28)	13.85	(0.11)	24.00	(1.63)	13.92	(0.14)
Rest of the UK	49.70	(1.52)	74.12	(0.15)	50.58	(1.91)	72.93	(0.22)
Partner in top 10%	15.08	(1.12)	3.12	(0.08)	21.26	(1.62)	4.47	(0.11)
Partner in top 1%	3.80	(0.60)	0.12	(0.02)	5.79	(0.90)	0.28	(0.03)
N (weighted, in millions)	1.197		59.758		1.278		67.733	

Notes: Based on 3-year pooled samples (1998/99–2000/01 and 2014/15–2016/17).
Source: authors' calculations based on FRS, HBAI and ONS data.

Table B3. Women: logit model coefficient estimates for probability of being in the top 1%, 1999 and 2015

	1999		2015	
	Coefficient estimate	SE	Coefficient estimate	SE
Age (years)	0.233***	(0.058)	0.336***	(0.084)
Age squared	–0.002***	(0.001)	–0.003***	(0.001)
Non-white	–0.676	(0.411)	–0.346	(0.298)
Family type (ref. is Single without children)				
Pensioner couple	–0.835	(0.574)	0.427	(0.395)
Single pensioner	1.014*	(0.562)	0.247	(0.717)
Couple with children	–0.719**	(0.293)	–0.395	(0.285)
Couple without children	–0.964***	(0.284)	–0.553**	(0.269)
Single with children	–0.671	(0.486)	0.067	(0.445)
Age completed full-time education	1.145***	(0.294)	1.054***	(0.255)
Age completed full-time education square	–0.023***	(0.007)	–0.021***	(0.006)
Employment Status (ref. is Employee)				
Self-employed	1.125***	(0.231)	0.434*	(0.230)
NILF or unemployed	–1.199***	(0.418)	–1.297***	(0.494)
Occupation (ref. is Other)				
Managers, directors & senior officials	1.952***	(0.297)	2.769***	(0.354)
Professional occupations	1.863***	(0.327)	2.116***	(0.341)
Associate prof. & technical occupations	1.266***	(0.354)	1.672***	(0.367)
Part-time worker	–1.113***	(0.279)	–1.023***	(0.215)
Industry (ref. is Other)				
Manufacturing	0.670**	(0.331)	0.239	(0.442)
Transport, storage and communications	0.623	(0.488)	1.024***	(0.353)
Financial intermediation	1.409***	0.369)	1.504***	(0.303)
Real estate, renting & business	0.823***	(0.277)	0.470*	(0.279)
Public administration and defence	–0.585	(0.513)	–0.607	(0.488)
Education	–1.934***	(0.507)	–1.265***	(0.402)
Health and social work	–0.119	(0.349)	–0.223	(0.303)
Government office region (ref. is London)				
South East	–0.698***	(0.253)	–0.334	(0.248)
Rest of the UK	–0.826***	(0.205)	–0.655***	(0.215)
Partner in top 10%	0.637**	(0.268)	0.590***	(0.226)
Partner in top 1%	1.854***	(0.267)	1.792***	(0.252)
Constant	–25.004***	(3.418)	–26.883***	(3.228)
Sample size	65,702		46,888	

Notes: Weighted estimates. Based on 3-year pooled samples (1998/99–2000/01 and 2014/15–2016/17).
* p < 0.10, ** p < 0.05, *** p < 0.01
Source: authors' calculations based on FRS, HBAI and ONS data.

Table B4. Men: logit model coefficient estimates for probability of being in the top 1%, 1999 and 2015

	1999		2015	
	Coefficient estimate	SE	Coefficient estimate	SE
Age (years)	0.240***	(0.029)	0.298***	(0.041)
Age squared	–0.002***	(0.000)	–0.003***	(0.000)
Non-white	–0.764***	(0.165)	–0.745***	(0.163)
Family type (ref. is Single without children)				
Pensioner couple	1.049***	(0.268)	1.245***	(0.281)
Single pensioner	1.012**	(0.400)	0.019	(0.518)
Couple with children	0.700***	(0.121)	0.594***	(0.159)
Couple without children	0.417***	(0.125)	0.503***	(0.161)
Single with children	0.563	(0.385)	0.379	(0.442)
Age completed full-time education	0.847***	(0.153)	0.885***	(0.149)
Age completed full-time education squared	–0.016***	(0.004)	–0.017***	(0.004)
Employment Status (ref. is Employee)				
Self-employed	0.800***	(0.087)	0.018	(0.126)
NILF or unemployed	–0.487**	(0.209)	–1.031***	(0.271)
Occupation (ref. is Other)				
Managers, directors & senior officials	2.039***	(0.120)	2.394***	(0.172)
Professional occupations	1.814***	(0.139)	1.687***	(0.175)
Associate prof. & technical occupations	1.339***	(0.150)	1.216***	(0.198)
Part-time worker	–1.023***	(0.193)	–0.892***	(0.184)
Industry (ref. is Other)				
Manufacturing	0.438***	(0.112)	–0.021	(0.173)
Transport, storage and communications	0.365**	(0.161)	–0.025	(0.154)
Financial intermediation	1.560***	(0.132)	1.540***	(0.138)
Real estate, renting & business	0.722***	(0.104)	0.330***	(0.126)
Public administration and defence	–0.203	(0.191)	–0.862***	(0.265)
Education	–1.373***	(0.249)	–1.412***	(0.270)
Health and social work	0.491***	(0.142)	0.221	(0.169)
Government office region (ref. is London)				
South East	–0.222**	(0.106)	–0.03	(0.132)
Rest of the UK	–0.809***	(0.093)	–0.524***	(0.113)
Partner in top 10%	0.091	(0.118)	0.428***	(0.126)
Partner in top 1%	1.982***	(0.306)	1.521***	(0.273)
Constant	–21.323***	(1.612)	–23.615***	(1.830)
Sample size	58,062		42,170	

Notes: Weighted estimates. Based on 3-year pooled samples (1998/99–2000/01 and 2014/15–2016/17).
* p < 0.10, ** p < 0.05, *** p < 0.01
Source: authors' calculations based on FRS, HBAI and ONS data.

Table B5. Women: average partial effects (APEs) for probability of being in the top 1%, 1999 and 2015

	1999 sample 1999 coefficients		1999 sample 2015 coefficients		2015 sample 1999 coefficients		2015 sample 2015 coefficients	
	APE (ppt)	SE	APE (ppt)	SE	APE (ppt)	SE	APE (ppt)	SE
Age	0.013***	(0.002)	0.017***	(0.004)	0.016***	(0.002)	0.020***	(0.003)
Non-white	–0.129**	(0.061)	–0.206**	(0.099)	–0.077	(0.059)	–0.120	(0.092)
Family type (ref. is single without children)								
Pensioner couple	–0.237*	(0.134)	–0.366*	(0.208)	0.168	(0.173)	0.250	(0.254)
Single pensioner	0.601	(0.450)	0.913	(0.668)	0.09	(0.287)	0.134	(0.425)
Couple with children	–0.213**	(0.095)	–0.328**	(0.144)	–0.111	(0.086)	–0.166	(0.126)
Couple without children	–0.261***	(0.090)	–0.403***	(0.136)	–0.146*	(0.079)	–0.218*	(0.116)
Single with children	–0.202	(0.124)	–0.312	(0.190)	0.023	(0.153)	0.034	(0.228)
Age completed full-time education†	0.057***	(0.007)	0.072***	(0.012)	0.055***	(0.007)	0.067***	(0.011)
Employment Status (ref. is Employee)								
Self-employed	0.460***	(0.125)	0.675***	(0.183)	0.139*	(0.085)	0.209*	(0.126)
NILF or unemployed	–0.174***	(0.050)	–0.261***	(0.073)	–0.202***	(0.056)	–0.309***	(0.081)
Occupation (ref. is other)								
Managers, directors & senior officials	0.485***	(0.095)	0.677***	(0.133)	0.775***	(0.155)	1.151***	(0.212)
Professional occupations	0.440***	(0.107)	0.614***	(0.147)	0.394***	(0.085)	0.591***	(0.116)
Associate prof. & technical occupations	0.214***	(0.075)	0.299***	(0.103)	0.238***	(0.067)	0.358***	(0.096)
Part-time worker	–0.214***	(0.042)	–0.347***	(0.075)	–0.208***	0.037)	–0.332***	(0.061)
Industry (ref. is Other)								
Manufacturing	0.208*	(0.113)	0.309*	(0.168)	0.067	(0.132)	0.097	(0.191)
Transport, storage and communications	0.189	(0.179)	0.281	(0.268)	0.414**	(0.174)	0.600**	(0.246)
Financial intermediation	0.624***	(0.229)	0.925***	(0.346)	0.767***	(0.189)	1.110***	(0.276)
Real estate, renting & business	0.275***	(0.097)	0.408***	(0.144)	0.146*	(0.087)	0.212*	(0.125)
Public administration and defence	–0.103	(0.079)	–0.155	(0.118)	–0.117	(0.084)	–0.171	(0.122)
Education	–0.205***	(0.053)	–0.308***	(0.082)	–0.189***	(0.060)	–0.275***	(0.090)
Health and social work	–0.026	(0.075)	–0.038	(0.112)	–0.051	(0.070)	–0.074	(0.102)
Government office region (ref. is London)								
South East	–0.207***	(0.076)	–0.316***	(0.116)	–0.105	(0.080)	–0.160	(0.121)
Rest of the UK	–0.233***	(0.069)	–0.356***	(0.106)	–0.182**	(0.071)	–0.277***	(0.105)
Partner in top 10%	0.169**	(0.079)	0.260**	(0.120)	0.158**	(0.069)	0.245**	(0.103)
Partner in top 1%	0.927***	(0.243)	1.424***	(0.368)	0.908***	(0.237)	1.392***	(0.344)
Sample size	65,702		46,888		65,702		46,888	

Notes Average partial effects in percentage points. † Effects account for the inclusion of the squared term in the model. Derived from logit model of the probability of being in the top 1% of the adult income distribution: see Table B4. * p < 0.10, ** p < 0.05, *** p < 0.01. *Source*: authors' calculations based on FRS, HBAI and ONS data.

Table B6. Men: average partial effects (APEs) for probability of being in the top 1%, 1999 and 2015

	1999 sample 1999 coefficients		1999 sample 2015 coefficients		2015 sample 1999 coefficients		2015 sample 2015 coefficients	
	APE (ppt)	SE	APE (ppt)	SE	APE (ppt)	SE	APE (ppt)	SE
Age	0.074***	(0.006)	0.073***	(0.009)	0.070***	(0.005)	0.067***	(0.007)
Non-white	–0.995***	(0.164)	–1.284***	(0.217)	–0.749***	(0.129)	–0.948***	(0.166)
Family type (ref. is Single without children)								
Pensioner couple	1.821***	(0.623)	2.199***	(0.737)	1.830***	(0.571)	2.179***	(0.660)
Single pensioner	1.729*	(0.940)	2.089*	(1.119)	0.017	(0.455)	0.020	(0.548)
Couple with children	1.052***	(0.161)	1.276***	(0.201)	0.661***	(0.157)	0.792***	(0.190)
Couple without children	0.558***	(0.156)	0.679***	(0.190)	0.539***	(0.157)	0.646***	(0.189)
Single with children	0.799	(0.658)	0.970	(0.796)	0.385	(0.508)	0.462	(0.609)
Age completed full-time education†	0.342***	(0.019)	0.373***	(0.024)	0.280***	(0.018)	0.301***	(0.021)
Employment Status (ref. is Employee)								
Self-employed	1.663***	(0.217)	2.071***	(0.262)	0.025	(0.174)	0.031	(0.214)
NILF or unemployed	–0.609***	(0.223)	–0.772***	(0.286)	–0.949***	(0.175)	–1.177***	(0.223)
Occupation (ref. is other)								
Managers, directors & senior officials	3.172***	(0.209)	4.095***	(0.274)	3.012***	(0.262)	4.087***	(0.338)
Professional occupations	2.505***	(0.240)	3.250***	(0.309)	1.427***	(0.162)	1.972***	(0.214)
Associate prof. & technical occupations	1.439***	(0.200)	1.884***	(0.262)	0.793***	(0.146)	1.107***	(0.202)
Part-time worker	–1.224***	(0.158)	–1.561***	(0.210)	–0.851***	(0.129)	–1.066***	(0.163)
Industry (ref. is Other)								
Manufacturing	0.679***	(0.179)	0.843***	(0.222)	–0.025	(0.208)	–0.030	(0.250)
Transport, storage and communications	0.550**	(0.267)	0.683**	(0.332)	–0.031	(0.185)	–0.037	(0.221)
Financial intermediation	3.854***	(0.444)	4.739***	(0.545)	3.527***	(0.414)	4.208***	(0.481)
Real estate, renting & business	1.263***	(0.190)	1.565***	(0.232)	0.461**	(0.182)	0.552**	(0.216)
Public administration and defence	–0.242	(0.214)	–0.301	(0.267)	–0.746***	(0.179)	–0.895***	(0.215)
Education	–1.038***	(0.135)	–1.300***	(0.174)	–0.995***	(0.138)	–1.195***	(0.168)
Health and social work	0.779***	(0.251)	0.967***	(0.310)	0.296	(0.237)	0.354	(0.283)
Government office region (ref. is London)								
South East	–0.509**	(0.248)	–0.619**	(0.301)	–0.049	(0.218)	–0.059	(0.263)
Rest of the UK	–1.514***	(0.211)	–1.851***	(0.254)	–0.720***	(0.178)	–0.871***	(0.212)
Partner in top 10%	0.155	(0.207)	0.191	(0.255)	0.630***	(0.211)	0.759***	(0.252)
Partner in top 1%	6.724***	(1.712)	8.056***	(2.011)	3.509***	(0.986)	4.194***	(1.158)
Sample size	58,062		42,170		58,062		42,170	

Notes Average partial effects in percentage points. † Effects account for the inclusion of the squared term in the model. Derived from logit model of the probability of being in the top 1% of the adult income distribution: see Table B4. * p < 0.10, ** p < 0.05, *** p < 0.01. *Source*: authors' calculations based on FRS, HBAI and ONS data.