Micklewright, John; Schnepf, Sylke Viola

Working Paper

How reliable are income data collected with a single question?

IZA Discussion Papers, No. 3177

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Micklewright, John; Schnepf, Sylke Viola (2007) : How reliable are income data collected with a single question?, IZA Discussion Papers, No. 3177, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/34744

DISCUSSION PAPER SERIES

IZA DP No. 3177

How Reliable Are Income Data Collected with a Single Question?

John Micklewright
Sylke V. Schnepf

November 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

How Reliable Are Income Data Collected with a Single Question?

John Micklewright
S3RI, University of Southampton
and IZA

Sylke V. Schnepf
S3RI, University of Southampton
and IZA

Discussion Paper No. 3177
November 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3177
November 2007

ABSTRACT

How Reliable Are Income Data Collected with a Single Question?*

Income is an important correlate for numerous phenomena in the social sciences. But many surveys collect data with just a single question covering all forms of income. This raises issues of quality, and these are heightened when individuals are asked about the household total rather than own income alone. Data are typically banded, implying a loss of information. We investigate the reliability of 'single-question' data using the ONS Omnibus and British Social Attitudes (BSA) surveys as examples. We first compare the distributions of income in these surveys – individual income in the Omnibus and household income in the BSA – with those in two other much larger UK surveys that measure income in much greater detail. Second, we investigate an implication of restricting the single question to individual income and interviewing only one adult per household: total income in respondents' households is unobserved. We therefore examine the relationship between individual and household income in one of the comparator surveys. Third, after imposing bands on comparator survey data, we measure the information loss from banding with Generalised Entropy indices. We then assess its impact on the use of income as a covariate. Disaggregation by gender proves fruitful in much of the analysis.

JEL Classification: D31, C8

Keywords: income data, banding, information loss, Omnibus survey, British Social Attitudes survey

Corresponding author:

John Micklewright
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield
Southampton SO17 1BJ
United Kingdom
E-mail: jm4@soton.ac.uk

* This research was supported by ESRC project grant 'Giving to Development' (RES-155-25-0061), which forms part of the Non-Governmental Public Action programme. We thank our co-investigator Tony Atkinson for comments and suggestions. We are grateful for information on the Omnibus survey to Theodore Joloza and Charles Lound of ONS, and for comments to Richard Berthoud, Stephen Jenkins, Steve Pudney, Holly Sutherland and participants at seminars at Southampton and Essex, and at the RSS 2007 annual conference and the GSS 2007 methodology conference.

1. Introduction

Income is an important correlate of numerous phenomena in the social sciences and information on income is therefore very commonly sought in social surveys. But in many surveys, income is not a principal focus of interest and limitations on the length of interview mean that detailed income questions cannot be asked. As a consequence, information is often collected with just a single question covering all forms of income. This raises questions of data quality. Problems may be exacerbated where respondents are asked not about their own individual income but about the household total. The data in single-question surveys are also typically banded, implying a loss of information on the within-band variation in incomes. There is a trade-off between the extent and detail of income questions and the accuracy of the resulting income estimates. This paper explores this trade-off. We illustrate the issues using two major UK surveys, the Office for National Statistics Omnibus survey (OMN) and the NatCen British Social Attitudes survey (BSA). These provide examples of the two main forms of 'single-question' survey, the OMN collecting information on individual income, the BSA seeking information from one individual on total income in his or her household.[1]

First, suppose that we are confined to a single question. What is the cost in terms of loss of accuracy? We address this by comparing the distributions of income in the single-question surveys with those in two surveys that collect income data in great detail: the Family Resources Survey (FRS) and the Expenditure and Food Survey (EFS). The FRS is the principal source of information on the distribution of income in the UK. The forerunner of the EFS, the Family Expenditure Survey (FES), had the same role until the mid-1990s. Our comparisons provide an indirect method of assessing the quality of the income data collected in single-question surveys. They complement the limited more direct evidence from comparisons of the same individuals' answers both to a single question and to a battery of questions on different forms of income (e.g. Foster and Lound 1993, Berthoud 2004), and research using cognitive methods to assess how people respond to single questions on income (e.g. Collins and White 1996).

Second, suppose we are limited to individual income, as a consequence of collecting just information on own income and of interviewing just one person per household – a common design in many single-question surveys. What do we lose compared with household

[1] Other examples of UK surveys that collect information on total income (individual or household) with a single question include the British Crime Survey, the British Election Study, the Citizenship Survey, the Health Survey for England, and the National Travel Survey. Examples of cross-national single-question surveys include the European Social Survey, Eurobarometer, and the International Social Survey Programme.

income? It is often household income that is of most interest to the analyst. If incomes are pooled within the household, individual welfare and behaviour will be affected by the incomes of the other household members. We examine the relationship between individual income and household income per adult in the FRS, which does interview all adults in the household.

Third, if a single question means that we only get banded data, how much information do we lose and what are the consequences? Compared to the situation in which (perfectly measured) data are collected in continuous form, banding results in a loss of information. Taking the OMN banding as one of our examples, we use Generalised Entropy measures to quantify the loss under different assumptions about the part of the distribution that is of most interest. We then discuss the implications of the loss for the use of income as a covariate, whether in descriptive analysis or as an explanatory variable in a regression model.

Sections 3-5 investigate the three questions just outlined. Section 2 paves the way by describing the surveys we use and their measurement of income. Section 6 concludes.

2. Data on Incomes

The single-question OMN and BSA are both long-running surveys. The Office for National Statistics (ONS) conducts the former every month. In common with surveys of this type in other countries, it is intended as 'a fast, cost-effective and reliable way of obtaining information on a variety of topics too brief to warrant a survey of their own' (ONS 2007). The survey receives the 'National Statistics' label, a quality marker applied to some of the UK's official statistics. The BSA has collected information since 1983 on social attitudes in Britain and like the OMN is drawn on by a wide range of different users.

Both surveys have conventional multi-stage probability designs. Both interview only one adult (selected at random) per sampled household. Adults are defined as aged 16 or over in the OMN and 18 or over in the BSA. The OMN response rate is typically around 65 percent, yielding an achieved sample size each month of about 1,250 persons. Our own interest in the OMN – and our motivation for investigating the income data – stems from a module of questions on charitable donations that is sponsored three times a year by the Charities Aid Foundation (CAF) and the National Council for Voluntary Organisations (NCVO). We analyse data only from those months in which this module was conducted in

2004/5: July and October 2004 and February 2005.[2] These months are spread relatively evenly through the financial year. We use the BSA sample for 2004, which covers June to September. The response rate was 57 percent. Data collection in both surveys is through face-to-face interview.

Our comparator surveys, the FRS and EFS, share the same sampling frame as the OMN and BSA (the Royal Mail Postcode Address File), also have multi-stage designs, and again use face-to-face interviewing. But both surveys interview all adults in responding households. They operate continuously through the year, with the interviews spread evenly, and have far larger sample sizes. We analyse microdata in both cases for the entire financial year 2004/5. The household response rates in 2004/5 were 63 percent in the FRS and 57 percent in the EFS. By these yardsticks, the levels of response in the OMN and BSA seem reasonable (albeit they refer to individuals rather than households).[3] The use of two comparator surveys emphasises that no one source provides 'the truth' – estimates of the income distribution from the FRS and EFS are known to differ somewhat (Department of Social Security 2000).

Since the OMN and BSA cover Great Britain (the UK excluding Northern Ireland), we limit analysis of the FRS and EFS to the same basis. Again for reasons of differing coverage, in all four surveys we analyse only people who are aged 19+. Imposing these criteria, we have unweighted sample sizes of 5,102 persons in the OMN, 3,162 persons in the BSA, 44,993 persons (in 26,073 households) in the FRS and 11,128 persons (in 6,261 households) in the EFS.

The OMN and BSA data contain a weight for each individual that adjusts for the higher probability of a person being interviewed in small households, which we apply throughout. The FRS and EFS have weights that take account of (measured) differential non-response, which we again apply. From 2005/6, the OMN also has non-response weights and we test their impact below for data from that year.[4]

Differences in composition between the achieved samples in the four surveys could help explain any differences found in the distribution of income. In the Appendix we focus on the comparison of the OMN with the FRS and consider gender, age, employment status and

[2] Results from the charitable donations module are reported in CAF and NCVO (2005, 2006) and Micklewright and Schnepf (2007).
[3] The OMN rates for the months we analyse were 67 percent in July, 66 percent in October and 64 percent in February.
[4] See http://www.statistics.gov.uk/about/services/omnibus/sample.asp.

education, all factors that have a considerable impact on income. Compositional differences do not impact clearly in one direction.

Measurement of income with a single question

Income data are collected in similar ways in the two single-question surveys, and in both cases the question refers to gross income, before deductions for tax and social insurance. OMN respondents are shown a card listing groups of annual income – 33 in 2004/5 – and 11 possible sources of income (intended to be exhaustive).[5] They are asked:

> 'Will you please look at this card and tell me which group represents your total income from all these sources before deductions from income tax, National Insurance etc."

Although the card lists annual amounts, respondents seem free to give an annual equivalent of their current weekly or monthly income if they wish to do so.

In the BSA, respondents are first asked whether they (or their partner) receive each of a large number of different state benefits. Next they are asked what is their main source of income from a card listing a number of possibilities (including earnings, various forms of pension, student loans etc.). Finally they are shown another card with 17 letters indicating both bands of annual income and their weekly equivalents and are asked:

> 'Which of the letters on this card represents the total income of your household from all sources before tax?'

The provision of both annual and weekly amounts on the card again suggests that respondents are free to choose the time period to which their reported incomes refer.

Although the methods of collection are similar, an important difference is that the OMN asks for information on *individual* income while the BSA seeks the total income of the *household*, a distinction that we take up in the next section.

[5] These are earnings from employment or self-employment, pension from former employer, personal/private pension, state pension, child benefit, income support, other state benefits, interest from savings, other kinds of regular allowance, other sources e.g. rent, and no source of income.

The two sets of income bands are shown in the Appendix. Band width increases with income. The OMN groups are clearly chosen so as to obtain a roughly even spread of the sample; apart from the top interval of £36,400+, which contains eight percent of persons responding to the income question, only one other group (£5,200 to £6,239) contains over five percent. From 2005/6, an additional six closed intervals were added for high incomes so that the top interval is now £52,000+. The top interval in the BSA of £56,000+, which refers to household income, contains 12 percent of the sample providing information on income.

The task for participants in a 'single-question' survey should not be underestimated:

> 'Firstly, the respondent has to interpret the question, specifically what is meant by gross [or net] income. Secondly, he or she must retrieve the information from memory, thirdly make a judgement about the information, and finally, find the appropriate answer category to tick….If respondents are paid at different intervals [time periods] to the intervals presented in the questions, they will have to convert their answers to the appropriate interval….For those with more than one source of income, the calculation of the amount becomes even more complex.' (Collins and White 1996: 3).

Even when a reminder is given of different possible types of income, as occurs in both the OMN and BSA, respondents may fail to consider all sources. If they attempt to give annual totals, they may fail to recall how their income varied over the previous 12 months on account of job changes, unemployment or sickness (most people in the UK do not need to submit annual tax returns). And, where the single question refers to the household total, as in the BSA, respondents may be very uncertain of incomes other than their own, compounding the problems of collecting data on individual income alone.

Non-response to the income question

Although the use of a single question is designed in part to reduce respondent burden and thus to increase response to an enquiry into incomes, some people decline to provide the requested information. This threatens data quality. The OMN and the BSA illustrate the problem: both suffer from item non-response for income and in neither case is the occurrence of missing data anywhere near random. In the OMN, 9.4 percent of respondents in our 2004/5 sample declined to answer the income question. Non-response is slightly higher in the BSA to the

question on household income: 11.7 percent in our sample. (These figures refer to unweighted data.)

Those answering the OMN income question are notably younger (with an average age of 48 compared to 57 for those not responding), better educated (18 percent with a degree compared to 12 percent for those not answering), in employment (61 percent compared to 48 percent), more likely to have supervisory positions in their current or last job (27 percent compared to 20 percent), and less likely to be self-employed (10 percent compared to 18 percent). Respondents to the BSA income question are again more likely to be employed and have a lower average age. Married people and those in smaller households (who presumably find it easier to report the household total requested in the BSA) are also more likely to provide data.

This pattern of non-response will have an impact on the measured distributions of income since several of the characteristics concerned are correlated with income. In the case of the OMN, we investigated this by first estimating regression models of the natural log of income for persons providing information. The banded nature of the data was allowed for with a model for grouped data estimated by maximum-likelihood.[6] The models were estimated separately for men and women and contained a wide range of explanatory variables often included in explanations of income based on human capital theory e.g. education, age, and occupation. Coefficients typically had the expected sign and were often highly significant (which itself speaks well of the data). We then used the results to impute income for persons not responding to the income question. Median imputed annual income for the non-respondents was £8,434, which is 30 percent below the £11,902 we estimate for respondents (we interpolate linearly in the critical band). However, since non-respondents represent less than 10 percent of the total sample, the bias in estimates of average income induced by excluding them is quite small.

We restrict analysis of the OMN and the BSA in the rest of the paper to persons providing income data. However, it is worth noting that users of these surveys, or the survey organisers, could impute the missing data in the manner above or in some other way. (Individuals could be assigned by ONS or NatCen to the appropriate income band on the basis of the imputed figure with a indicator variable included in the microdata to show if imputation

[6] The *intreg* procedure in Stata ®. (Persons reporting zero income are treated as having annual income of £1.) Results are available from the authors on request.

had occurred).[7] Imputation is already undertaken by ONS to deal with non-response to individual income questions in the FRS and EFS – see below.

Measurement of income in the comparator surveys

The income data in the FRS and the EFS are collected in much more detail than in the OMN and BSA and there is little doubt that these two surveys should provide superior measures of individual and household income. Both surveys collect information separately on each possible income source. They both collect exact amounts, rather than requesting information in banded form. The information provided by respondents is verified during interview where possible. For example, 60 percent of earnings data in the FRS was verified from payslips in 1998/9 (Frosztega 2000, para 5.2).

However, neither the FRS nor the EFS provide perfect yardsticks. First, despite the care and attention paid to collection of income in both surveys, both are known to measure income imperfectly. For example, both surveys are said not to identify about a third of investment income (Department of Social Security 2000: 17).[8] The same study reported that the FRS 'provided an income distribution that understated' the distribution in the FES (the forerunner to the EFS, measuring income in a very similar way) and that the FRS 'suggests over-representation of some low income households and under-representation of some types of high income households' (*ibid*.: 12, 14).

Second, to the extent that OMN and BSA respondents do report incomes over a 12 month period (which is subject to doubt), one should recognise that the FRS and EFS data refer to a shorter period.[9] ONS has long eschewed collecting annual income data in these surveys in favour of weekly or monthly figures (respondents may provide figures for earnings, for example, on either basis).[10] Annual income has a lower variance than weekly or monthly income (Böheim and Jenkins 2006). We convert all income variables in the FRS and

[7] An alternative would be to construct a weight based on an estimated model for the response probability.
[8] Atkinson and Micklewright (1983) discuss difficulties in comparing survey income aggregates with totals from National Accounts.
[9] For the FRS, we use the gross weekly individual income variable ('indinc'), which is the sum of the totals for each separate income source. For the EFS, we use a variable that measures total 'normal' gross weekly income ('P051'), where the definition of 'normal' by long-standing convention is left to the respondent. As in the FRS, it is the sum of all separate individual income sources. We measure household income as the sum across all individuals in the household of these variables. Some types of income are not strictly personal, notably Housing Benefit for low income households. In the FRS, this is attributed to the household reference person. We assume the analogous person in the OMN (if sampled within the household) would include this sort of income in his or her personal total.
[10] An important reason for this is the difficulty in collecting information on annual amounts via recall that is able to be verified (see, for example, Kemsley *et al.* 1980: 71).

EFS to their annual equivalents. There are also differences in the timing of the surveys. In particular the BSA data are collected in June to September while the FRS and EFS survey continuously through the year. We experimented with taking FRS data from just June-September for the comparison but the results were very similar to the full year data on which we focus. (Greater differences would be found in a high inflation period.)

Finally, the item non-response to income questions in FRS and EFS is a reminder that this phenomenon is hard to avoid in any survey, no matter how detailed the collection of income. Some 14 percent of the income information was obtained by proxy from other household members in the 1998/9 FRS and interest from assets and savings was imputed in 13 percent of cases where respondents refused or did not know the required information (Frosztega 2000, paras 4.2 and 5.1). Missing data on interest on savings are also imputed in the EFS, although proxy responses are not allowed.

3 Comparisons of Income Distributions

A comparison of the distribution from a single-question survey with that from a survey collecting detailed information on incomes is an indirect method of assessing the reliability of the former's data. Research comparing directly the same individuals' answers with both methods has found a mix of under- and over-statement in responses to a single question (Foster and Lound 1993). The differences we find in the distributions from two surveys using contrasting methods of collection will reflect the net effect of under and over-statement, as well as other factors (such as differences in time period and composition of samples). The 'direct' research suggests that the net effect may be dominated by understatement, may be larger for individuals with more complicated incomes, and may be greater for household income than individual income.

Individual income

Figure 1 graphs the cumulative frequency distributions of gross individual income in the OMN, FRS and EFS. (The Appendix gives the underlying data.) Table 1 gives estimates of selected quantiles. We assume a uniform distribution within the bands concerned to obtain the estimates for the OMN. This assumption is fairly innocuous given the band-widths and densities and we apply it for all estimates from the OMN and the BSA in this section. We do

not interpolate in the top unbounded interval, and this determines the choice in each case of the highest percentile to estimate.

Looking at the men, the first impression from the graph is of a high degree of similarity between the three sources. The differences in the cumulative percentages between the OMN and the two other surveys exceed two percentage points for only two groups for the FRS and three groups for the EFS. Consider the tails of the distributions: the percentage of men with no income is 1.4 in the OMN, 2.7 in the FRS and 1.9 in the EFS, while the figures for the top group of £36,400 or more are remarkably similar, 12.9 percent, 13.2 percent and 13.1 percent respectively. The quantiles in Table 1 reveal the general pattern more clearly. The 5th percentile in the OMN is above those from the other two surveys, while elsewhere the OMN gives the lowest estimate of the three but often not by much. The 95 percent confidence interval for the percentage in the OMN up to £16,640 (50.3 percent) contains the FRS figure. The difference between the 5th percentiles in the FRS and EFS shows that the choice of yardstick can influence the picture obtained: the OMN is much closer to the EFS. However, elsewhere the FRS and EFS are in closer agreement with each other than with the OMN. The larger difference at the bottom of the distribution is consistent with the hypothesis that OMN respondents are indeed reporting annual incomes rather than annualising current incomes (which have a higher variance) but this does not explain the similarity towards the top of the distribution.

The comparison is different for women. The larger differences between the OMN and both the other two surveys are clearly visible in Figure 1. And all the OMN quantiles are below those in both the FRS and EFS. The seven intervals from £4,680–£5,119 to £10,400–£11,439 have cumulative frequencies in the OMN that average 6.5 percentage points higher than those in the FRS and 4.5 points higher than those in the EFS. However, as for the men, the distributions converge at high levels of income so that the percentages in the top income group are again remarkably similar in the three surveys (3 to 4 percent). The percentages with zero income also differ very little.[11]

These results refer to 2004/05, before ONS provided weights to partially correct for unit non-response to the OMN. Their use with the 2005/06 data pushes up the estimate of the median for men by about 1 percent and moves that for women slightly down.[12] Our findings

[11] With the exception of the zero income group, these results imply that there is first order stochastic dominance of the distributions with the cumulative OMN percentages higher than those in the other surveys at all income levels.

[12] These results refer to the months in 2005/06 when the CAF/NCVO charitable donations module was included in the survey: June and October 2005 and February 2006.

on item non-response for income in 2004/05 reported above imply that weights to account for this would bring the estimated medians down, probably by around 3 or 4 percent. The OMN medians would fall still further below those in the other surveys.

Figure 2 probes the different picture for men and women in more detail, focusing on the comparison with the FRS and separating the samples by age and labour force status. The distribution for active men aged less than 65 is very similar in the two sources. However, for inactive men of this age and for men aged 65 and over the distributions differ quite a lot, especially in the middle two thirds for the former and the bottom two thirds for the latter. The OMN medians are 78 percent and 89 percent respectively of the FRS estimates. This is consistent with the hypothesis that a single question on income produces more accurate answers from people with earnings from employment than it does from those not in work and reliant on benefit income or pensions. Age may also be a factor. However, for the women, sizeable differences between the distributions are found for the sub-sample of active persons below statutory retirement age as well as for the inactive and those above retirement age.

We further disaggregated the active women aged less than 60 into those with dependent children present in the household and those without. (Women cannot be linked with their own children in the OMN.) The same broad pattern as for all active women less than 60 was found (results not shown). But the distributions are much closer for women in households without children: the OMN median is 95 percent of the FRS figure compared to 86 percent for the women with children. (The distinction is not important for men.) We hypothesise that women with children are failing to include state benefit income associated with the children, such as child benefit (a universal benefit received in respect of all children and paid to the mother).

Household income

A single question on the total income in the household raises issues that go beyond the measurement of individual income alone. Knowledge can be expected to be less for others' income than for one's own. Even where couples pool all income in a joint bank account, partners may have imperfect information on each other's gross, pre-tax figures (the account receiving net salaries, net benefit payments, net interest and dividends etc). The result seems more likely to be under-reporting than over-reporting.

We view the switch from reporting individual income to household income as a 'treatment' and consider its effect within a quasi-experimental evaluation framework. We

compare the estimated quantiles in the two single-question surveys – OMN and BSA – for multi-adult households, in each case as a percentage of the corresponding FRS quantiles for individual and household gross income respectively. This comparison involves a change in survey as well as the 'treatment' of reporting a different income concept. We therefore also compare the quantiles for a 'control' group, the single-adult households in the two surveys. Individual and household income are the same for this group and the difference between the two sets of quantiles (again given as a percentage of corresponding FRS quantiles) is unaffected by the treatment. In effect we evaluate the effect of the treatment by considering a 'difference in differences'.

The final column of Table 2 shows that BSA quantiles for the multi-adult households, as a percentage of those in the FRS, are well below the corresponding figures for individual income in the OMN shown in the penultimate column. The 'treatment' of asking for household rather than individual income appears to have a negative net impact on the group's ability to report income data. The difference between the surveys is much larger towards the bottom of the distribution. In the top half of the distribution, the household income figures in the BSA, relative to the FRS, are about 10 percentage points below the individual figures in the OMN. (See also the cumulative percentages for the income bands in Table A2.) This difference is about the same for men and women. Although the household figures for women are in general lower than those for men (that is, the BSA quantiles as a percentage of the FRS quantiles are typically lower), they are no worse than they are for individual income – where they are also lower.

However, we have yet to take into account the picture for the control group of single-adult households for whom individual and household income are the same by definition. For them, the BSA figures in the second column are *higher* than the OMN figures in the first column, substantially so for men in particular. In other words, the effect of the change in survey alone from OMN to BSA appears to result in higher figures being reported. We cannot rule out that this reflects a difference in composition between the BSA and OMN samples of single adults. But this possibility aside, the results suggests that the effect of the 'treatment' on multi-adult households is even larger than suggested above. The slightly different wording of the income question in the BSA and, in particular, the prior questioning of respondents about receipt of different state benefits and income sources, may improve reporting of income *per se*, something that is only revealed when looking at the single-adult households.

4. The Relationship between Individual and Household Income

A survey that collects information on individual income and that interviews only one adult per household, as in the OMN, will not observe total income in the household (except for single-adult households). Some people will have access to more income than their own. Others will need to share their income with other people in the household, such as a spouse who does not work. Some households pool all their income: rightly or wrongly this is the assumption typically made in most analyses of individual wellbeing and behaviour in the social sciences (see e.g. Burton *et al* 2007). Users of data in surveys like the OMN therefore need a guide to the relationship between individual and household income. What is lost by the focus on individual income?

We address this question by comparing the two measures in the FRS for 2004-5. The analysis is again restricted to adults in Britain aged 19 or over and the income concept is again gross income from all sources, expressed in annual terms. Household income is the sum across all persons in the household of the individual figures.

Table 3 summarises the distributions of individual income and of household income per adult (the household total divided by the number of adults).[13] The results for men are not surprising: both the mean and the variance of household income per adult are substantially lower than for individual income, men sharing households with people who on average have lower incomes than themselves. Conversely, the mean for women rises in the switch to household income per adult but – and this was less predictable – so does the variance.

The key issue of interest to any user of survey data such as these is the relationship between the different income concepts at the individual level. Figure 3 plots the natural logs of the two variables. The lower correlation coefficient for women summarises the weaker association between the two income concepts in their case. The data points on the 45 degree line are for persons living in single adult households, for whom individual income equals household income per adult by definition. (Their exclusion has relatively little impact on the correlation coefficients.) Including this group in the figures (about a fifth of the sample), around 40 percent of either sex has household income per adult that is within 20 percent of individual income. However, while just over a half of women have income per adult that exceeds their individual income, this is true of only a quarter of men (the pattern reflecting the means in Table 3).

[13] We are not trying to measure household welfare in this exercise, so we do not adjust the total income for the household with an equivalence scale that takes account of size (including children) or composition.

The diagrams show clearly the many persons, especially women, who have low individual income but substantially higher household income per adult. Among women with below £2,500 of individual income (1 in 10 of the sample), as many as a half have household income per adult of £10,000+, and the mean household income per adult is *above* that for women with individual income of £2,500–£10,000. (At £13,450, it in fact equals the median for all women.) Many fewer men with individual income in this category have high levels of household income per adult. Low income women are more likely to be living in high income households than are low income men. Moving to the other end of the distribution, however, only a minority of persons with individual income of £25,000 or more have household income per adult that is much below this level. And almost all persons at this level of individual income have household income per adult that is above the median. On average, women 'gain' more than men in the switch to household income per head and at higher levels of income they 'lose' less. High income women are more likely to be living in households with other higher income people (often their partners) than are high income men.

What are the implications of these comparisons for the user of a survey like the OMN who believes that household rather than individual income is relevant for a topic under investigation? First, low individual income needs to be treated with caution, especially in the case of women. Many women with low income of their own live in households with income per adult that is much higher. Second, high individual income *is* generally associated with high household income (by definition, once individual income is high enough). FRS data show that virtually all persons in the top 20 percent of the distribution of individual income (taking men and women together) are in the top half of the distribution of household income per adult, and 3 out of 4 are above the top quartile. Third, it is worth remembering that individual income and household income are the same for single-adult households. Hence, this group is worth investigating alone (e.g. see Micklewright and Schnepf 2007), although the behaviour and circumstances of people living with no other adults may differ from other people in a number of ways. Users pooling several months of OMN data will obtain reasonable-sized samples of persons in single-adult households.

5. Information Loss through Income Banding and its Implications

Survey designers can economise by asking for income information in bands; indeed with a single question it may not be possible to obtain greater precision. Hence banded data are almost invariably found in the single-question surveys. Collection in bands implies a loss of

information compared to collection on a continuous scale. We estimate the extent of that loss and then assess its implications. A maintained assumption is that the banded variable is perfectly measured, i.e. respondents do assign themselves to the correct income category. (We return to this assumption at the end of the section.)

The extent of information loss

We illustrate the problem by dividing the EFS sample into the groups of gross individual income defined by the OMN bands. We can now think of there being within-group information and between-group information in the EFS measure. The former would be lost within the OMN or in any other survey collecting banded data. The latter would be retained (provided one is prepared to estimate the band means by assuming a form for the within-band distributions).

A natural way to measure the loss of information is to split the variation of income into within-band and between-band components. We do this with the Generalised Entropy (GE) class of indices. These indices are commonly used to decompose income inequality into between-group and within-group components, where the groups could be areas of a country, ethnic groups etc (see e.g. Jenkins 1991, Cowell 1995). The GE indices are 'additively decomposable', meaning that total inequality (or 'variation' in the present application) can be expressed as the sum of the inequality within groups and that between groups, the latter being a function of the group means. The general formula of the GE indices is given below. A further attraction of this class is embodied in the parameter, *a*, that indicates the weight to be given to distances between incomes at different parts of the distribution. The most frequently used values are $a = 0$, 1, and 2, which result respectively in the GE index corresponding to the mean log deviation, the Theil index, and half the square of the coefficient of variation (CV).

$$\frac{1}{a^2 - a}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i / \bar{y}\right)^a - 1\right] \qquad (1)$$

By choosing $a = 2$, the analyst gives most weight in the calculation to income differences at the top of the distribution. With $a = 0$, most weight is given to the differences at the bottom of the distribution, while $a = 1$ represents an intermediate position.[14]

[14] In our case the incomes of the groups do not overlap (unlike incomes of people in different regions or ethnic groups). We could therefore have decomposed the Gini coefficient, the most popular measure of income

Table 4 shows the results of decomposing gross individual income in the EFS using these three cases, taking men and women together. We use three different variants of the bands: the 33 bands used in the 2004/5 OMN, the 39 bands used since 2005/6, which include six additional bands for higher incomes, and a variant in which we drastically collapse the 2005/6 bands into just 8 bands, chosen so as to contain roughly equal numbers of individuals. In each case we report the percentage of the variation of income that is within-band – the information that would be lost in a banded variable. The figure in brackets is the proportion of the within-band variation that comes from the top unbounded interval.

The results are striking. If variation is measured with the half-CV^2, we conclude that the 2004/5 OMN banding loses as much as a half of the variation in income. But almost all of this is in the top unbounded interval (which contains 1 in 12 of the sample). On the other hand, use of the mean log deviation, implying most interest in differences at the bottom of the distribution, results in almost all the variation being between band. The banding in this case implies only a very small lost of information. Use of the Theil index, a popular measure of income inequality, leads to the conclusion that little more than 10 percent of the variation would be lost through banding. Looking down the rows in the table reveals that the adding of the extra income categories in the 2005/6 OMN will only have reduced the loss of within-band variation by a moderate amount as measured by the half-CV^2. The final line shows that a large reduction in the number of income groups (albeit one designed to achieve an even spread of the sample across the new bands) would lead to only very modest increases in information loss.

The precise figures in Table 4 depend on the particular setting in terms of the distribution concerned and the sets of income bands evaluated. However, two general messages are clear. First, a conclusion on the extent of information loss depends on whether one's principal interest lies in differences towards the top of distribution or towards the bottom. Second, a large amount of variation is between band even when there is only a small number of income bands. In choosing bands for single-question surveys, designers can benefit from a literature on optimal grouping e.g. Aghevli and Mehran (1981) and Davies and Shorrocks (1989).[15] This addresses a converse problem: statistical offices that publish

---

inequality, since in the case of non-overlapping groups this index too is additively decomposable. However, we prefer the GE indices since by varying the parameter *a* we can allow for different views on the part of the distribution that is of most interest.

[15] See also Cox (1957), who illustrates the problem with the example of a continuous variable measuring health status, noting that the choice of bands may be influenced by views on the desirability of certain values, as in the case of blood pressure. Interest in grouping issues stretches back to Pearson (1920) and beyond.

tabulated data on income distributions based on continuously measured variables need to choose income ranges that preserve as much variation in the data as possible.

Implications of the loss

What are the implications of the banding for users of the data? The bands should pose little problem if income is to be a dependent variable. Standard computer packages now contain procedures for handling grouped dependent variables – see the discussion of item non-response in Section 2. But most users of single-question surveys will see income as a potential correlate of another variable under investigation, i.e. income will be seen as an explanatory variable, which is our focus here.

Many of these users will be unconcerned. For example, the researcher who wishes to cross-tabulate a categorical variable of interest (e.g. voting intentions, method of travel to work, etc) against income would be perfectly happy with the Omnibus survey's 33 bands. Indeed the detail would be substantially greater than needed, with the income categories being collapsed for the analysis. The banding is sufficiently fine for approximate quintile groups or decile groups to be identified. In fact, were a continuous variable available, as in the FRS or EFS, income groups would need to be created for the cross-tabulation and hence information discarded.

The more interesting case is the user who does want a continuous measure but is confronted with a categorical one. The classic example is the researcher wanting to do regression analysis using income as one of the explanatory variables. A common practice is to create a continuous variable by allocating individuals to the mid-point of their income groups, with individuals in the top unbounded interval assigned to an estimate of the group mean (which might be taken from external sources, e.g. the FRS in the case of the OMN or BSA top intervals).[16]

It may be tempting to conclude that since the 'mid-point' variable measures the unobserved continuous income variable with error, there must be attenuation bias in its estimated coefficient. In a simple OLS regression model with one explanatory variable, the text book result is that classical measurement error in that variable leads to a downwards bias in the parameter estimate – the 'iron law of econometrics' (Hausman 2001). But the

[16] Another common practice is to use a set of dummy variables corresponding to the income bands. But this is impractical if the bands are large in number (as in the OMN and BSA) and many users in any case may be unable to resist their urge for a continuous variable.

measurement error in this case is *not* classical in form, an IID variable uncorrelated with the unobserved true values. Define the unobserved continuous measure of income across which bands have been placed as X_i and the 'observed' mid-points, Z_i. Within each band the measurement error, $u_i = Z_i - X_i$, which is absorbed into the regression equation's error term, has perfect negative correlation with X_i. Given a uniform distribution within each interval, the set of mid-points, Z_i, have zero correlation with the values of u_i, other than through the top interval where the errors above the chosen mid-point are unbounded. The equation's error term is therefore virtually uncorrelated with Z_i. In practice, the within-interval distributions are not uniform, but in our experiment reported below we still find correlations of Z_i and u_i that are very close to zero.

The properties of regression estimates using banded data are well-established.[17] Hsiao (1983) provides a clear summary – underlining the importance of the uniform distribution for consistency of the OLS estimator. Manski and Tanner (2002) is a recent extension. Table 5 illustrates using EFS data. We model individuals' expenditures on alcohol and on clothing (recorded in the two-week EFS expenditure diaries) as a function of their gross individual income (in each case selecting only individuals with positive expenditures). We first regress (log) expenditure on the 'true' continuous income variable in the survey, X_i, and then on the 'mid-point' variable, Z_i. We show results first with the mid-points based on the 2004/5 OMN income bands and then on the greatly collapsed bands used in Table 4 (less one, since the top band is now set at £36,400). We experiment with both double-log and semi-log functional forms, excluding individuals with very low incomes and very high incomes respectively to improve the model fit.

The estimated coefficients when the mid-point variables are used are very close to those obtained with the continuous variable, reflecting the properties established in the literature. (The estimated standard errors are also very similar but caution is needed in their interpretation as the measurement error introduces heteroskedasticity – through the increasing width of the bands as income rises – that has not been allowed for.)

However, several caveats are needed. First, the exclusion of individuals with very low or very high incomes is essential for the pattern of results in this example. When we estimate models that include all individuals, irrespective of their level of income, the estimated

[17] The original motivation was the need to estimate models in an era where computers could not handle the 'embarrassingly large quantity of information' (Prais and Aitchison 1958) present in microdata. Hence observations on dependent and independent variables, both measured continuously, were grouped into cell frequencies in cross-tabulations, with the mean values of the observations in each cell employed in the regression.

coefficients on X_i and on Z_i differ substantially. This is a reminder that the properties established in the literature for 'mid-point' regressions apply only to well-specified models, where the analyst is estimating the right model for the data. This underlines the need for careful exploration of the data before selecting a functional form.[18] Second, the maintained assumption throughout this section, that individuals do report income in the correct band, is a strong one. Both our earlier results comparing distributions and the 'direct' literature comparing answers to different types of questions cast considerable doubt on its suitability. As a consequence, in practice there is very likely to be measurement error bias in parameter estimates obtained with 'mid-point' variables, even in models that apparently fit the observed mid-point data satisfactorily, and the form of that bias is hard to judge. Third, the properties summarised in Hsiao (1983) and illustrated with our empirical findings in Table 5 refer only to regression models for continuous dependent variables. Many outcome variables of interest in social surveys are categorical (e.g. voting behaviour, attitudes, employment status). Manski and Tanner (2002, Table VII) compare results from a binary logit model for home ownership estimated using a mid-point income variable and a conventional maximum likelihood approach with results from models estimated with modified minimum distance and maximum score methods. The latter do not assign individuals to the mid-points of the income bands and impose no assumptions on the within-band distribution of income. The comparisons underline the need for caution when using banded data to explain categorial outcome variables.

6. Conclusions

Single-question surveys of income are common given both the importance of income to the investigation of many social phenomena and the competing demands from other topics in the design of questionnaires for social surveys. In the UK, there is also a recurring debate on whether to include a question on income in the decennial census (e.g. Collins and White 1996, ONS 2006).[19] It is therefore important that the quality and nature of single-question data are assessed.

There is an important distinction between a single-question on individual income and one on the household total. We find from comparisons of distributions in single-question surveys with those in surveys collecting detailed income data that the household total appears

[18] A more suitable functional form, fitting the whole sample, could certainly be found for the OMN data, involving more than a single parameter in income or log income.
[19] The 2007 Census Test included a question to each individual in the household on total gross income, with 8 income bands. See http://www.statistics.gov.uk/census2001/pdfs/2007_test_H1_form.pdf.

to be collected much less well than individual income. The differences are especially notable at the bottom of the distribution. The comparisons for men show that single-questions on individual income can result in distributions that correspond very closely to those based on detailed income data, even at the top of the distribution.

While individual income appears better measured, there are notable differences between men and women, between people of working age and the elderly and, among women, between those with children in the household and those in childless households. These differences suggest groups where greater probing or reminding of possible income sources prior to the single question may be especially useful.

Although collection of individual income alone is likely to produce more accurate answers, the user of the data is then left without an estimate of the household total if, as is common, only one person in each household is interviewed. We showed the relationship in the UK between individual and household income. The results are again less encouraging for women. Women with low individual income are much more likely to be in households where there are other substantial sources of income than are low income men. Individual income and household income per adult have a lower correlation for women than for men. Nevertheless, we argued on several grounds that individual income data have considerable value.

Lastly, we analysed the banding of single-question data. This results in a loss of information that must be balanced against the much reduced costs of data collection. But the loss may be quite small, although we showed how the verdict depends on what part of the distribution is of most interest. We argued that the loss will matter little to many users of the data. We then summarised the implications for users who want continuous measurement of income for use as an explanatory variable in regression models.

In addressing these issues, we posed the situation as a trade-off. On the one hand, detailed questions on income lead to greater accuracy and more information but at greater survey cost. On the other, single-question surveys collect income data at much less cost but there are losses of accuracy and information. Irrespective of how one views this trade-off, the detailed-question surveys will always have a place, as they provide information on each component of income, as well as on the total for an individual or a household. As far as the single-question surveys are concerned, there are definite losses but they do not seem to be catastrophic. We need therefore to quantify the losses, which is what the paper has done.

Appendix

Composition of OMN and FRS samples

We report differences between the composition of our selected samples of 2004/5 OMN and FRS data.

Gender. The OMN includes a slightly higher share of men, 46.4 percent compared to 48.2 percent in FRS.

Age. The average age of women is the same in the two surveys, 48.7 years; however, men are on average over two years older in the OMN, 49.4 years compared to 47.1 years in the FRS. The share of men that are aged up to 40 is seven percentage points less in the OMN.

Employment status. Both surveys contain a variable measuring ILO activity status: employed, unemployed or inactive. The percentage in each category is strikingly similar in the two surveys for both men and women. (Differences appear only in the first decimal place.) However, once we focus on the people of working age (defined as up to age 59) the OMN figure for men in employment is three percentage points higher, 85.7 percent compared to 82.7 percent in the FRS. A comparison of income measurement in FRS and the Family Expenditure Survey (the forerunner of the EFS) noted that the FRS 'tends to over-state the numbers not in employment' (Department of Social Security 2000: 12).

Education. While the OMN has quite detailed information on respondents' educational attainment, the FRS collects only limited information: whether the respondent has a degree, 'another kind of qualification', or neither. The percentage with a degree is very similar in the two surveys, 17.4 in the OMN and 18.7 in the FRS.

These results show the composition of the OMN and FRS samples – in terms of the variables concerned – to be very similar. The lower share of younger men in the OMN may be expected to slightly reduce estimates of mean income relative to the FRS. However, the OMN has a slightly higher share of working age men in employment, which should have the opposite impact.

The OMN and BSA Income Bands

Tables A1 and A2 give the bands and the distribution of our samples across them.

Table A1: Cumulative frequencies (percent), individual income, OMN, FRS and EFS

Income band	Men			Women		
(£s pa)	OMN	FRS	EFS	OMN	FRS	EFS
Zero	1.4	2.5	1.9	2.2	2.9	2.3
less than £520	2.1	3.3	2.9	4.3	4.1	3.9
£520 to less than £1,040	2.3	3.7	3.3	6.4	5.5	5.4
£1,040 to less than £1,560	2.6	3.9	3.6	8.3	7.2	6.8
£1,560 to less than £2,080	3.3	4.2	4.0	9.8	8.7	8.2
£2,080 to less than £2,600	4.1	5.1	4.7	14.5	10.7	11.1
£2,600 to less than £3,120	5.2	6.0	6.0	17.2	14.9	15.2
£3,120 to less than £3,640	6.2	6.6	6.9	20.1	17.2	17.4
£3,640 to less than £4,160	7.6	7.5	8.2	22.8	19.7	20.0
£4,160 to less than £4,680	9.6	8.5	9.6	26.0	22.1	22.9
£4,680 to less than £5,200	11.4	9.6	10.6	30.7	24.8	25.3
£5,200 to less than £6,240	15.3	12.2	13.8	38.0	30.0	33.3
£6,240 to less than £7,280	17.9	15.2	17.5	44.2	35.7	39.1
£7,280 to less than £8,320	21.2	19.1	20.9	48.4	40.9	44.4
£8,320 to less than £9,360	24.8	23.0	24.3	53.1	46.8	49.1
£9,360 to less than £10,400	28.4	26.9	27.4	58.0	51.9	53.8
£10,400 to less than £11,440	32.8	30.8	31.0	61.9	56.6	58.0
£11,440 to less than £12,480	36.1	34.9	34.0	66.1	61.1	61.8
£12,480 to less than £13,520	39.5	38.6	37.7	68.9	65.0	65.2
£13,520 to less than £14,560	42.0	42.2	41.2	72.3	68.9	68.9
£14,560 to less than £15,600	46.4	45.6	44.8	75.4	72.2	71.7
£15,600 to less than £16,640	50.3	49.0	47.6	77.8	74.9	74.8
£16,640 to less than £17,680	52.6	52.2	50.8	80.0	77.5	77.2
£17,680 to less than £18,720	55.5	55.3	54.1	81.8	79.9	79.4
£18,720 to less than £19,760	58.1	58.2	57.2	83.1	82.1	81.4
£19,760 to less than £20,800	62.3	61.1	60.0	85.5	84.1	83.3
£20,800 to less than £23,400	68.5	67.1	66.6	88.0	87.5	86.8
£23,400 to less than £26,000	74.1	72.6	72.6	90.9	90.3	89.5
£26,000 to less than £28,600	77.5	77.2	77.6	93.0	92.4	91.4
£28,600 to less than £31,200	81.0	81.1	81.3	94.5	94.1	93.5
£31,200 to less than £33,800	83.8	84.1	84.0	95.8	95.3	95.0
£33,800 to less than £36,400	87.1	86.7	86.8	96.6	96.4	96.2
£36,400 or more	100.0	100.0	100.0	100.0	100.0	100.0

Table A2: Cumulative frequencies (percent), household income, BSA and FRS (individuals in multi-adult households)

	Men		Women	
Income band (£s pa)	BSA	FRS	BSA	FRS
less than 4,000	1.2	0.6	1.5	0.6
4,000 to 5,999	3.0	1.0	3.8	1.0
6,000 to 7,999	5.7	1.8	6.9	1.8
8,000 to 9,999	10.0	4.0	11.7	3.9
10,000 to 11,999	14.5	7.0	16.3	7.0
12,000 to 14,999	21.4	13.1	22.6	13.5
15,000 to 17,999	26.4	19.0	28.6	19.7
18,000 to 19,999	30.9	22.7	34.8	23.7
20,000 to 22,999	36.9	29.0	40.4	30.3
23,000 to 25,999	44.0	34.8	47.3	36.4
26,000 to 28,999	48.3	41.0	54.9	42.6
29,000 to 31,999	55.2	46.9	61.4	48.7
32,000 to 37,999	64.8	58.2	68.8	59.9
38,000 to 43,999	72.0	67.4	76.6	69.1
44,000 to 49,999	80.2	74.6	81.9	76.2
50,000 to 55,999	85.5	80.2	86.1	81.4
56,000 +	100.0	100.0	100.0	100.0

References

Atkinson A B and Micklewright J (1983) 'On the reliability of income data in the Family Expenditure Survey', *Journal of the Royal Statistical Society*, Series A.

Aghevli B and Mehran F (1981) 'Optimal grouping of income distribution data' *Journal of the American Statistical Association*, 76 (373): 22-6

Berthoud R (2004) *Patterns of Poverty in Europe*, Bristol: The Policy Press

Böheim R and Jenkins S P (2006) 'A comparison of current and annual measures of income in the British Household Panel Survey', *Journal of Official Statistics*, 22(4):1-27.

Burton P, Phipps S and Woolley F (2007) 'Inequality within the household reconsidered' in S P Jenkins and J Micklewright (eds.), *Poverty and Inequality Re-examined*, Oxford: Oxford University Press

CAF and NCVO (2005) *UK Giving 2004/05*, West Malling: CAF.

CAF and NCVO (2006) *UK Giving 2005/06*, West Malling: CAF.

Collins D and White A (1996) 'In search of an income question for the 2001 Census' *Survey Methodology Bulletin* 39(7): 2-10

Cowell F A (1995), *Measuring inequality*, (2nd ed.) Prentice Hall/Harvester Wheatsheef

Cox D R (1957) 'Note on grouping' *Journal of the American Statistical Association*, 52 (280): 543-7

Davies J B and Shorrocks A F (1989) 'Optimal grouping of income and wealth data' *Journal of Econometrics* 42: 97-108

Department of Social Security (2000) *Comparisons of income data between the Family Expenditure Survey and the Family Resources Survey*, London: Government Statistical Service Methodology Series 18

Foster K and Lound C (1993) 'A comparison of questions for classifying income' *Survey Methodology Bulletin* 32(1): 1-7

Frosztega M (2000) 'Income distribution data for Great Britain: robustness assessment report', Department of Social Security.

Hausman, J. (2001), 'Mismeasured variables in econometric analysis: problems from the right and problems from the left', *Journal of Economic Perspectives*, 15: 57-67

Hsiao C (1983), 'Regression analysis with a categorized explanatory variable' in festschrift for Ted Anderson edited by S Karlin, T Amemiya and L Goodman (eds.), *Studies in Econometrics, Time Series, and Multivariate Statistics*, New York: Academic Press

Jenkins, S P (1991) 'The measurement of income inequality' in L Osberg (ed.) *Economic Inequality and Poverty: International Perspectives*, ME Sharpe, Armonk NY and London

Kemsley W, Redpath R and Holmes, M (1980) *Family Expenditure Survey Handbook*, London: HMSO

Manski C and Tanner E, (2002) 'Inference on regressions with interval data on a regressor or outcome', 2002 *Econometrica*, 70(2): 519-46

Micklewright J and Schnepf S V (2007) 'Who gives for overseas development?', Southampton Statistical Sciences Research Institute, Working Paper A07/05, University of Southampton

ONS (2006) 'The 2011 Census: assessment of initial user requirements on content for England and Wales – income'.
www.statistics.gov.uk/about/consultations/downloads/2011Census_assessment_income.pdf

ONS (2007) 'The Omnibus Survey'
http://www.statistics.gov.uk/about/services/omnibus/default.asp, accessed 15.01.07

Pearson K (1920) 'On the probable errors of frequency constants' *Biometrika*, 13(1): 113-32

Prais S J and Aitchison J (1954) 'The grouping of observations in regression analysis', *Review of the International Statistical Institute*, 22: 1-22

Thomas R (1999) 'Question Bank commentary: income', The Question Bank, University of Surrey, http://qb.soc.surrey.ac.uk/topics/income/thomas.htm

Table 1: Percentiles of individual gross income, OMN, FRS and EFS

Men	OMN	FRS	EFS	OMN as % of FRS	OMN as % of EFS
P5	3,025	2,340	2,808	129.3	107.7
P10	4,809	5,304	4,926	90.7	97.6
P25	9,407	9,828	9,580	95.7	98.2
P50	16,563	16,900	17,349	98.0	95.5
P75	26,640	27,248	27,336	97.8	97.5
P85	34,644	34,632	34,701	100.0	99.8

Women	OMN	FRS	EFS	OMN as % of FRS	OMN as % of EFS
P5	676	884	858	76.5	78.8
P10	2,103	2,444	2,520	86.0	83.5
P25	4,516	5,200	5,150	86.8	87.7
P50	8,657	9,984	9,566	86.7	90.5
P75	15,467	16,640	16,713	93.0	92.5
P90	25,074	25,688	26,731	97.6	93.8
P95	32,073	33,020	33,789	97.1	94.9

Note: OMN percentiles estimated with the assumption of a uniform distribution in the relevant range.

Table 2: Percentiles of individual and household gross income, OMN, BSA and FRS

	'Control group': individuals in single-adult households		'Treatment group': individuals in multi-adult households	
	OMN as % of FRS	BSA as % of FRS	OMN as % of FRS	BSA as % of FRS
Income concept	individual	household	individual	household
Men				
P5	59.5	76.8	146.3	70.3
P10	61.6	78.0	103.6	74.4
P25	69.3	92.1	101.7	81.6
P50	81.4	125.3	102.1	88.7
P75	87.4	118.7	100.2	91.5
P85	89.4	114.6	100.9	88.6
Women:				
P5	67.7	80.0	148.7	63.1
P10	68.8	72.0	80.3	69.5
P25	67.4	72.2	91.2	78.9
P50	77.9	84.5	93.9	83.0
P75	83.8	104.5	95.4	87.3
P85	94.7	107.1	96.9	89.1

Table 3: Individual gross income and household gross income per adult, FRS (£s pa)

	Men		Women	
	Mean	Std. dev.	Mean	Std dev.
Individual income	22,372	29,650	12,742	14,678
Household income per adult	17,906	18,758	16,687	16,386

Note: the unit of analysis in each case is the individual.

Table 4: Percentage of variation in income that is within-band and the proportion of within-band variation generated by the top interval, (EFS)

Banding	Mean log deviation ($a = 0$)	Theil index ($a = 1$)	½ CV^2 ($a = 2$)	% individuals in top interval
OMN bands 2004/05	3.3	12.4	50.2	8.4
	(0.50)	(0.99)	(0.99)	
OMN bands 2005/06	2.3	7.7	41.9	3.5
	(0.30)	(0.99)	(0.99)	
8 bands, top interval as for OMN 2005/06	8.8	10.5	43.3	3.5
	(0.08)	(0.72)	(0.97)	

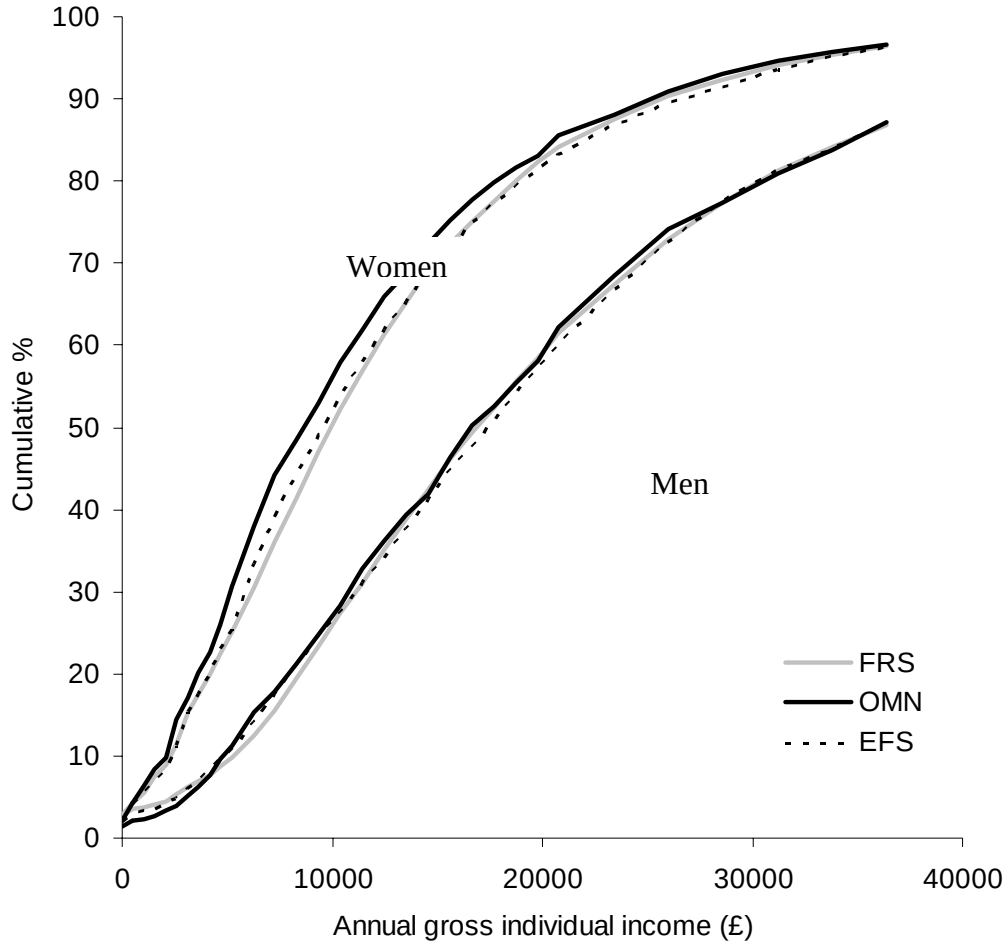
Note: the proportion of the within-band variation that is generated by the top band is shown in brackets. The top-band OMN band in 2004/05 starts at £36,400 and in 2005/06 at £52,000.

Table 5: Estimated coefficients on individual income in regressions of individual alcohol and clothing expenditure, EFS

	Income		Log income	
	Alcohol	Clothing	Alcohol	Clothing
'true' (continuous)	11.609	11.838	0.261	0.215
	(0.982)	(1.321)	(0.020)	(0.025)
'mid-point' 33 bands	11.699	12.180	0.249	0.210
	(1.023)	(1.343)	(0.019)	(0.024)
'mid-point' 7 bands	11.594	11.787	0.248	0.208
	(1.012)	(1.320)	(0.019)	(0.024)
Sample size	6,180	5,283	5,693	4,704

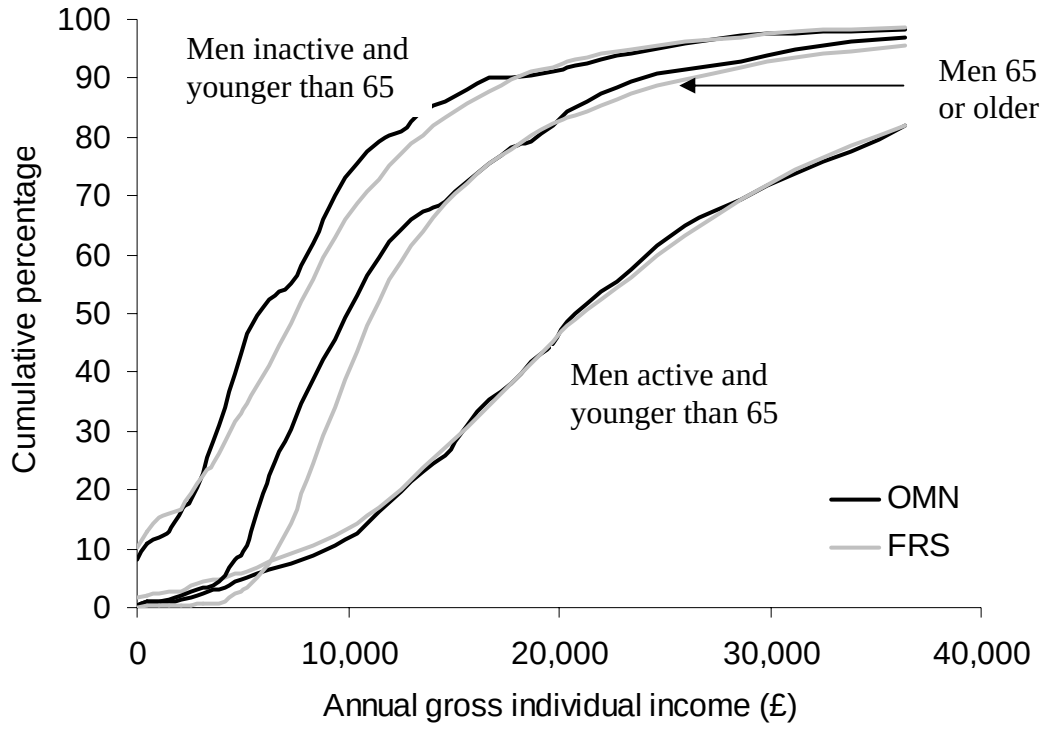
Note. Estimated standard errors are given in brackets. Alcohol and clothing expenditure relate to a two week period and are in logs. In the model with income in levels, individuals with income above the 99th percentile are excluded. In the model with income in logs, individuals with income below £3,120 p.a. are excluded. The 7 band variable has a top interval starting at £36,400, but otherwise is as the collapsed band variable used in Table 4.

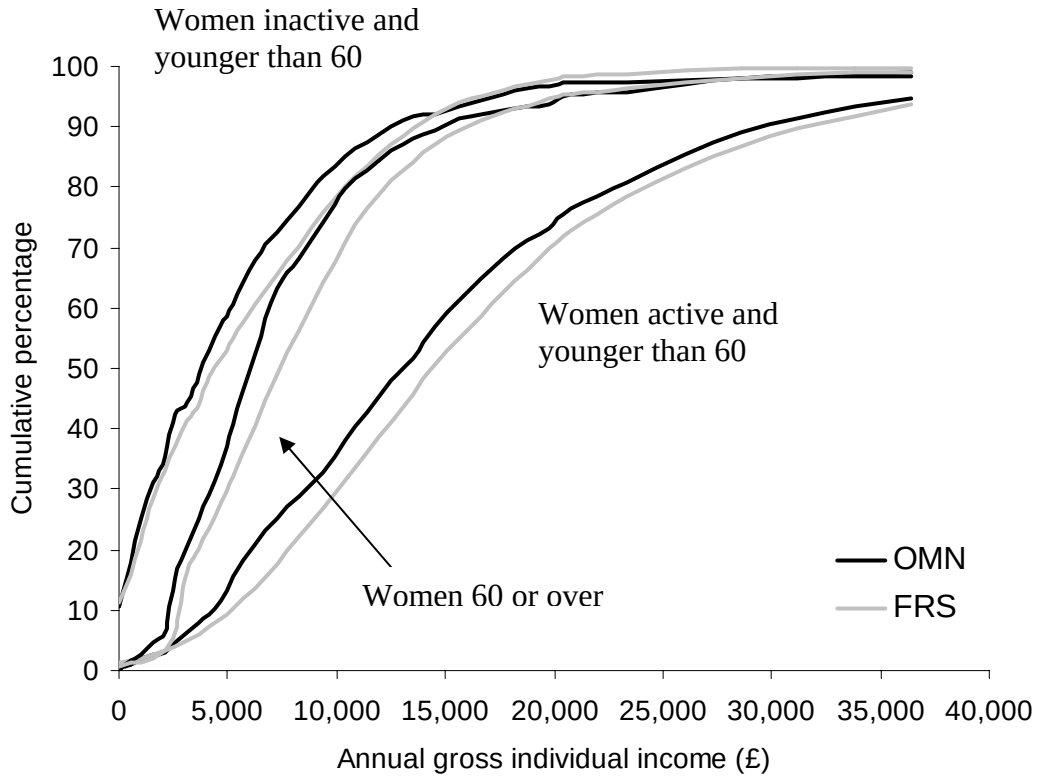
Figure 1: Distribution of individual gross income, cumulative frequencies (percent), OMN, FRS and EFS



Note: we include any negative amounts (caused by losses from self-employment) with the zeros in the FRS and EFS.

Figure 2: Individual income distribution by activity and age, cumulative frequencies (percent), OMN and FRS





Note: OMN sample sizes for men are 262 inactive and aged under 65, 492 retired, and 1307 active and under 65. Sample sizes for women are 450, 1276, and 834 respectively.

Figure 3: Individual income and household income per adult, FRS (£s pa, logs)

Men

Women

Note: Individuals with zero income are assigned £52. The correlation coefficients when the two variables are in levels rather than logs is 0.84 for men and 0.47 for women.