Session Number: Parallel Session 7A
Time: Friday, August 27, AM

Paper Prepared for the 31st General Conference of
The International Association for Research in Income and Wealth

St. Gallen, Switzerland, August 22-28, 2010

Income Reporting Behaviour in the SHIW

Andrea Neri and Roberta Zizza

For additional information please contact:

Name: Andrea Nerri
Affiliation: Bank of Italy

Email Address: ANDREA.NERI@bancaditalia.it

Income reporting behaviour in the SHIW *

Andrea Neri
Bank of Italy

Roberta Zizza
Bank of Italy

July 2010

Abstract

This paper analyzes the accuracy of income data in the Survey on Household Income and Wealth (SHIW) and presents a method to correct for response error. Survey data relating to the number of earning recipients and to amounts received are validated using external information from administrative and statistical sources. Our findings suggest that misreporting is particularly severe for income from self-employment, financial assets and rents, as well as from secondary jobs. As to the distribution of response error, about 15 per cent of respondents show a high probability of misreporting. Misreporting is more diffuse among males, the older, the self-employed and respondents at the higher end of the earnings distribution.

JEL classification numbers: J65, D21, L11.
Keywords: income distribution, response error, item response theory, SHIW, data accuracy.

[*] We wish to thank Andrea Brandolini, Giovanni D'Alessio, Leandro D'Aurizio, Andrea Filippone, Stefano Iezzi, Paolo Sestito, Raffaele Tartaglia-Polcini and two anonymous referees for their comments and suggestions. The authors are responsible for any mistakes. The views expressed here are our own and do not necessarily reflect those of the Bank of Italy. Correspondence: Bank of Italy, Economics, Research and International Relations Area, via Nazionale 91, 00184 Rome, Italy. Tel: +39 06 4792 3872 Fax: +39 06 4792 4764. Email: andrea.neri@bancaditalia.it, roberta.zizza@bancaditalia.it.

1 Introduction

Household surveys are a classic source of data on income. For one thing, they improve upon alternative sources, such as social security or tax records, in many respects: they do not suffer from the censoring of low-income earners (e.g. due to tax thresholds); the concept of income is not restricted to the base relevant for tax or social contribution payments but can be the most inclusive one (e.g. including in-kind receipts); administrative sources are usually found lacking in terms of an individual's characteristics and the unit is the individual, with no possibility of matching with information on other household members. On the other hand, household surveys are affected by non-sampling errors. A bias due to the selection into participation, which is usually not compulsory, is often at work. Moreover, response error may arise because of underreporting, recording errors, difficulties in understanding questions, and so on.

The paper studies the accuracy of data in economic surveys, in particular income data. This is an important aspect in empirical research. It helps data users to understand the results and limits of their analysis: for instance, response error could bias regression parameters for variables whose absolute values are biased; in cross-country comparisons, it is important to know whether heterogeneities are due to differences in data quality or to economic behaviours. Moreover, several policy recommendations are based on the analysis of survey data and reliability issues are thus of great importance, not only for the scientific community, but also for policy makers and the general public. Finally, having a good understanding of the distribution of response errors - whether these errors are concentrated in a few respondents or not, what are the characteristics of misreporters - can in fact help data producers to improve the accuracy of their data.

Response error is defined as the difference between the objective truth relating to a respondent's income and what is reported about that income. The consequences of response errors are twofold. The first is bias, which affects the size and the sign of the discrepancy between the actual and the reported amount earned. The second is variability or reliability, which refers to the distribution of errors around their average values. Here we focus on the issue of the potential bias due to response error. As we lack administrative records to match exactly at the individual level, we develop a set of statistical exercises (some original, some borrowed from previous work) to derive a set of proxies of response error for each respondent.

The data set we rely upon is the Bank of Italy's Survey on Household Income and Wealth (hereinafter the SHIW), which is conducted biennially and is representative of the Italian population (Banca d'Italia, 2006).

To motivate the paper, we assess the magnitude of response error in the 2004 edition of the SHIW using external information published by the Italian National Statistical Institute (Istat) as a yardstick. A comparison of income estimates from the SHIW with those from the national accounts (NA) shows that the distance between the two sources varies substantially with the income type. In particular, data on payroll income and on income from pensions replicate quite well the corresponding macro figures, while data on income from self-employment and capital tend to be less accurate (table 1).[1] The same has been found in a study by Biancotti *et al.* (2008), with regard to the issue of reliability. As to the occupational status, table 2 shows that the number of workers (broken down into dependent and self-employed) in the SHIW is of the same order of magnitude as that in the Labour Force Survey and in the NA[2]; on the contrary, the number of dependent jobs held equaled around 20 million in the NA, against 16.6 million in the SHIW (underestimated by almost 20 per cent; table 3). People earning from self-employment totaled 6.3 million in the SHIW and around 11 million in the NA (underestimated by more than 40 per cent). Since the micro and macro estimates for the main job are in line, this suggests that multiple positions are underestimated.

The paper is organized as follows: Section 2 reviews the literature; Section 3 describes the adjustment procedure and Section 4 shows its results, with a focus on the characteristics of the misreporters; Section 5 concludes.

[1] See the appendix for details on the methodology and on the results of the comparison exercise.

[2] The occupational status in the SHIW refers to the prevailing condition, that in the Labour Force Survey to the condition in a reference week.

2 Income response error: a review of the literature

Income reporting in surveys is generally a two-stage process involving the reporting of income sources in the first and of the corresponding amounts in the second (see Moore *et al.*, 2000). Errors can occur at either stage. An entire source of income can be omitted (reported) even if it has (not) actually been received. Or, the source of income may be reported correctly but the amount received for that source can be misreported.

A theoretical framework for the cognitive factors in reporting is proposed by Tourangeau *et al.* (1984), who distinguish three stages in the cognitive process required to answer a survey question: understanding, retrieval and response production.

In either of these stages there are areas of potential error. One reason for this is the understanding of terms. Tourangeau *et al.* (2000) identify seven major types of comprehension problems: grammatical ambiguity, excessive complexity, faulty presuppositions, vague concepts, vague quantifiers, unfamiliar terms and false inferences. For instance, for questions relating to non-cash benefits it is not always possible to include in the questionnaire a fully exhaustive list of examples to help the respondent. Therefore, it may not be clear to him or her which items to include.

A second reason has to do with retrieval problems, that is the failure to bring to mind information stored in long term memory (Groves *et al.*, 2004). Even if the respondents have fully understood the question, they may fail to retrieve the correct information. Lack of knowledge is the primary cause. For instance, in the SHIW the respondent is often asked to report the income sources earned by all the household members. Even if the respondent is selected as the most knowledgeable person in the household, he or she may not know detailed income proceedings of the other components (i.e. in the case of a working son still living with his parents). Recall problems often arise as a result of the presence of multiple jobs and/or of the low level of some incomes, e.g. in case of temporary jobs or small forms of financial assistance. As a matter of fact, classical interference and information-processing theories suggest that as the number of similar or related events occurring to an individual increases, the probability of recalling any one of those events declines (Groves *et al.*, 2004). Moreover, when the required information is not directly retrievable from their memory, respondents adopt a reconstruction strategy which is unlikely to be completely successful. For instance, it can lead to a sort of rounding (Pudney, 2008).

In the final stage, after recalling the requested information, the respondent adopts a response strategy. Deliberate underreporting is probably the major cause of response error at this stage. Nonetheless, besides deliberate prevarication, there are further possible sources of error. Questions on household income may be considered impolite or very personal, making the respondent more reticent. Another source of error may come from the interaction between the interviewer and the respondent. For instance, if the respondent belongs to a very rich household he or she may decide to underreport their income because of a desire for “social conformability” with the interviewer, or because of a fear of being robbed. This could be considered as a special case of the so-called “social desirability bias” (Bagozzi, 1994), namely, the tendency for individuals to present themselves in a way that makes them look positive in respect of cultural norms or standards. On the flip side, response error and overreporting in particular, may arise from the respondent wanting to impress the interviewer. Other studies show that the “social presence” of the interviewer stimulates the respondents to consider social norms at the judgment phase of their response, leading to response errors (Turner *et al.*, 1998). Interviewers also play a motivational role, setting expectations for respondent performance (Groves *et al.*, 2004; Fowler and Mangione, 1990). The higher the interviewer’s concern about accuracy and their request for “precise answers” rather than “general ideas”, the higher the quality of the data provided by the respondents. Indeed, the practice of matching interviewers and respondents on demographic traits is also widespread in surveys, especially in the case of sensitive questions.

Response error is also affected by the data collection method. For instance, according to the existing literature, computer-assisted personal interviewing (CAPI) should reduce errors (Couper *et al.*, 1998).[3] On the flip side, the respondent might not feel at ease if a computer is used during the interview.

[3] Interviewers enter responses into a computer file. The interview software ensures that questionnaire skip patterns are followed correctly and that entered responses are reasonable. If not, the interviewer is asked for further confirmation.

3 The adjustment procedure for response bias

There is an extensive literature on models dealing with response or measurement error. We can separate the different approaches to response errors into two broad categories, according to whether or not identification is achieved with external information. In the absence of external information, identification can be achieved in some particular cases. For instance, a large literature gives general identification results for the non-parametric regression model with response error under the classical assumption (for example, Schennach, Hu and Lewbel, 2007). When point estimation of a parameter is not feasible, another solution is the identification of bounds or of intervals in which the parameter lies (see, among others, Bollinger, 1996; Imbens and Manski, 2004).

When external information is available, several possibilities arise. For instance, identification can be achieved with repeated observations of a variable within the same survey (Kan and Pudney, 2008) or using panel data (Griliches and Hausman, 1986). The most common approach consists of matching survey and administrative data. Matching is done by either asking respondents for their relevant identification number directly, or using statistical matching based on personal data such as name, location and date of birth. For example, Bollinger and David (2005) match data from the Survey of Income and Program Participation (SIPP) with those in the Food Stamps register to study respondents' latent propensity to cooperate in the Survey. Another example is Pedace and Bates (2001) who study income misreporting by matching SIPP to social security earning records.

Our method lies in between those two approaches. As we lack respondents' personal identifiers we cannot perform any exact linkage with administrative data. Nonetheless, we have collected several external data on earners and on income earned that can be used for validating income survey data. Namely, our identification strategy rests on a sequence of statistical exercises devised according to the kind of external information available. In each step of the adjustment, we use the adjusted data from the previous steps. We also perform a robustness check to assess whether the ordering of the adjustment steps affects the results.

Among the findings from our assessment on data accuracy are: a serious underestimation of income from self-employment and of income from financial assets and rents; the underreporting of secondary income sources. In the rest of this section we cope with these issues by either implementing an ad hoc adjustment procedure (for secondary jobs) or modifying the standard approach (for self-employment income) or using methodologies already developed in the literature (for capital income). Preliminary to this, we cope with the issue of nonresponse, which is relevant in most household surveys.

3.1 Adjustment for unit nonresponse

The first step of the adjustment relates unit nonresponse. Here we propose a new set of weights which exploits recently available external information drawn from the EU-SILC (*European Union Statistics on Income and Living Conditions*) survey. For Italy and starting from the 2004 wave, EU-SILC survey data are linked with administrative records containing information on self-employment income, on wages and salaries and on public pensions. Survey data are then adjusted both on recipients and on amounts. When survey respondents fail to report a source of income, it is imputed using administrative records. Moreover, self-employment income is set to the maximum value between the net income resulting from the tax source and the net income reported in the survey (Consolini *et al.*, 2006; Istat, 2009).

We use the adjusted information on income recipients to build a new set of weights for SHIW survey modifying the actual final raking step (for a description see Faiella and Gambacorta, 2007). Present raking forces weights to conform marginal distribution with respect to gender, age (4 classes), geographical area (3 classes) and municipality size (4 classes). We modify this scheme including, as a further constraint, the distribution of individuals according to their main source of income: wages and salaries, self-employment income, pensions, capital income, financial assistance[4].

The proposed new weighting scheme aligns the two populations along a pivotal variable (*main source of income*), which is related to respondents' income. Due to the new weighting scheme the percentage of recipients increases for any source of income, at the expense of the

[4] This requires that all incomes are available at the individual level. In the EU-SILC survey, only income from financial assets is collected at the household level, while in the SHIW also income from real investments is available at the household level. In order to make data comparable, the latter revenues are attributed to the members of the household using the following assignment rules: in both surveys, income from financial assets is equally redistributed among earners; in the SHIW, income from non-financial assets has been divided among owners using available information on shares owned by each member.

percentage of non-earners (table 4). This is because wealthier households tend to have a lower propensity to participate in the survey (D'Alessio and Faiella, 2002), as corroborated by the fact that the new weight is related positively to education and negatively to household size. For income from self-employment and from property the increase is respectively equal to 3.4 and 5.6 percentage points. The increase in the coefficients of variation is not significant.

The new weights will be used in all the steps of the adjustment procedure. This will enable us to separate the effect of response error from the effect of non-participation if the following assumptions hold: first, the two sources of error are independent; second, unit non-response is correctly accounted for by the new weighting scheme.

3.2 Misreporting on secondary sources of income

In order to adjust income data from secondary jobs we develop an original procedure which again exploits the Italian section of EU-SILC as an external source of information. Due to the record linkage with administrative and tax records, we can assume that the number of income recipients is correctly estimated in EU-SILC and perform a statistical matching. We focus on underreporting only, as overreporting is unlikely since it increases respondents' burden significantly: as a matter of fact, for each income source a thorough set of information must be provided.

Using the new set of weights, we apply the following procedure. Let I_{it} be an indicator equal to 1 if the respondent i has reported the secondary source of income t ($I_{it=1}$) and 0 otherwise. Let I^*_{it} be the true latent respondent's status. We assume that whenever a respondent reports a source of income, his or her answer is "correct".

On the contrary, $I_{it=0}$ is a mix of "correct" answers and of response error. Namely, $I_{it=0}(\mathbf{x}) = I^*_{it=0}(\mathbf{x})\pi_{t0}(\mathbf{x}) + [1 - I^*_{it=0}(\mathbf{x})][1 - \pi_{t0}(\mathbf{x})]$ where $\pi_{t0}(\mathbf{x})$ is the (mixing) probability of correct reporting conditional on respondents' vector of characteristics $\mathbf{x}$. This probability is unknown for the SHIW, but can be estimated through EU-SILC using a set of individual characteristics $\mathbf{x}$, also available in the SHIW (gender, age, number of household members, educational attainment, geographical area and main source of income). It can then be extrapolated to SHIW respondents. Once a fitted probability $\widehat{\pi}_{t0}(\mathbf{x})$ is obtained for each respondent in the SHIW, a random experiment is used to impute secondary sources of income. The expected value of the random variable is set equal to the difference between the percentages of recipients calculated from the two surveys. This ensures that in the end the expected number of recipients is aligned.

Two underlying assumptions are needed. First, the two surveys are representative of the same population: the marginal distributions of the conditioning variables are not significantly different (also because of post-stratification). Second, conditional on common variables, the streams of income from the two surveys are independent, as commonly assumed in statistical matching.

In particular, we estimate the following models:

- for respondents whose main source of income is payroll, we estimate their probability of having a secondary source from either self-employment or pensions or other forms of assistance (table 5);
- for respondents whose main source of income derives from self-employment we estimate their probability of perceiving also pensions (table 5);
- for those whose main source of income is either pensions or capital we estimate the probability of also having a payroll or self-employment income (table 6).

As a whole having a secondary source of income is more likely among men, except in the case of employees receiving financial assistance, and in the North, except for retirees receiving a payroll income and for employees receiving a pension or financial assistance. Graduates are more likely to have a secondary source of income if their main income is from capital or from pensions or for employees who are also self-employed. The probability of receiving pensions pr any form of assistance is higher for less-skilled persons.

Table 7 summarizes the percentages of recipients by income source before and after the adjustment process. The share of self-employment and payroll income earners increases by 3.9 and 2.6 percentage points respectively. The extent of the adjustment increases with the respondent's education and decreases with household size. Moreover, it is higher for the Central and the Northern regions. It appears, therefore, to be positively related to the respondents' economic status.

Finally, income from these fitted secondary jobs are assigned with a random selection of income figures from the EU-SILC data (stratified according to gender and geographical area). Only secondary sources of income are used. In the case of payroll income and pensions, this imputation lowers the mean values (table 8): this is because income perceived in secondary dependent jobs is generally lower than in the principal ones. The opposite holds for self-employment income, whose average value increases after the imputation.

3.3 Income from self-employment

As a second step, we deal with the misreporting behaviour of the self-employed as regards their main source of income. The adjustment procedure exploits the relation between market value of the main residence and income from work, hence modifying the traditional approach developed by Pissarides and Weber (1989), based on the relation between food expenditure and income. This relation is estimated on a validation sample and then extrapolated onto the self-employed, under the assumption that the value of the main residence is correctly reported by all income groups while income reporting is accurate only in the validation sample. Previous research shows that house prices collected in the SHIW are quite in line with those from external sources (Cannari and Faiella, 2007). The adjustment allows for both overreporting and underreporting.

Our validation sample (346 observations) includes employees working in the public sector whose responses are deemed highly reliable by the interviewers. At the end of the interview the interviewer is in fact asked to evaluate the respondent's reliability in income reporting and his ability to understand income questions, as well as to rate the overall climate of the interview. For the validation sample we selected respondents with the maximum score (10) in these three items.

We use the following specification:

$$E\left[\frac{Y_i}{W_i} \mid \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\right] = \alpha + \mathbf{X}_1'\beta_1 + \mathbf{X}_2'\beta_2 + \mathbf{X}_3'\beta_3$$

in which the dependent variable is the ratio of individual income from labour Y to market value of main residence W. We prefer a relation at the individual rather than at the household level firstly because some of the right-hand side variables are individual (e.g. age, labour income class). Secondly, because in this way we can avoid deciding whether households composed of both self-employed and public sector employees must be included or not in the validation sample; that decision would have involved looking at the magnitudes of each income source, being in turn heavily influenced by the misreporting behaviour itself.

$\mathbf{X}_1$ contains three dummy variables indicating the labour income quartile the respondent belongs to, meant to capture any non-linearities in the relation. $\mathbf{X}_2$ includes the size of the house, a dummy variable for luxury dwellings and one indicating whether the dwelling has been (totally or partially) received as a bequest or gift. Last, $\mathbf{X}_3$ includes other control variables, such as age, geographical area, size of the municipality, the educational attainment of the interviewer. We also tried to account for mixed households by including dummy variables for the presence of a self-employed or for the existence of other sources of income, but no explicative power was added to the previous specification.

In order to make the two groups more similar to each other we build a new set of weights that align the distribution of validation sample to the estimated distribution of self-employed by geographical area (3 classes) and household income (5 classes). This auxiliary information is obtained from the SHIW data using the method by D'Alessio and Faiella (2002) to adjust for nonresponse.

As expected, the income-to-house-value ratio increases in the upper income brackets, while it decreases in the North, for larger and valuables dwellings (table 9). Estimated coefficients are then fitted to the self-employed in order to predict their expected income from work given the value of their house and their characteristics. Moreover, a random component from the distribution of the residuals is added to preserve the observed variability.

Income from self-employment rises on average by around 36 per cent due to the adjustment, 11 percentage points more than found in Cannari and Violi (1995), who applied the standard Pissarides and Weber approach to Italy.[5] The adjustment is higher for large households, for people living in the South and in the Centre, for women, for the youngsters

[5] As a robustness check, we have also replicated the standard method with our data and obtained an upward revision by about 30 per cent.

or for those with a low level of education (table 10).

3.4 Net income from financial assets and liabilities and property income

Income from financial assets is not directly collected in the SHIW, but computed applying different rates of return to corresponding categories of assets. Therefore our adjustment relates to household financial wealth and is based on the approach developed by D'Aurizio *et al.* (2006). They dealt with the issue of underreporting of financial assets and liabilities using a survey conducted in 2003 by a leading Italian banking group on its customers as a supplementary source of information. Each respondent answered questions on financial assets held at the bank; then, survey data were linked to the bank databases using an exact matching procedure, hence allowing a comparison between self-reported wealth and that resulting from administrative records.[6]

This correction allows for both over- and underreporting. As a matter of fact, the increasing complexity of households' financial portfolios makes the respondents' task difficult. Some respondents may confuse different financial instruments, thus providing "false positive" answers. Moreover, overreporting does not increase a respondent's burden significantly, since the only information asked is the asset value.

The average adjusted value is about 3 times the unadjusted one. The correction is higher for young respondents, the lowly educated and for those living in the South (table 11). The adjustment is not therefore positively related to the respondents' economic status. The adjustment procedure also has an impact on interest paid on financial liabilities, though to a smaller extent (table 12).

Regarding property income, previous studies, in particular Cannari and D'Alessio (1990, 2008), show that while the survey estimate of the number of total dwellings used as a principal residence is in line with the macro figure, this is not the case for other dwellings. The method we borrow from Cannari and D'Alessio (1990) to adjust income from rents relies on the assumption that the empirical distribution of the number of dwellings not used as a principal residence recorded in the SHIW is a discrete Poisson distribution conditional

[6] See the appendix for details on the methodology.

on a set of characteristics. In the absence of more precise information, all dwellings not used as a principal residence are assumed equally likely to be declared by the owners. The probability that one of these dwellings is declared in the SHIW can then be described by a binomial distribution. This distribution can be estimated and used to impute ownership of the missing dwellings, that is the difference between the number of dwellings owned by the households (excluding the main residence) recorded in the SHIW and the corresponding "true" figure derived from the Census.[7] As for secondary positions, the adjustment process does not allow for overreporting, that is held unlikely because it increases response burden significantly: in fact, respondents know that for each dwelling they report, they will be asked a further set of detailed questions.

On average, the percentage of owners increases by 2.4 percentage points. The average amount they hold increases by 11 per cent. There are no clear trends regarding the extent of the adjustment: corrections are higher for men, for those older than 64 years and for graduates (table 13).

4 Results of the adjustment procedure. An analysis of the characteristics of the misreporters

We first assess by how much the adjusted SHIW provides income aggregates which are closer to those estimated in the national accounts. The correction steps performed so far lead to an overall alignment of micro and macro figures (table 1). Major improvements are achieved for self-employment income and for entrepreneurial income and income from financial assets. Regarding the former, the SHIW now accounts for 88 per cent of the corresponding national account aggregate; it was 43 per cent before. Original SHIW data on income from financial wealth and from entrepreneurial activity accounted for only 13 per cent of the corresponding macro figure; corrected data represent 72 per cent of the same aggregate. Wages and salaries also benefit from the adjustment procedure, though to a smaller extent: originally 88 per cent of the corresponding national accounts' estimate

[7] The value for 2004 has been computed by inflating the figure for 2001 (the year of the latest available Census) with the net growth of the number of dwellings (see for details Banca d'Italia, 2008).

in 2004, they now stand at 94 per cent. Pensions and financial assistance are much less affected, though a slight gain is also obtained for this source of income.

With regard to the income distribution, non-parametric estimates performed before and after the adjustment show that modifications of the shape of the distribution vary according to the income type (figure 1). As a whole, the "ex-post" distributions are slightly more dispersed than the "ex-ante", with higher probability mass at higher income levels; this is particularly apparent for actual rents, for income from self-employment and for income from financial assets.

With reference to inequality, we look at two standard measures - the interdecile ratio and the Gini index - applied to disposable income and separately to each income source. These measures are calculated only on individuals with a given income source; hence, any variation after the adjustment can be due either to the inclusion of an entire income source originally omitted or to the correction of a misreported amount.[8] Total earnings, as well as most income sources, turn out to be less equally distributed after the adjustment; for income from financial assets only we obtain opposing indications from the two indices (table 14). Inequality rises considerably for income from independent work, due to the fact that we add many recipients of this income source as a secondary activity and revalue significantly underreported amounts through the "modified" Pissarides and Weber (1989) procedure. Payroll income and pensions and financial assistance are now slightly more unequally distributed, as we impute many small incomes of this kind from secondary positions. For actual and imputed rents inequality rises, as expected, as non-reported dwellings are imputed on the basis of the probability distribution of declared houses. Overall, the distribution of disposable income widens: the decile ratio rises from 6.8 to 7.9 while the Gini index rises from 0.385 to 0.427. A similar trend occurs when equivalent income, computed using the OECD modified equivalence scale, is considered.

Finally, to test whether the results are affected by the ordering of the adjustment steps, we fit our procedure with two alternative procedures: in the first one (alternative 1) we change the order of the steps, in particular by shifting at the end the correction for self-

[8] The figures relating to unadjusted income may differ from those in Banca d'Italia (2008) since we use a different method for computing capital income from financial assets (see Section 3.4).

employment income; in the second the order is left as in the baseline case, but adjusted values from previous steps are not considered (alternative 2). The results show that differences are negligible (table 15).

4.1 Who are the misreporters?

As a result of the adjustment procedure each individual is characterized by a set of 14 proxy indicators of misreporting regarding all their income sources. These indicators can help to assess respondents' misreporting behaviour in terms of both propensity and magnitude and to describe the salient characteristics of the misreporters through the framework of the item response theory. As many of these characteristics are not strictly exogenous to the event of misreporting (above all, the occupational choice) we do not give a causal interpretation to the results of these estimations; rather, we use them for descriptive purposes.

We consider two models. The first is the random intercept logistic model that is akin to the Rasch model (Rasch, 1960). This model specifies the probability of a correct response to item t ($Y_t = 1$) conditional on the unobserved "ability" of subject i (θ_i)

$$\Pr(Y_{it} = 1|\theta_i) = \frac{1}{1 + e^{[-(\theta_i - \delta_t)]}}.$$

The coefficient δ_t is basically the proportion of errors relating to item t. Therefore, the greater the δ_t, the more difficult the item. In our model we include a dummy indicator for each item to account for its level of difficulty. If a subject's trait level θ_i exceeds the difficulty of the item δ_t, then the probability of a correct response is greater than 0.5. Since in our application we model the probability of an "incorrect answer", signs are reversed and δ_t can be interpreted as the easiness of the item. Thus, given the respondent's latent trait, the easier the item the lower the probability of an "incorrect" answer.

In practice we estimate the model

$$\text{logit}[\,\Pr(Y_{it} = 1|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \theta_i)] = \alpha + \mathbf{X}_1'\beta_1 + \mathbf{X}_2'\beta_2 + \mathbf{X}_3'\beta_3$$

where Y_{it} is a dummy variable taking value 1 if the $i - th$ respondent is estimated to misreport on item t ($t = 1, .., 14$).

We construct the following 14 response dummy variables Y_{it}:

Items	
Y_{i1} = Additional payroll income	Y_{i8} = Income from shares
Y_{i2} = Additional self-empl. income	Y_{i9} = Income from mutual funds
Y_{i3} = Primary self-empl. income	Y_{i10} = Income from managed savings
Y_{i4} = Additional income from pensions	Y_{i11} = Interests on financial liabilites
Y_{i5} = Income from deposits	Y_{i12} = Income from actual rents
Y_{i6} = Income from government bonds	Y_{i13} = Income from imputed rents
Y_{i7} = Income from private bonds	Y_{i14} = Income from financial assistance

The second model exploits the information on the estimated amount of response error. Once a given answer to item t is marked as "incorrect", not only do we impute the ownership of that source of income, but also its expected value. It is therefore possible to end up with a distribution of misreported amounts for each item. This enables us to classify the response errors into four categories, according to the quartile of the estimated misreported amount on item t they belong to. We specify the following random-intercept proportional odds model:

$$\text{logit}[\,\Pr(Y_{it} > s|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \theta_i)] = \alpha + \mathbf{X}_1'\beta_1 + \mathbf{X}_2'\beta_2 + \mathbf{X}_3'\beta_3$$

where S are the ordinal categories with index s ($s = 1, 2, 3, 4$), based on the quartiles of the distribution of response error.

One important caveat is that we cannot assess whether the respondent has reported the "true" value of a given income source. We are only able to estimate the probability of their providing a wrong answer. The final set of outcome variables we construct are therefore proxies of response error. Each outcome can be written as $Y_{it} = Y_{it}^* + e_{it}$ where Y_{it}^* is the true latent outcome and e_{it} is the error we may add because of the adjustment process. Our results are unbiased only if the following three conditions hold:

1. the expected value of the error term is zero: $E(e_{it}) = 0$;

2. the error term e_{it} and the observable characteristics used in the adjustment are not correlated, namely $E(e_{it}|\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = 0$. This assumption does not appear to be very restrictive. It implies that all the adjustments do not produce biased results for a given set of characteristics. Indeed, this should not be the case in our application

since we employ several independent sources of validation data and it is unlikely that all of them produce biased estimates of the same coefficients;

3. e_{it} is not correlated with some unobservable characteristics which are not used in the adjustment process. Unfortunately this assumption cannot be made with certainty. Nevertheless, the measures of goodness of fit of our models are satisfactory in most cases.

Turning to the covariates used in the analysis, X_1 is a set of indicator variables, one for each income item. Their coefficients can be interpreted as the "item difficulties", reflecting the proportion of individuals who are estimated to provide an incorrect answer to a given item. The second group of regressors X_2 includes a set of classical controls: age and its square, education, occupational status, area of residence, household wealth. Finally, X_3 includes variables accounting for the data collection process referred to in Section 2. In particular, we include two dummies for proxy respondents and one dummy for the use of the CAPI method. Moreover, we include a variable representing the interviewer's assessment of the respondent's level of understanding of the questions (on a scale of 1 to 10; *comprehension*) in order to proxy for respondents' cognitive ability. Finally, we include three dummy variables indicating whether the interviewer and the respondent match in terms of gender, education and age. Namely, the variable $edu - inter$ takes value 1 if one person between the interviewer and the respondent holds at least an upper secondary school diploma, while the other has at most a lower secondary school certificate. The variable $age - inter$ takes value 1 when the difference in age is greater than ten years. These variables are meant to capture unobserved interactions between interviewer and respondent which are likely to happen during the interview influencing a respondent's attitude towards response error.

Tables 16 and 17 summarize the results. As to the item difficulty parameters in table 16, the higher their value, the higher the percentage of response errors. The overall picture is consistent with the preliminary analysis presented in the Introduction. The items relating to financial wealth components show higher percentages of response errors. One possible explanation for this is the increasing complexity of household financial portfolios, resulting in a higher cost of information retrieval for respondents. Income from financial assistance has the lowest misreported percentage, mainly because they only refer to a marginal share of the population.

Respondents' characteristics show a similar association with response error when considering either income sources or corresponding amounts. Misreporting is more likely among male respondents and among those living in the North. It also increases with age, probably because of the greater difficulties older people face to retrieve the necessary information to reply. It also increase with the level of educational attainment. This result is probably due to the adjustment process we use for secondary dwellings (see Section 3.4). As expected those who are self-employed show higher response errors than employees. Not only, as confirmed by previous research, independent workers are likely to have a higher propensity towards underreporting, but their income has a higher variability and therefore implies greater recall difficulties. Finally, greater errors are to be expected among more affluent households. That result is in line with prior expectations, not least because those households have a higher number of items to report on during the interview. It is therefore likely that some errors may occur at some stage.

It is worth stressing that the results of the models are strongly driven by the results of the adjustment procedure described in Section 3. In a sense, these models can be seen as tools that summarize and show the final results of the adjustment process. The within-subject correlation among the outcomes of the 14 items is affected by the sign and the magnitudes of the observable characteristics used in the analysis. For instance, if for each adjustment a positive association between self-employment and response error is found, this will result in a higher probability of finding "incorrect responses" along all the 14 items.

The use of the CAPI data collection technique reduces the chance of response error with respect to the paper-and-pencil interviewing (PAPI) method. On the contrary, the probability of misreporting increases in the case of proxy respondents or when respondents' cognitive ability of understanding questions are judged unsatisfactory by the interviewer.

Finally, all proxies for the existence of significant demographic differences between respondent and interviewer suggest that those differences increase the probability of response errors. As anticipated in Section 2, the respondent might feel more at ease with an interviewer whose social status more closely mirrors their own and the resulting climate might be more confidential; overreporting arising from the respondent wanting to make a positive impress on the interviewer will be then less likely. On the contrary, wealthy respondents might be tempted to underreport their income in the presence of an interviewer coming from a lower social class. Similar mechanisms may be at work when there are sizeable age differences or when they are of opposite gender. Our working assumption is that the fewer the social differences between interviewer and respondent, the lower the probability of response error.

Table 18 shows the distribution of respondents according to their expected number of reporting errors along the 14 items. Each respondent is classified as a “misreporter” in a given item if his or her estimated probability exceeds the average probability computed for that same item. These indicators of misreporting behaviour are then summed at the respondents’ level to provide a synthetic indicator of their behaviour.

Our findings suggest that respondents can be grouped into three clusters on the basis of their latent trait θ_i. Some 27 per cent of them can be deemed highly reliable, their expected probability of misreporting being close to zero. At the opposite side of the distribution, about 15 per cent of respondents are likely to misreport on at least four items.

The average amount of response error is estimated to be about € 4,000 per year (table 19). The distribution looks highly concentrated with a Gini index of 0.799. Around 10 per cent of respondents are estimated to have an average response error of about € 46,000 and they are responsible for two thirds of total misreporting.

Robustness checks involving the use of synthetic indicators - the number of misreported items and the average value of the adjustment - and the estimation of several specifications provide a consistent picture.[9]

5 Conclusions

In this paper we investigate misreporting behaviour in survey data on personal income. Data from the Bank of Italy’s 2004 SHIW have been checked against external sources in

[9] Estimates, not reported for the sake of brevity, are available upon request.

order to assess if, and to what extent, information provided by respondents were correctly reported. Our main results may be summarized as follows:

First, underreporting issues emerged as particularly serious with regard to income from financial assets and from self-employment: the adjusted income is estimated to be as large as three and 1.4 times the reported data respectively.

Second, we found that the estimated response error has a significant influence on income distribution. The Gini index increases for most of the income sources, but not for income from financial assets. The inequality level of total disposable income distribution is higher for the adjusted data than for the unadjusted ones.

Third, our findings suggest that respondents are predisposed to provide either accurate answers or poor data. Namely, around 15 per cent of respondents are likely to be highly unreliable. The misreporting pattern is higher among males, the well-educated, the self-employed and respondents at the higher end of the earnings distribution. Moreover, it increases when respondents and interviewers exhibit significant demographic differences.

Finally, the distribution of the magnitude of response error is highly concentrated. Around ten per cent of respondents are estimated to account for two thirds of the total amount.

6 Bibliography

Agostinelli, A. “Il reddito disponibile delle famiglie”, Istat, seminario “I conti economici nazionali per settore istituzionale: le nuove stime secondo il Sec95”, Rome, 2003.

Agostinelli, A. and Di Veroli, N. “L’attribuzione dei redditi primari: gli utili distribuiti dalle imprese e gli altri redditi da capitale”, Istat, seminario “I conti economici nazionali per settore istituzionale: le nuove stime secondo il Sec95”, Rome, 2003.

Bagozzi, R. (eds.) “Measuring in market research: basic principal of questionnaire design”, in: Principles of Marketing Research, Blackwell, Business, 1-49, 1994.

Banca d’Italia “Household income and wealth in 2004”, Supplement to the Statistical Bulletin, Year 16, no. 7, 2006.

Banca d'Italia "Household Wealth in Italy 2007", Supplement to the Statistical Bulletin, Volume 18, no. 76, 2008.

Biancotti, C., D'Alessio, G. and Neri, A. "Measurement Error In The Bank Of Italy's Survey Of Household Income And Wealth", Review of Income and Wealth, Blackwell Publishing, 54, 466-493, 2008.

Bollinger, C.R. "Bounding mean regressions when a binary regressor is mismeasured", Journal of Econometrics, 73, 387-399, 1996.

Bollinger, C.R. and David, M.H. "I didn't tell, and I won't tell: dynamic response error in the SIPP", Journal of Applied Econometrics, 20, 563-569, 2005.

Brandolini, A. "The distribution of personal income in post-war Italy: source description, data quality, and the time pattern of income inequality", Bank of Italy, Discussion paper no. 350, 1999.

Cannari, L. and D'Alessio, G. "Housing Assets in the Bank of Italy's Survey of Household Income and Wealth", in Dagum and Zenga (eds.), Income and Wealth Distribution, Inequality and Poverty, Berlin, Springer-Verlag, 1990.

Cannari, L. and D'Alessio, G. "L'indagine sui bilanci delle famiglie. Metodi, qualità, linee evolutive", mimeo, Banca d'Italia, 2008.

Cannari, L. and Faiella I. "House prices and housing wealth in Italy", in the proceedings of the conference Housing wealth in Italy, held in Perugia, 2007.

Cannari, L. and Violi, R. "Reporting Behaviour in the Bank of Italy's Survey of Italian Household Income and Wealth", Research on Economic Inequality, 6, 117-130, 1995.

Consolini, P., Di Marco, M., Ricci, R. and Vitaletti, S. "Administrative and Survey Microdata on Self-Employment: the Italian Experience with the EU-SILC Project", presented at the 29th IARIW Conference, Joensuu, Finland, 2006.

Couper, M.P. and others (eds.) "Computer Assisted Survey Information Collection", John Wiley and Sons, New York, 1998.

Coromaldi, M. and Guerrera, D. "Modello di microsimulazione ECONLAV: la costruzione del data-set di input", ISFOL, 2006.

D'Alessio, G. and Faiella, I. "Non-response Behaviour in the Bank of Italy's Survey of Household Income and Wealth", Bank of Italy, Discussion paper no. 462, 2002.

D'Aurizio, L., Faiella, I., Iezzi, S. and Neri, A. "The underreporting of financial wealth in the Survey on Household Income and Wealth", Bank of Italy, Discussion paper no. 610, 2006.

Faiella, I. and Gambacorta, R. "The weighting process in the SHIW", Bank of Italy, Discussion paper no. 636, 2007.

Fowler, F. and Mangione, T. "Standardized Survey Interviewing: Minimizing Interviewer-Related Error", Beverly Hills, CA: Sage Publications, 1990.

Griliches, Z. and Hausman, J.A. "Errors in variables in panel data", Journal of Econometrics, 31, 93-118, 1986.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. "Survey Methodology", John Wiley and Sons, New York, 2004.

Imbens, G.W. and Manski, C.F. "Confidence Intervals for Partially Identified Parameters", Econometrica, Econometric Society, vol. 72(6), 1845-1857, 2004.

Istat "Metodologie di stima degli aggregati di contabilità nazionale a prezzi correnti", 2004.

Istat "Integrazione di dati campionari EU-SILC con dati di fonte amministrativa", Metodi e Norme, 38, 2009.

Kan, M.Y.Y. and Pudney S. "Measurement error in stylized and diary data on time use", Sociological Methodology, 38, 101-132, 2008.

Moore, J.C., Stinson, L.L. and Welniak, E.J. "Income measurement error in surveys: a review", Journal of Official Statistics, 16, 331-361, 2000.

Pedace, R. and Bates, N. "Using administrative records to assess earnings reporting error in the survey of income and program participation", Journal of Economic and Social Measurement, 26, 173-192, 2000.

Pissarides, C.A. and Weber, G. "An expenditure-based estimate of Britain's black economy", Journal of Public Economics, 39, 17-32, 1989.

Pudney, S. "Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure", ISER working paper no. 9, 2008.

Rasch, G. "Probabilistic models for some intelligence and attainment tests", Danish Institute for Educational Research, Copenhagen, 1960.

Schennach, S., Hu, Y. and Lewbel, A. “Nonparametric identification of the classical errors-in-variables model without side information”, CeMMAP working papers CWP14/07, Institute for Fiscal Studies, 2007.

Tourangeau, R. “Cognitive Science and Survey Methods”, in: Jabine, T., Straf, M., Tanur, J. and Tourangeau, R. (eds.): “Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines”, Washington DC, National Academy Press, 1984.

Tourangeau, R., Rips, L. and Rasinski, K. “The Psychology of Survey Response”, Cambridge: Cambridge University Press, 2000.

Turner, C., Forsyth, B., O’Reilly, J., Cooley, P., Smith, T., Rogers, S. and Miller, H. “Automated Self-interviewing and the Survey of Measurement of Sensitive Behaviours”, in Couper, M.P. and others (eds.) “Computer Assisted Survey Information Collection”, John Wiley and Sons, New York, 1998.

Wooldridge, J.M. “Introductory Econometrics”, 2nd edition, Thomson, 2003.

7 Methodological appendix

7.1 Income and jobs: a comparison between the SHIW and the NA

In order to assess the degree of accuracy of microeconomic data on income collected by the SHIW we compare them with corresponding macro figures from the National Accounts. We only consider the year 2004; comparisons referred to previous years can be found in Brandolini (1999), in Coromaldi and Guerrera (2006) and in Cannari and D'Alessio (2008).

The two sources have been matched according to the following criteria. Payroll income includes net wages and salaries and fringe benefits. Pensions and net transfers include pensions, arrears and wage supplementation. The item "Income from self-employment in economic producing units with 5 dependents or less and actual rents"compares with the Italian NA item "Mixed income share transferred from productive to consumer households" not featuring in the ESA95 framework, where income from activity in producing households (unincorporated enterprises) is classified. The item "Income from self-employment in economic producing units with more than 5 dependents"compares with the NA item "Withdrawals from income of quasi-corporations“. The item "Imputed rents"is contrasted with the NA item "Gross operating surplus"since it is essentially (by more than 80 per cent; Agostinelli, 2003) represented by imputed rents from owner-occupied dwellings, with the residual part being domestic services, agricultural production for own use and maintenance made by the owner. The item "Entrepreneurial income and income from financial assets"includes three NA items, "Dividends" "Other distributed income of corporations"(not featuring in the ESA95 manual) and "Net interest“. Social contributions paid by dependent and independent workers have been subtracted from aggregates in the NA. Taxes on income have been assigned pro quota to the different income categories excluding imputed rents (as in Brandolini, 1999). Depreciation is taken into account for the items "Income from self-employment in economic producing units with 5 dependents or less and actual rents"and for "Imputed rents“. Further details on NA definitions can be found in Agostinelli (2003) and in Agostinelli and Di Veroli (2003).

A number of caveats must be introduced as the comparison, while necessary, is inherently limited in scope: the reference population is different in the two sources (residents net of institutionalized persons in the SHIW versus residents and non-residents in the NA); the SHIW suffers from the typical problems of sample surveys on households (lack of representativeness of some categories of income recipients, reticence, missing responses that were not randomly distributed); NA aggregates are subject to frequent revisions as new information becomes available, and include explicit adjustments to cope with undeclared work, which is likely to be only partly captured by the survey; income definitions sometimes do not match perfectly in the two sources. On this point, the SHIW asks for data on income net of taxes and social contributions, while the NA are built on a gross basis. The assignment of taxes and social contributions to each separate income category in the NA is an unavoidable source of measurement error. For all these reasons, although the NA are the best available benchmark, it is worth stressing that our adjustment procedure does not aim at aligning microdata to them exactly.

Table 1 shows that in 2004 the micro estimates of payroll income and of income from pensions and net transfers were lower than the corresponding macro figures by 12 and 31 per cent, respectively. In addition, the SHIW estimate of income from self-employment was about 43 per cent of the corresponding macro figure, mainly because of income generated in the largest units (29 per cent). Entrepreneurial income and income from financial assets fared even worse (13 per cent of the corresponding NA aggregate). By contrast, we detected a sizeable overestimation of the imputed rents, presumably reflecting the estimation procedure that Istat adopts to meet specific requirements of the ESA95 (Istat, 2004). In particular, imputed rents are estimated not by relying on subjective evaluations collected through the Survey on Consumption Expenditure (which provides an estimate of the imputed rents that is broadly comparable with the SHIW estimate), but using the value of actual rents for similar dwellings.

7.2 The adjustment procedure for financial assets: methodological details

In order to improve upon data comparability, the design and implementation of the survey conducted by the banking group were planned to be as similar as possible to those of the SHIW. The reference population was made up of customers who authorized the disclosure of their data for research purposes, as required under Italian law. The population was stratified according to its geographical area of residence, municipality size and, most importantly, to the financial wealth held in the bank. The survey collected data on 1,834 households. Assuming that respondents were representative of customers of other banks, this information can be extrapolated to the SHIW data. In order to strengthen that assumption the sampling weights were post-stratified to reproduce the distribution of the Italian population of banks' customers (e.g. geographical area).

The econometric framework is based on the hurdle models (Wooldridge, 2003). The first step estimates the non-reporting of ownership: the response variable, obtained from the administrative records, is a dummy for the actual holding of an asset. The probability of non-reporting is estimated by including among the covariates a dummy for the declared asset ownership in the interview. The analysis is carried out separately for six financial assets (deposits and repos, government bonds, private bonds, quoted shares, mutual funds and managed savings) and for financial liabilities. The second step models the underreporting of the amount held, defined as the ratio between the actual and the reported amount for each class of assets. This ratio is computed at the individual level and is assumed to be a proxy for reticence at the household level. The log of the ratio is regressed on the household declared amount, its square, and a set of sociodemographic characteristics. The third step fits the preceding estimates to the SHIW data. For each financial instrument, the estimated probability of holding a given asset is fitted at the household level. A random experiment is then used to impute ownership to households which are likely to possess an asset, whether they declare it or not. We reconstruct for every asset the amount owned by the households to whom the experiment attributes ownership, even if they did not declare it. Finally, the estimated coefficients of misreporting on amounts are fitted to SHIW data to obtain an inflation factor (less often a deflation factor) for the declared amount.

8 Tables and figures

Table 1.National Accounts and SHIW income estimates

Source of income	NA	SHIW(1)	SHIW(2)	SHIW(1)/NA	SHIW(2)/NA
	Millions of euro			Percentages	
Payroll income	303,466	267,962	285,656	88.3	94.1
Imputed rents	57,278	125,733	164,575	219.5	287.3
Income from self-employment in units					
with up to 5 employees and actual rents	167,406	79,440	162,689	47.5	97.2
with more than 5 employees	52,418	15,447	30,297	29.5	57.8
Entrepren. income, income from financ. assets	111,877	14,951	80,553	13.4	72.0
Pensions and net transfers	222,754	153,969	158,831	69.1	71.3
	NA	SHIW(1)	SHIW(2)	SHIW(1)-NA	SHIW(2)-NA
	Share of total income (%)			Percentage points	
Payroll income	33.2	40.8	32.4	7.6	−0.8
Imputed rents	6.3	19.1	18.6	12.8	12.3
Income from self-employment in units					
with up to 5 employees and actual rents	18.3	12.1	18.4	−6.2	0.1
with more than 5 employees	5.7	2.3	3.4	−3.4	−2.3
Entrepren. income, income from financ. assets	12.2	2.3	9.1	−9.9	−3.1
Pensions and net transfers	24.3	23.4	18.0	−0.9	−6.3

SHIW(1) = unadjusted figures; SHIW(2) = adjusted figures.

Table 2. Workers by main income source

Main source of income	LFS*	NA	SHIW	SHIW/LFS	NA/LFS
From employment	16,117,254	18,029,080	16,115,090	100.0	89.4
From self-employment	6,287,176	6,226,750	6,008,486	95.6	96.5
Total	22,404,430	24,255,830	22,123,576	98.7	91.2

* Labour force survey (yearly average).

Table 3. Jobs held

	NA	SHIW	SHIW/NA
Employees	20,055,000	16,633,430	82.9
Self-employed	10,976,300	6,292,250	57.3
Total	31,031,300	22,925,680	73.9

Table 4. Effect of the new weighting scheme

	Recipients (%)		Mean values (€)		Coeff. of variation	
Source of income	Initial	Final	Initial	Final	Initial	Final
Payroll income	30.9	32.6	14,458	15,080	1.2%	1.2%
Income from self-employment	8.4	11.8	20,436	20,929	5.3%	4.1%
Pensions	24.3	25.6	10,417	10,825	1.1%	0.9%
Property income	65.0	70.6	4,283	3,993	2.4%	2.0%
Financial assistance	3.9	5.1	1,562	1,605	13.1%	13.5%

Table 5. Probability of a secondary source of income by main source

Main source	Payroll income						Self -empl.	
Secondary source	Self-empl.		Pensions		Financ. assistance		Pensions	
Geogr. area (base: South/ Islands)	Coeff.	St. err.	Coeff.	St. err.	Coeff.	St. err.	Coeff.	St. err.
North	0.018	0.001	−0.153	0.002	−0.118	0.002	0.317	0.002
Center	0.080	0.001	0.033	0.002	−0.306	0.002	0.145	0.003
Education (base: $\geqq$ degree)								
None/primary	−0.558	0.003	0.292	0.003	0.834	0.002	0.663	0.003
Secondary school	−0.360	0.002	0.180	0.002	0.013	0.002	0.074	0.003
Age class (base: > 64)								
< 25	−0.612	0.004	−1.998	0.007	−0.134	0.012	−2.484	0.020
25 - 44	−0.006	0.003	−1.183	0.003	1.071	0.011	−1.680	0.008
45 - 64	−0.112	0.003	0.515	0.003	0.704	0.011	0.711	0.007
Household size (base: > 4)								
1 member	0.126	0.002	0.072	0.003	−0.334	0.003	0.300	0.004
2 members	−0.095	0.002	0.439	0.002	0.135	0.002	0.508	0.003
3 members	−0.085	0.002	0.164	0.002	−0.040	0.002	0.156	0.003
4 members	0.013	0.002	−0.545	0.003	0.199	0.002	−0.448	0.004
Gender								
Male	0.158	0.001	0.052	0.001	−0.190	0.001	0.143	0.002
Intercept	−2.395	0.003	−3.186	0.003	−4.044	0.011	−2.747	0.007
Observations	17,908						5,714	

Table 6. Probability of a secondary source of income by main source

Main source	Pensions				Capital income			
Secondary source	Payroll income		Self-emp. inc.		Payroll income		Self-emp. inc.	
Geogr. area (base: South/ Islands)	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.	Coef.	St. err.
North	−0.083	0.002	0.204	0.002	0.343	0.002	0.360	0.002
Center	0.205	0.002	0.091	0.002	0.147	0.002	0.206	0.002
Education (base: ≧ degree)								
None/primary	−0.564	0.003	−0.745	0.002	−0.760	0.003	−0.524	0.002
Secondary school	0.110	0.002	−0.209	0.002	−0.019	0.002	−0.137	0.002
Age class (base: > 64)								
< 25	0.266	0.007	−0.414	0.014	1.103	0.005	−0.055	0.005
25 - 44	0.685	0.005	−0.156	0.010	1.245	0.004	0.598	0.003
45 - 64	0.594	0.003	0.497	0.006	0.577	0.004	0.130	0.002
Household size (base: > 4)								
1 member	−0.202	0.004	−0.427	0.004	0.507	0.003	−0.031	0.003
2 members	−0.263	0.003	0.010	0.003	0.003	0.003	0.058	0.003
3 members	0.069	0.003	0.329	0.003	−0.069	0.003	0.053	0.003
4 members	0.608	0.003	0.296	0.004	−0.174	0.003	0.071	0.003
Gender								
Male	0.277	0.002	0.480	0.002	0.208	0.002	0.858	0.001
Intercept	−2.552	0.004	−3.206	0.006	−2.417	0.004	−1.629	0.002
Observations	12,832				5,256			

Table 7. Income recipients by source (percentages)

	Payroll		Self-empl.		Pensions		Fin. assistance	
Characteristics	Initial	Final	Initial	Final	Initial	Final	Initial	Final
Gender								
Male	38.4	41.5	15.3	19.8	23.9	25.4	5.7	5.9
Female	27.1	29.2	8.5	11.9	27.2	28.0	4.6	4.8
Age class								
<25	14.8	15.6	2.9	4.3	0.5	0.5	4.5	4.5
25-44	61.4	63.1	20.0	25.3	1.5	1.8	6.4	7.0
45-64	39.3	43.9	17.9	22.8	27.9	32.1	5.8	6.0
>64	0.4	4.3	2.1	5.7	93.2	93.2	3.1	3.1
Education								
None/primary school	6.9	9.0	3.5	5.5	47.7	48.4	3.8	3.9
Secondary school	44.9	47.7	14.8	19.6	14.5	15.8	5.6	5.9
University degree	53.0	56.1	26.8	31.9	13.1	14.6	7.4	7.9
Household size								
1 member	29.8	33.3	14.3	20.1	53.7	54.7	10.7	10.7
2 members	29.5	33.1	9.9	14.7	49.1	50.7	5.0	5.4
3 members	37.1	40.2	13.2	17.3	19.0	20.4	5.0	5.2
4 members	34.9	36.3	12.4	15.4	7.6	8.3	3.6	4.0
5 or more members	26.8	28.1	8.6	10.7	7.9	8.7	3.3	3.5
Geographical area								
North	37.3	40.2	13.1	17.2	26.2	27.6	4.8	5.1
Center	35.7	38.6	12.9	19.1	27.5	28.5	6.4	6.5
South and Islands	25.0	27.1	9.6	12.1	23.8	24.7	4.8	5.1
Total	32.6	35.2	11.8	15.7	25.6	26.8	5.1	5.4

Table 8. Individual income by source (€)

	Payroll		Self-empl.		Pensions		Fin. assistance	
Characteristics	Initial	Final	Initial	Final	Initial	Final	Initial	Final
Gender								
Male	15,648	15,007	23,530	23,352	12,523	11,813	561	2,323
Female	12,868	12,392	15,101	18,850	8,675	8,434	2,715	2,925
Age class								
<25	9,160	8,885	8,408	16,011	4,082	4,082	1,997	2,020
25-44	14,370	14,221	20,855	20,488	6,120	5,233	1,660	2,582
45-64	16,563	15,504	20,876	26,280	11,708	10,238	1,259	3,102
>64	14,584	7,458	29,440	10,162	10,060	10,060	1,176	2,503
Education								
None/primary school	11,549	10,281	14,290	21,889	8,663	8,550	1,988	2,586
Secondary school	13,806	13,392	18,524	23,021	12,582	11,546	1,639	2,456
University degree	20,670	19,984	32,504	14,320	20,731	18,525	68	3,536
Household size								
1 member	16,688	15,540	26,133	19,650	10,300	10,115	−931	1,762
2 members	14,908	14,039	23,995	19,030	10,218	9,894	1,459	2,575
3 members	14,390	13,743	15,649	17,128	10,693	9,971	2,697	2,959
4 members	13,928	13,649	20,545	28,154	11,328	10,407	3,214	3,286
5 or more members	12,735	12,646	17,196	24,608	10,173	9,387	2,433	2,449
Geographical area								
North	15,137	14,558	23,806	19,823	11,138	10,607	1,623	2,867
Center	15,523	14,958	20,113	27,550	11,000	10,601	717	2,134
South and Islands	12,360	11,813	14,796	19,742	9,053	8,742	2,091	2,597
Total	14,458	13,888	20,412	21,601	10,417	9,991	1,562	2,603

Table 9. Labour income and market value of main residence

Dep. variable: ratio of labour income to market value of main residence		
	Coeff.	P-value
Geogr. area (base: South and Islands)		
North	−0.050	0.00
Center	0.040	0.01
Municipality size (base: medium)		
Small municipalities	−0.048	0.58
Large municipalities	0.063	< .0001
Male	−0.008	0.48
Education (base: secondary school)		
Primary	−0.066	0.25
Tertiary	−0.006	0.86
Labour income quartile (base: First)		
Second	0.037	0.32
Third	0.093	0.01
Fourth	0.182	< .0001
Age class (base: <30)		
30-40	0.041	0.11
40-50	−0.039	0.01
>50	0.037	0.00
Luxury dwelling	−0.072	< .0001
Logarithm of house surface	−0.099	< .0001
Low-rate dwelling	0.125	0.07
Interviewer's characteristics		
free-lance	0.021	0.20
graduate	−0.208	< .0001
main job as interviewer	0.017	0.25
Inheritance/gift	−0.050	< .0001
Intercept	0.662	< .0001
Observations	346	
Adj R-square	0.72	

Table 10. Adjustment of income from self-employment

	Mean values		Variation
Characteristics	Initial	Final	%
Gender			
Male	26,600	32,858	23.5
Female	17,770	30,201	70.0
Age class			
<30	9,485	23,794	150.9
30-40	23,244	29,265	25.9
40-50	24,442	38,107	55.9
>50	54,384	12,392	−77.2
Education			
None/primary school	16,870	34,333	103.5
Secondary school	21,110	34,445	63.2
University degree	38,046	18,783	−50.6
Household size			
1 member	30,950	32,445	4.8
2 members	28,848	29,887	3.6
3 members	18,340	26,112	42.4
4 members	22,316	36,600	64.0
5 or more members	19,521	35,517	81.9
Geographical area			
North	28,069	28,754	2.4
Center	23,049	47,943	108.0
South and Islands	16,446	25,911	57.6
Total	23,438	31,907	36.1

Table 11. Adjustment of income from financial assets

	Percent. of owners		Mean values (€)	
Characteristics	Initial	Final	Initial	Final
Gender				
Male	65.3	66.5	550	1, 514
Female	52.5	53.4	472	1, 577
Age class				
<25	15.3	15.5	162	1, 380
25-44	72.8	74.1	443	1, 141
45-64	71.9	73.4	592	1, 635
>64	74.3	75.6	618	2, 098
Education				
None/primary school	42.4	43.4	244	1, 439
Secondary school	64.9	65.9	481	1, 427
University degree	84.9	86.1	1, 347	2, 506
Household size				
1 member	75.9	77.2	654	2, 419
2 members	73.6	74.8	571	1, 726
3 members	59.6	60.3	572	1, 447
4 members	48.9	49.8	350	989
5 or more members	34	35.2	304	988
Geographical area				
North	72	72.4	644	1, 660
Center	64.1	65	484	1, 760
South and Islands	39.1	41.1	239	1, 098
Total	58.7	59.8	514	1, 543

Table 12. Adjustment of interest on financial liabilities

	Percent. of owners		Mean values (€)	
Characteristics	Initial	Final	Initial	Final
Gender				
Male	22.6	26.4	1,165	1,246
Female	15.1	18.2	1,104	1,239
Age class				
<25	6.2	7.6	827	958
25-44	31.1	35.5	1,426	1,581
45-64	24.8	29.8	921	1,000
>64	6.8	9.7	466	531
Education				
None/primary school	7.2	9.3	556	646
Secondary school	23.9	27.9	1,178	1,234
University degree	30.2	36	1,545	2,013
Household size				
1 member	12.7	15.9	1,372	1,683
2 members	17.1	21.3	1,535	1,490
3 members	21.3	25.6	812	1,008
4 members	22.1	24.7	1,156	1,235
5 or more members	14.6	17.6	976	982
Geographical area				
North	22.4	26	1,380	1,472
Center	18.9	24	936	1,104
South and Islands	13.9	16.4	799	894
Total	18.7	22.2	1,140	1,243

Table 13. Adjustment of income from rents

	Percent. of owners		Mean values (€)	
Characteristics	Initial	Final	Initial	Final
Gender				
Male	47.8	50.3	6, 131	6, 904
Female	48.3	50.7	5, 226	5, 745
Age class				
<25	4.5	5.1	5, 676	5, 309
25-44	45.8	49.2	5, 326	5, 979
45-64	72.5	75.9	5, 926	6, 549
>64	75.1	77.2	5, 657	6, 405
Education				
None/primary school	45.8	47.3	4, 227	4, 751
Secondary school	47.8	50.6	5, 976	6, 552
University degree	60.9	64.6	8, 690	10, 027
Household size				
1 member	75.9	77.6	6, 959	7, 693
2 members	63.9	67.9	5, 537	6, 309
3 members	46	48	5, 302	6, 007
4 members	35.2	37.6	5, 381	5, 812
5 or more members	23.4	24.6	4, 366	4, 660
Geographical area				
North	50.3	53.4	6, 480	7, 220
Center	52.4	54.7	6, 874	7, 795
South and Islands	42.9	44.6	3, 655	3, 936
Total	48.1	50.5	5, 663	6, 305

Table 14. Interdecile ratio (IDR) and the Gini index

Source of income	Unadjusted IDR	Adjusted IDR	Unadjusted Gini	Adjusted Gini
Payroll income	3.49	4.40	0.262	0.291
Income from self-employment	9.50	39.75	0.493	0.564
Pensions and financial assistance	4.15	5.20	0.317	0.330
Actual rents	17.07	54.54	0.553	0.601
Imputed rents	7.75	9.17	0.418	0.471
Net income from financial assets	100.76	153.42	0.857	0.727
Disposable personal income	6.85	7.95	0.385	0.427
Equivalent income (OECD modified scale)	4.42	5.18	0.341	0.364

Table 15. Results with alternative models of adjustment

Source of income	Baseline*	Alternative 1**	Alternative 2***
	Mean values (€, millions)		
Payroll income	285,656	285,656	283,546
Imputed rents	164,575	165,021	165,073
Income from self-employment in units			
with up to 5 employees and actual rents	162,689	163,217	158,404
with more than 5 employees	30,297	31,146	31,147
Entrepren. income and income from financ. assets	80,553	80,645	79,931
Pensions and net transfers	158,831	158,831	158,831

* Order as indicated in the paper. Adj. values at each step are used in the following steps.

** Income from self-empl. as the final step of the process. Adj. values at each step are used in the following steps.

*** Order as indicated in the paper. Adj. values at each step are not used in the following steps.

Table 16. Probability of misreporting: random intercept logistic model

Variables	Coef.	P>|z|	95% Conf. Interval	
Item				
Additional payroll income	−8.424	0.000	−8.690	−8.157
Additional self-empl. income	−8.053	0.000	−8.315	−7.791
Primary self-empl. income	−7.209	0.000	−7.464	−6.953
Additional income from pension	−9.062	0.000	−9.341	−8.782
Income from deposits	−3.715	0.000	−3.960	−3.471
Income from government bonds	−6.976	0.000	−7.230	−6.722
Income from private bonds	−5.882	0.000	−6.132	−5.632
Income from shares	−6.575	0.000	−6.827	−6.323
Income from mutual funds	−5.696	0.000	−5.946	−5.447
Income from managed savings	−7.947	0.000	−8.208	−7.687
Interests on financial liabilities	−5.753	0.000	−6.003	−5.503
Income from actual rents	−7.558	0.000	−7.815	−7.301
Income from imputed rents	−7.243	0.000	−7.498	−6.988
Income from financial assistance	−10.617	0.000	−10.981	−10.252
Age	0.081	0.000	0.074	0.088
Age squared	0.000	0.000	0.000	0.000
Educational qualification (base: none)				
Primary school certificate	0.276	0.000	0.173	0.380
Lower secondary school certificate	1.098	0.000	0.961	1.235
Upper secondary school diploma	1.381	0.000	1.241	1.521
University degree	1.503	0.000	1.354	1.652
Geographical area (base: North)				
Center	−0.215	0.000	−0.263	−0.168
South and Islands	−0.956	0.000	−1.004	−0.907
Gender Female (base: male)	−0.277	0.000	−0.317	−0.236
Work status (base: payroll employee)				
Self-employed	0.567	0.000	0.498	0.636
Not employed	−1.340	0.000	−1.399	−1.281
Quintiles of household wealth (base: 1st quintile)				
2nd quintile	0.290	0.000	0.198	0.382
3rd quintile	0.339	0.000	0.261	0.417
4th quintile	0.394	0.000	0.331	0.457
5th quintile	0.603	0.000	0.544	0.663
Proxy respondent on payroll income	0.751	0.000	0.681	0.822
Proxy respondent on self-employment income	0.672	0.000	0.529	0.814
CAPI	−0.081	0.000	−0.125	−0.038
Edu-inter	0.798	0.000	0.699	0.897
Sex-inter	0.323	0.000	0.279	0.366
Age-inter	0.201	0.000	0.152	0.251
Comprehension	0.077	0.000	0.064	0.089
Standard error. random intercept	0.5325			

Table 17. Probability of misreporting: random intercept ordinal logistic model

Variables	Coef.	P>|z|	95% Conf. Interval	
Item				
Additional payroll income	2.194	0.000	1.912	2.475
Additional self-empl. income	2.561	0.000	2.284	2.839
Primary self-empl. income	3.432	0.000	3.161	3.704
Additional income from pension	1.528	0.000	1.235	1.822
Income from deposits	6.421	0.000	6.154	6.689
Income from government bonds	3.653	0.000	3.382	3.923
Income from private bonds	4.714	0.000	4.445	4.982
Income from shares	4.031	0.000	3.761	4.301
Income from mutual funds	4.869	0.000	4.600	5.137
Income from managed savings	2.667	0.000	2.391	2.944
Interests on financial liabilities	4.839	0.000	4.571	5.107
Income from actual rents	3.066	0.000	2.793	3.340
Income from imputed rents	3.370	0.000	3.098	3.641
Age	0.072	0.000	0.066	0.079
Age squared	0.000	0.000	0.000	0.000
Educational qualification (base: none)				
Primary school certificate	0.286	0.000	0.190	0.381
Lower secondary school certificate	1.058	0.000	0.931	1.185
Upper secondary school diploma	1.332	0.000	1.202	1.461
University degree	1.471	0.000	1.333	1.609
Geographical area (base: North)				
Center	−0.189	0.000	−0.233	−0.146
South and Islands	−0.766	0.000	−0.811	−0.722
Gender Female (base: male)	−0.259	0.000	−0.297	−0.222
Work status (base: payroll employee)				
Self-employed	0.636	0.000	0.573	0.698
Not employed	−1.172	0.000	−1.226	−1.118
Quintiles of household wealth (base: 1st quintile)				
2nd quintile	0.304	0.000	0.219	0.389
3rd quintile	0.388	0.000	0.316	0.460
4th quintile	0.443	0.000	0.385	0.501
5th quintile	0.649	0.000	0.594	0.704
Proxy respondent on payroll income	0.617	0.000	0.552	0.681
Proxy respondent on self-employment income	0.469	0.000	0.340	0.599
CAPI	−0.072	0.000	−0.113	−0.032
Edu-inter	0.710	0.000	0.619	0.802
Sex-inter	0.297	0.000	0.257	0.337
Age-inter	0.192	0.000	0.147	0.237
Comprehension	0.046	0.000	0.035	0.057

(1) Income from financial transfers is omitted for multicollinearity. (2) Cut-offs and their standard errors are not displayed.

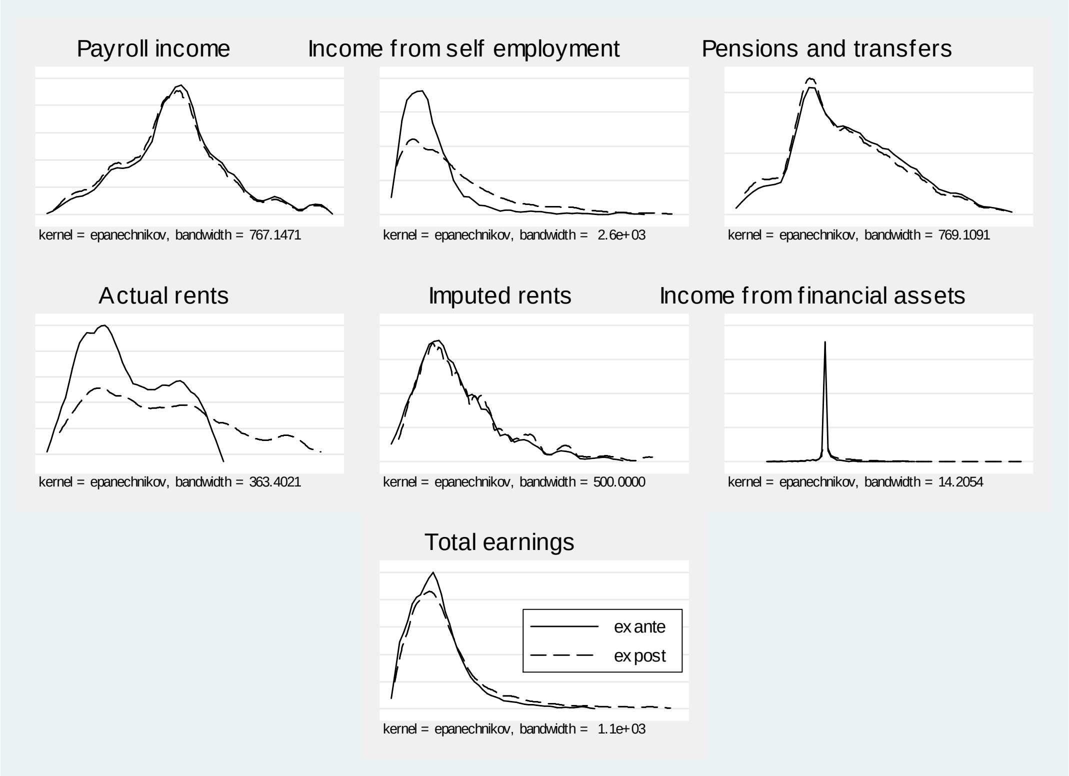
Table 18. Distribution of respondents according to the estimated number of response errors

Number of response errors	percent.	cum. percent.
0	27.5	27.5
1	18.4	45.9
2	21.2	70.1
3	17.4	84.5
4	8.8	93.3
5	3.8	97.1
> 5	2.9	100.0

Table 19. Response error by household tenths

Household tenths	percent. share of response error	mean (€) response error
up to 1st decile	0.0	10
from 1st to 2nd decile	0.1	40
from 2nd to 3rd decile	0.3	192
from 3rd to 4th decile	0.6	430
from 4th to 5th decile	1.2	843
from 5th to 6th decile	2.2	$1,523$
from 6th to 7th decile	4.0	$2,720$
from 7th to 8th decile	7.7	$5,265$
from 8th to 9th decile	16.7	$11,392$
over the 9th decile	67.2	$45,815$
Average response error=$4,013$; Gini = 0.799.		

Figure 1: Change in earnings distribution (kernel density estimate)


Payroll income
Income from self employment
Pensions and transfers
kernel = epanechnikov, bandwidth = 767.1471
kernel = epanechnikov, bandwidth = 2.6e+03
kernel = epanechnikov, bandwidth = 769.1091
Actual rents
Imputed rents
Income from financial assets
kernel = epanechnikov, bandwidth = 363.4021
kernel = epanechnikov, bandwidth = 500.0000
kernel = epanechnikov, bandwidth = 14.2054
Total earnings
ex ante
ex post
kernel = epanechnikov, bandwidth = 1.1e+03