

PERSONAL INCOMES

A. INTRODUCTORY NOTE

1. This section provides a selection of tables on the incomes of United Kingdom (UK) taxpayers with some information on the income tax to which they are liable. The way in which income tax liabilities are calculated is described in more detail in the section 'Income tax', which also gives information on coverage of income and the classification as Earnings and Pensions income or Trading and Other income. The tables are derived from the annual Survey of Personal Incomes (SPI), which is described in sub-section B below.

B. THE SURVEY OF PERSONAL INCOMES

2. The sample survey is based on information held by HM Revenue & Customs (HMRC) tax offices on persons who could be liable to UK tax. It is carried out annually and covers the income assessable for tax in each tax year. The tables in this section are based on the surveys for 2007-08 and earlier.
3. Samples were selected from three HMRC operational IT systems, which are as follows:

COP: this covers all employees and occupational or personal pension recipients with a PAYE record;

CESA: this covers the self assessment (SA) population; those with self-employment, rent or untaxed investment income, directors and other people with complex tax affairs or very high incomes (over £100k). Some people have both a COP and CESA record, although after the refinement of many higher rate employees out of Self Assessment this group has reduced.

Claims: this covers people without COP or CESA records who have had too much tax deducted at source and claim repayment.

The approximate sample sizes for recent survey years are as follows:

1999-00	150,000
2000-01	200,000
2001-02	300,000
2002-03	400,000
2003-04	430,000
2004-05	520,000
2005-06	540,000
2006-07	570,000
2007-08	600,000

4. COP data has been captured electronically since the 1998-99 survey. Previously the COP part of the sample was collected by a paper survey. In the 2007-08 survey, the COP population was stratified by sex, marital status (men only) and pay or non-state pension in the previous year. The sampling fractions varied from 1 in 5 for individuals with high incomes and rare allowances to about 1 in 250 for people with low combined pay and pensions. Estimates derived from other data sources have been used where it appears that the information on the COP system is incomplete.
5. The source of the data for the SA part of the sample is the 2007-08 SA Return, held on CESA. An extract of 2007-08 records on CESA was used to stratify the sample, with the sampling fractions varying from 1 in 1 for cases with very high incomes to about 1 in 210 for employees and occupational pensioners with smaller tax liabilities.

6. The self assessment sample cases in each stratum are grossed up to a forecast of the number of people in that stratum who will eventually file 2007-08 tax returns and pay the tax due.
7. Data for the claims cases in the SPI sample relate to 2007-08. This information was captured directly from an extract of the Claims database. A sampling fraction of about 1 in 30 was used to select the sample.

C. NOTES ON THE TABLES

8. In all of the tables in this section, the following conventions have been used:
 - .. not available, due to small sample size
 - negligible
 - . not applicable
 Individual figures have been rounded independently to three significant figures. Therefore, the sum of component items may not necessarily add to the totals shown.
9. The tables in this section only cover individuals shown by HMRC records to have some liability to tax. There may be no record if an individual's income is less than the personal allowance (£5,225 in 2007-08). The lowest level of total income in the tables start at these levels and no attempt has been made to estimate the numbers of cases below the tax threshold or the amount of their incomes.
10. Mortgage Interest Relief at Source (MIRAS) ceased on 6th April 2000 and has been excluded from all the relevant tables in this section from 2000-01 onwards.
11. Some components of investment income (e.g. interest and dividends) are estimated for the sample cases drawn from COP because the information is not held on that HMRC business system. For the components requiring this imputation the amount for each SPI case:
 - is known for cases in self assessment from the amount declared on the Self Assessment Return
 - can be inferred or estimated reasonably for COP cases where there is an adjustment to tax code for higher rate taxpayers
 - is supplemented with information from interest paying institutions
 - is unknown for COP cases where there is no coding adjustment- typically no liability at higher rate.

Where no information at case level is available from HMRC administrative systems, estimated values are imputed to cases so that the population as a whole has amounts consistent with evidence from other sources. For example, amounts of tax accounted for by deposit takers and the propensity to hold interest bearing accounts as indicated by household surveys.

Starting from control totals at UK level; for the number of cases with interest and the total amount of that interest, the numbers of cases and amounts of interest in Self Assessment cases and those COP cases with coding adjustments are deducted to leave targets for the remainder of the taxpayer population. These targets are at UK level – no attempt is made to control the targets to sub-UK geographical units. The cases to which amounts are attached by the imputation process and the amounts attached are determined by probabilistic methods with just the UK targets and distributions in mind.

12. Alternative breakdowns, at UK or Country level, of the information in the tables in this section may be available on request subject to sample size and confidentiality considerations. For example, information could be provided on pay instead of total earned income. Also, information on income, deductions and tax items may be available, tabulated by other items, e.g. age, sex, industry group for the self employed.

Estimates at sub-UK level (Tables 2.2, 3.11, 3.12, 3.13, 3.14 and 3.15).

13. The sample for the SPI is drawn at random from records held by HMRC of persons who could be liable to UK tax, irrespective of where they live. The population of records is not grouped (stratified) by geographical region before the sample is selected. The geography indicators are attached only to the selected sample based on address and postcode. It follows that sub-UK estimates based on SPI are not controlled to known and fixed taxpayer population figures for each area. They are subject to random error caused by sampling and, where sample size is small, estimates for a geographical area can be subject to large sampling errors.

Confidence Intervals around estimates at sub-UK level (Tables 3.13a, 3.14a and 3.15a).

14. A new series of tables for 2003-04 onwards, Tables 3.13a, 3.14a and 3.15a, show the confidence intervals associated with the sub-UK level estimates. Tables 3.13a, 3.14a and 3.15a respectively show the upper and lower limits of the 95% confidence interval.
15. Confidence intervals for taxpayer numbers, mean incomes and amounts are symmetric about the central estimate (i.e. the upper and lower confidence limits are equidistant from the central estimate), but the confidence intervals for median incomes is asymmetric. The reason for this is that, after estimating the median income level (M), which 50% of the population is expected to exceed, a 95% confidence interval is calculated for the proportion of the population with incomes in excess of M. This confidence interval is symmetric about 50% -say (50+x)% to (50-x)%. The income levels associated with 50+x% and 50-x% are estimated from the sample, say M+ and M-. In typical income distributions, the distance from M to M+ will usually exceed the distance from M to M- because there are progressively fewer individuals at higher levels of income.
16. The confidence interval estimates have been calculated without rounding. When displayed using the standard rounding convention, the confidence intervals (for numbers, means and amounts) may not appear to be symmetrical around the central estimate.
17. The accuracy of the confidence intervals will depend on the sample selected being fully representative of all the characteristics of the population. For survey years to 2001-02, an initial sample of COP cases was drawn from the business system by selecting all cases where the last two digits of the National Insurance Number (NINO) matched five of the possible 100 combinations. These cases were grouped (stratified) by gender, marital status and income. A sub-sample was drawn from each group by using specific combinations of two, three or four digits from the end of the NINO, with sampling rates which varied by category.
18. For 2002-03 and subsequent surveys, the initial sample of COP cases comprises all cases whose last two NINO digits match 10 of the possible 100 combinations. The sub-sampling process uses NINO digits prior to the last two, in which no geographical clustering has been identified. Sub-sampling ignores the last two NINO digits.
19. The sub-sampling process was changed from 2002-03 due to an uneven geographical distribution of the cases supplied with the five NINO ends for the 2001-02 survey. This was traced back to an apparently incomplete extract of PAYE data in April 2003 from one of the twelve Regional Processing Centres. The data extracted in April 2003 was used for both the 2001-02 and 2002-03 surveys. From an extract of about 3.3 million records for the UK, it is estimated that about 40 thousand were missing. The shortfall is around 1% for UK and a little less for each of England, Scotland and Wales. However a disproportionate number of these records relate to Northern Ireland individuals for which the shortfall is around 12%. It is not possible to measure local variations in the shortfall for lower level geographical areas within England, Scotland or Wales. Consequently, all sub-UK tables based on the 2002-03 survey and earlier years are published in their entirety but the figures should be treated with caution whether viewing levels and distributions within a year or making comparisons across years.

D. ENQUIRIES AND FURTHER INFORMATION

20. Tables 3.11 to 3.15 (including 3.13a, 3.14a and 3.15a) for 2007/08 will be published at the end of February 2010.
21. The tables in this section are expected to be updated to incorporate information from the 2008-09 survey in March 2011, with the exception of tables 3.11 to 3.15 (including 3.13a, 3.14a and 3.15a) which are expected to be updated in April 2011. This delay is necessary to allow the development of new processes for handling PAYE data to be delivered from a new business system, the National Insurance and PAYE system (NPS) which has replaced COP.
22. Requests for further information should be addressed to Gordon McGregor, KAI: Corporate Business Intelligence – Strategy & Production, HMRC, 3rd Floor South Spur, South West Bush House, Strand,

London WC2B 4RD. A telephone enquiry number for this section is given in the [Update calendar and enquiry points page](#).

E. SAMPLE ESTIMATES AND MEASURES OF PRECISION

Population and sample design

23. The SPI aims to cover all individuals with a UK income tax liability. The sample drawn from HMRC operational systems will include some cases where income is less than allowances so no tax liability arises. Paragraphs 3 to 7 above, explain how the records in each operational system are grouped (stratified) before the sample is selected. A random sample of records is drawn from each grouping (stratum) - the proportion of cases varies from stratum to stratum.

Reliability of estimates

24. As with all sample surveys, estimates from the SPI have a sampling error attached to them. A statistic (e.g. an estimate of a mean or a total from a random sample) will be subject to sampling fluctuation - its value will vary from one sample to the next if repeated random samples are drawn. The Standard Error of the statistic measures the extent of the variability. It depends on how much spread exists in the observations from the sample and the size of the sample.
25. In general, the larger the sample size, the smaller the standard error. To a lesser extent, the standard error of the statistic will decline as the proportion of the population surveyed increases, but only by taking measurements for the whole population can sampling error for the statistic be removed entirely.
26. A Confidence Interval for the statistic is constructed from the standard error. It gives an estimated range of values which is likely to include the unknown population parameter that has been estimated by the statistic. If independent samples are taken repeatedly from the same population and the confidence interval is calculated for each sample, then a proportion (known as the Confidence Level) of such intervals will include the unknown population parameter.
27. A 95% Confidence Interval is one that if compiled repeatedly would encompass the population parameter 19 times in 20. For a given sample size, narrower intervals can be compiled if a greater risk of failing to encompass the true population value is acceptable, whereas if greater certainty of including the true value is required, the interval will be wider.
28. The Upper and Lower boundaries of the Confidence Interval are called the Confidence limits. They are a function of the statistic, the standard error of the statistic and the degree of confidence required of the interval.
29. The calculation of sampling errors assumes a simple random sampling method but can be extended to more complex sample designs. The sample for the SPI, as described earlier, is selected using a stratified sample.

Precision of estimates: sub-UK areas.

30. The population is not stratified by geographical area before the SPI sample is selected. Estimates of taxpayer numbers for low level geographical areas of the UK depend on measuring the proportion of the UK population which belong to the area. Typically these proportions are very small and to ensure high precision for any estimated proportion, the sample size across the UK needs to be large.
31. The table below gives an indication of the level of precision which may be assumed, with 95% confidence, for an estimate of taxpayer numbers from a simple random sample as large as the 2003-04 SPI. It shows that for estimated populations of 2.5 million or more, the estimate will be within 1% of the true population with 95% confidence. As the estimated population falls, the level of confidence interval reduces more slowly, thus increasing in size relative to the estimate. For a typical Parliamentary Constituency with an estimated taxpayer count of 46 thousand, the true figure may lie between 42 thousand and 50 thousand. The error could be +/- 8% of this estimate. For a large Parliamentary Constituency, the error may be about 4 thousand (7% of the estimate), while for a small constituency, the error may be 3 thousand (about 9%).

**Survey of Personal incomes:
Confidence intervals for estimates of taxpayer numbers(1)**

Geographical area	Estimated value '000	Confidence Limits		95% confidence interval (+/-) '000	As % of estimate
		Lower limit '000	Upper limit		
Government Office Region (medium)	2,500	2,475	2,525	25.0	1.00%
County (large)	500	488	512	11.6	2.32%
County (small)	200	193	207	7.4	3.69%
Parliamentary Constituency					
Large	57	53	61	4.0	6.93%
Medium	46	42	50	3.6	7.72%
Small	34	31	37	3.1	8.98%

(1) Assumes a taxpayer population of 29 million, a simple random sample of 400 thousand. In practice, estimates will reflect the more complex SPI sample design.

32. Broadly speaking, as sample size changes by a factor x , the confidence interval will change by a factor $1/\sqrt{x}$, so a fourfold increase in sample size will halve the confidence interval.
33. Year on year changes in published estimates of taxpayer numbers within small geographical areas (e.g. districts and constituencies) should be viewed with caution. They involve measuring small differences between two very small proportions. The confidence interval for the difference could be large relative to the measured difference, so any observed change may be due to sampling fluctuation alone.
34. Similar precision, or relative precision, to those shown in the table above in estimates for subsets of taxpayers (e.g. pensioners or higher rate taxpayers) in small geographical areas (e.g. districts and constituencies) requires even greater national samples, far in excess of the present sample size of the SPI. Estimates in such detail are not considered sufficiently reliable to be published.

Notes updated January 2010